SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 1–28

TRANSMISSION, REFLECTION, AND SECOND-HARMONIC GENERATION IN A NONLINEAR WAVEGUIDE*

ROBERTO CAMASSA†, ALP FINDIKOGLU‡, AND GRANT LYTHE§

Abstract. We present an experimental, analytical, and numerical investigation of the passage of electromagnetic signals through a device with voltage-dependent differential capacitance. This dependence gives rise to the device's nonlinear response, which can then be tuned by an externally applied static electric field. The system is modeled with a wave equation for the current and the charge density with continuity conditions at the boundaries between two linear regions and the nonlinear medium they sandwich. We derive asymptotic formulae for transmission and reflection coefficients of a monochromatic signal and its nonlinearity-induced second harmonics. Predictions based on this analysis are then compared with numerical and experimental results, across a range of parameter values, including those tuning the nonlinearity by means of an imposed voltage. The experiments are carried out at microwave frequencies using 1cm^2 devices consisting of a superconducting thin film meandering waveguide on a nonlinear dielectric substrate.

Key words. transmission and reflection, tunable nonlinearity, finite length nonlinear medium, second harmonic, differential capacitance, boundaries, method of characteristics, noise, numerical solution

AMS subject classifications. 35-04, 35L70, 78-05, 78M20

DOI. 10.1137/040615183

1. Introduction.

In this paper we study transmission, reflection, and generation of harmonics when an electromagnetic signal passes through a finite-length region, which represents a nonlinear medium coupled via realistic boundary conditions to its environment. Our theory is developed with actual experimental devices in mind, but it is applicable to any system where waves pass through a medium with voltage-dependent differential capacitance; i.e., the charge does not simply increase linearly with voltage [1] and where losses are not so strong as to overwhelm nonlinear effects. Of particular interest in this class of media is the consequence that the functional form of nonlinearity is tunable: the device response can be controlled with an external bias voltage, which enables exploration of dynamical behaviors characteristic of quadratic and cubic nonlinear media within the same apparatus [2, 3, 4, 5, 6, 7, 8, 9, 10, 11].

The experimental data we collect are obtained by sending electromagnetic signals along an 8cm length meandering waveguide consisting of superconducting electrodes on the surface of a 10mm × 10mm × 0.5mm nonlinear dielectric crystal of strontium titanate [10, 11, 12, 13, 14, 15]. Electromagnetic waves in the waveguide with wave-

*Received by the editors September 15, 2004; accepted for publication (in revised form) March 23, 2005; published electronically October 3, 2005.

http://www.siam.org/journals/siap/66-1/61518.html

†Department of Mathematics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (camassa@amath.unc.edu). This author was supported by NSF grant DMS-0104329 and from DOE CCPP and BES programs through the Theoretical Division of Los Alamos National Laboratory.

‡Superconductivity Technology Center, Los Alamos National Laboratory, Los Alamos, NM 87545 (findik@lanl.gov).

§Department of Applied Mathematics, University of Leeds, Leeds LS2 9JT, UK (grant@maths.leeds.ac.uk).

lengths of a few cm have frequencies of a few hundred MHz (microwaves). We model this experimental setup with a wave equation coupling the evolution of voltage and current through a distributed capacitance that depends on voltage and temperature throughout the nonlinear region. Section 2 formulates this model in its nondimensional form used for analytical and numerical studies. After introducing the nonlinear wave equation and the definition of differential capacitance, we express it in terms of its characteristic variables and the corresponding Riemann invariants [16]. Because of the boundary conditions, this form cannot be immediately used to provide closed form solutions. We then proceed by carrying out an asymptotic expansion in these variables in the limit of low-amplitude incident signals, through a corresponding expansion of the boundary conditions. The analysis is simplified by the absence of dispersion and dissipation in the nonlinear wave equation and produces closed form expressions for the nonlinear contributions to experimentally measurable quantities such as transmission and reflection coefficients.

In section 3, we study the concrete case of a sinusoidal input from the left of the region. In the weakly nonlinear regime, we carry out an asymptotic expansion of the characteristics solution in the parameter a/v^*, where a is the amplitude of the input signal and v^* is the characteristic voltage associated with the variation of the differential capacitance in the nonlinear central region. The calculation of the second-harmonic generated signal is given in some detail in section 3 and Appendix A, since this establishes a basis for further research on this class of system. The analysis yields explicit formulae that characterize the frequency-doubling response caused by nonlinearity.

Section 4 reports on numerical simulations of the governing equations: two coupled PDEs for the charge and current fields in three regions. The charge and current fields are updated in such a way as to keep current and voltage continuous across the boundaries, even though the charge-voltage relationship has a discontinuity. Sinusoidal signals and noise, band-limited or white, are continuously input from the left. Fourier transforms of numerical time series in the steady state, at points in the left and right regions, are used to evaluate the transmitted and reflected signals. We find good quantitative agreement with analytical results. In particular, we are able to predict the amplitude of the second harmonic generated by the nonlinearity.

In section 5 we compare the analytical and numerical results with experimental data. These are collected for microwave signals and noise passing through a compact device that operates at temperatures that are easily attained using liquid nitrogen or helium and whose response can be tuned with a bias voltage of a few volts. The waveguide is a patterned superconducting thin film; because its lateral dimensions are much smaller than the wavelengths of the input signals, wave propagation is effectively one-dimensional. The source of the nonlinearity is the nonlinear dielectric substrate that the superconducting waveguide rests on [19, 20, 21]. We find very good agreement with all the major measurable quantities of interest, while qualitative agreement is achieved when effects neglected by our model, but known to become relevant in certain regimes (like high frequencies), come into play.

2. The governing equations. The dynamics of wave propagation along the transmission line is described by the wave equation for the current $i(x,t)$ (Coulomb s^{-1}) and the charge density $q(x,t)$ (Coulomb m^{-1})

$$\frac{\partial q(x,t)}{\partial t} = -\frac{\partial i(x,t)}{\partial x},$$

$$L\frac{\partial i(x,t)}{\partial t} = -\frac{\partial v(x,t)}{\partial x}, \tag{2.1}$$

where L is the inductance per unit length. A relationship between $q(x,t)$ and the voltage $v(x,t)$ is needed to close the set of equations (2.1). Let

$$q(x,t) = \mathcal{Q}(v(x,t)). \tag{2.2}$$

The "differential capacitance" [1] is the derivative of $\mathcal{Q}(v)$:

$$C_{\rm d}(v) = \frac{{\rm d}\mathcal{Q}(v)}{{\rm d}v}. \tag{2.3}$$

In a linear medium $q(x,t) = Cv(x,t)$, where C is constant. Then the differential capacitance is the constant C and the wave equation (2.1) is linear:

$$\frac{\partial^2}{\partial t^2}v(x,t) = (LC)^{-1}\frac{\partial^2}{\partial x^2}v(x,t). \tag{2.4}$$

The nonlinear wave equation we study in this work can be written as

$$LC_{\rm d}(v(x,t))\frac{\partial^2}{\partial t^2}v(x,t) + L\frac{\partial C_{\rm d}(v(x,t))}{\partial v}\frac{\partial v(x,t)}{\partial t} = \frac{\partial^2}{\partial x^2}v(x,t). \tag{2.5}$$

Motivated by the configuration of the experiments and by analogy with classical transmission-reflection problems, we study the propagation of waves through a nonlinear region (region II, $0 \le x \le l$) sandwiched between two semi-infinite regions (region I, $-\infty < x < 0$ and region III, $l < x < \infty$) where the wave equation is linear. Current and voltage are continuous at the boundaries. The situation is illustrated in Figure 2.1.

We shall assume that the relationship (2.2) is such that we can define its inverse: $v(x,t) = \mathcal{Q}^{-1}(q(x,t))$. The inductance and differential capacitances per unit length are as follows:

$$L = \begin{cases} L_{\rm I}, \\ \\ L_{\rm II}, \end{cases} \qquad C_{\rm d}(v) = \begin{cases} C_{\rm I}, & \text{regions I and III,} \\ \\ C_{\rm m}(v), & \text{region II,} \end{cases} \tag{2.6}$$

where $L_{\rm I}$, $L_{\rm II}$, and $C_{\rm I}$ are constants. We assume that $C_{\rm m}(v)$ is a positive even function with maximum at $v = 0$. See Figure 2.2. Since $C'_{\rm m}(0) = 0$ and $C''_{\rm m}(0) < 0$ we define the characteristic voltage associated with the differential capacitance curve by

$$v^* = \left(\frac{2C_{\rm m}(0)}{|C''_{\rm m}(0)|}\right)^{\frac{1}{2}}. \tag{2.7}$$

A constant "bias" voltage $v_{\rm b}$ is applied across the three regions. The voltage at time t and position x is thus the sum of $v_{\rm b}$ and the time-dependent voltages due to the input signals and their interactions. The input signals consist of one or more sinusoidal signals and broadband noise. Noise effects are interesting in their own right [11], but in this paper we shall use noise input as a technique to explore simultaneously the response at numerous frequencies.

In section 3 we consider the case where the input signal (the right-going wavetrain in region I) is given by $v_{\rm in} = a\cos(2\pi f(t - x/u))$. Our analysis is based on the

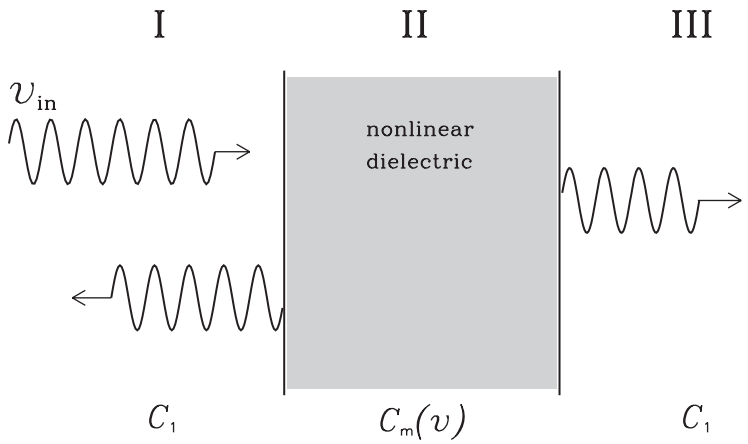


FIG. 2.1. *Schematic diagram for solution of the nonlinear wave equation in three one-dimensional regions. A constant bias voltage v_b is applied across all three regions and the right-going wavetrain in region* I *is the input signal, v_{in}. The late-time solution has the form of an incident and reflected wavetrain in region* I *and a transmitted wavetrain in region* III.

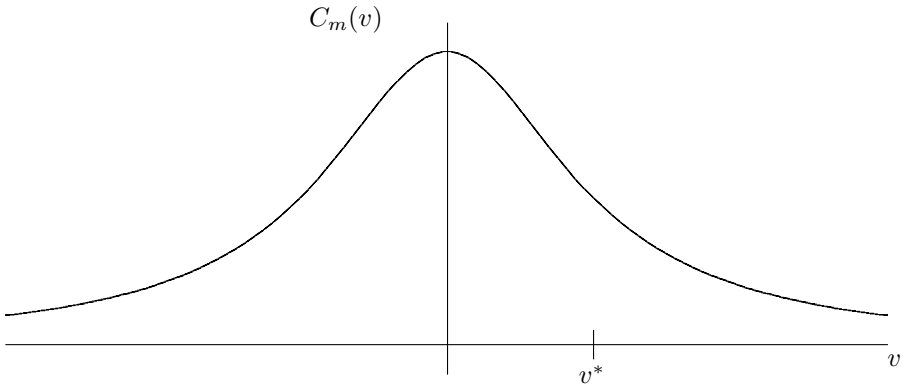


FIG. 2.2. *A typical graph of differential capacitance versus voltage. The voltage v^* is defined in* (2.7).

assumption that $a/v^* \ll 1$. The bias voltage is not assumed small but is fixed for any one experimental or numerical run. Thus the voltage values attained in any one experiment are in a small interval of the differential capacitance curve, but the full catalogue of nonlinear behaviors can be explored by performing runs at various values of the bias voltage.

We define the dimensionless fields

$$
\begin{aligned}
V(x,t) &= \frac{v(x,t)}{v^*}, \\
Q(x,t) &= \frac{q(x,t)}{C_{\text{II}} v^*}, \\
I(x,t) &= \left(\frac{L_{\text{II}}}{C_{\text{II}}}\right)^{\frac{1}{2}} \frac{i(x,t)}{v^*},
\end{aligned}
\tag{2.8}
$$

where

$$C_{\rm II} = C_{\rm m}(0). \tag{2.9}$$

In these variables, (2.1) is

$$\frac{\partial}{\partial t}\begin{pmatrix} Q(x,t) \\ I(x,t) \end{pmatrix} = -\left(L_{\rm II}C_{\rm II}\right)^{-\frac{1}{2}} \frac{\partial}{\partial x}\begin{pmatrix} I(x,t) \\ V(x,t) \end{pmatrix}, \tag{2.10}$$

where $V(x,t)$ is given as a function of $Q(x,t)$ by

$$V(x,t) = \begin{cases} \dfrac{L_{\rm II}C_{\rm II}}{L_{\rm I}C_{\rm I}}\, Q(x,t), & \text{regions I and III,} \\ \dfrac{1}{v^*}\mathcal{Q}^{-1}(C_{\rm II}v^*Q(x,t)), & \text{region II.} \end{cases}$$

We also nondimensionalize space and time by dividing by the length l of the waveguide and the transit time at the velocity of light in a vacuum, c:

$$X = \frac{x}{l}, \qquad T = t\frac{c}{l}. \tag{2.11}$$

So region I is $-\infty < X < 0$, region II is $0 \le X \le 1$, and region III is $1 < X < \infty$. Then (2.10) becomes

$$\frac{\partial}{\partial T}\begin{pmatrix} Q(X,T) \\ I(X,T) \end{pmatrix} = -U\frac{\partial}{\partial X}\begin{pmatrix} I(X,T) \\ V(X,T) \end{pmatrix}, \tag{2.12}$$

where

$$U = \frac{1}{c}\left(L_{\rm II}C_{\rm II}\right)^{-\frac{1}{2}}. \tag{2.13}$$

The propagation speed of small-amplitude signals in region II, $u_{\rm m}(v_{\rm b})$, is

$$u_{\rm m}(v_{\rm b}) = \left(L_{\rm II}C_{\rm m}(v_{\rm b})\right)^{-\frac{1}{2}}; \tag{2.14}$$

the constant $U = u_{\rm m}(0)/c$ is the dimensionless speed at zero bias. In our numerical work, we solve the nonlinear PDEs in the form (2.12).

Our analytical work, using the method of characteristics, proceeds by rewriting (2.12), using

$$\frac{{\rm d}Q}{{\rm d}T} = \frac{{\rm d}Q}{{\rm d}V}\frac{{\rm d}V}{{\rm d}T}, \tag{2.15}$$

in the form

$$\frac{\partial}{\partial X}\begin{pmatrix} I \\ V \end{pmatrix} = -U^{-1}\begin{pmatrix} 0 & G(V) \\ 1 & 0 \end{pmatrix}\frac{\partial}{\partial T}\begin{pmatrix} I \\ V \end{pmatrix}, \tag{2.16}$$

where the arguments (X,T) for the fields V and I have been suppressed and

$$G(V) = \begin{cases} G_{\rm l}, & \text{regions I and III,} \\ G_{\rm n}(V), & \text{region II,} \end{cases}$$

where

$$G_{\mathrm{l}} = \frac{L_{\mathrm{I}} C_{\mathrm{I}}}{L_{\mathrm{II}} C_{\mathrm{II}}}, \tag{2.17}$$

a constant, and

$$G_{\mathrm{n}}(V) = \frac{C_{\mathrm{m}}(v^* V)}{C_{\mathrm{II}}}. \tag{2.18}$$

As $V \to 0$, $G_{\mathrm{n}}(V) \to 1 - V^2 + \mathcal{O}(V^4)$.

In order to use the method of characteristics [16] in region II, we introduce the fields $\Gamma_+(X,T)$ and $\Gamma_-(X,T)$, defined as

$$\Gamma_\pm = H(V) \pm I, \tag{2.19}$$

where

$$H'(V) = (G_{\mathrm{n}}(V))^{\frac{1}{2}}. \tag{2.20}$$

The fields $\Gamma_\pm(X,T)$ are constants on two characteristic curves in the (X,T)-plane defined by the solutions $T_\pm(X)$ of the equations

$$\frac{\mathrm{d}T_\pm}{\mathrm{d}X} = \pm U^{-1}(G_{\mathrm{n}}(V(X,T_+)))^{\frac{1}{2}}. \tag{2.21}$$

We shall impose conservation of the fields $\Gamma_\pm(X,T)$ along characteristics using "initial times functions": given that a characteristic curve passes through $X = 1$ at time S, the time at which it passes through $X = 0$ is

$$\tau_\pm(S) = T_\pm(0)|_{T(1)=S}. \tag{2.22}$$

Then

$$\Gamma_\pm(0, \tau_\pm(S)) = \Gamma_\pm(1, S). \tag{2.23}$$

See Figure 2.3. At the boundary between region I and region II, $X = 0$, we denote

$$\begin{aligned} V(0,T) &= V_{\mathrm{I}}(T), \\ I(0,T) &= I_{\mathrm{I}}(T). \end{aligned} \tag{2.24}$$

At the boundary between region II and region III, $X = 1$, we denote

$$\begin{aligned} V(1,T) &= V_{\mathrm{III}}(T), \\ I(1,T) &= I_{\mathrm{III}}(T). \end{aligned} \tag{2.25}$$

The conditions (2.23) can be rewritten

$$H(V_{\mathrm{I}}(\tau_\pm(S))) \pm I_{\mathrm{I}}(\tau_\pm(S)) = H(V_{\mathrm{III}}(S)) \pm I_{\mathrm{III}}(S). \tag{2.26}$$

In section 3, we evaluate the transmitted signal as a function of the input signal using (2.26).

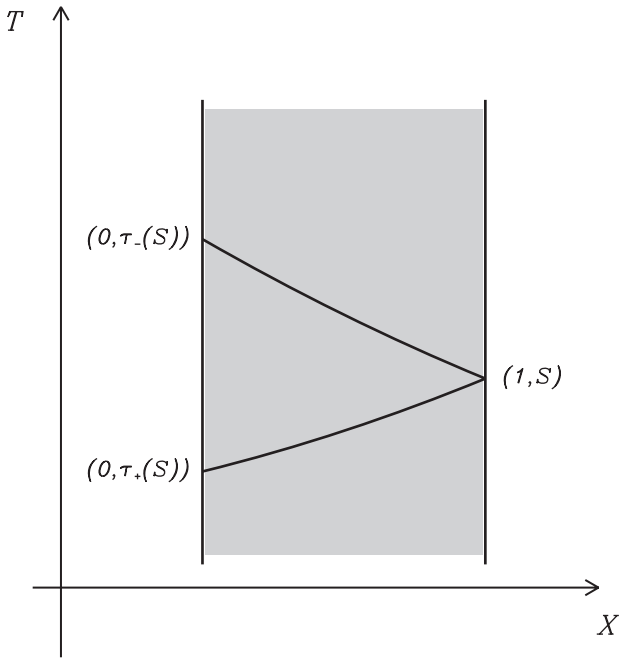


FIG. 2.3. *The "initial times" functions. The two characteristic curves that cross at* $(1, S)$ *follow paths that intersect* $X = 0$ *at times* $\tau_+(S)$ *and* $\tau_-(S)$.

3. Transmission, reflection, and harmonic generation: Analytical results. Let the input signal be sinusoidal with amplitude a at frequency f, that is,

$$v_{\text{in}}(x,t) = a\cos(2\pi f(t - x/u)), \tag{3.1}$$

where $u = (L_{\text{I}}C_{\text{I}})^{-\frac{1}{2}}$. We nondimensionalize the input amplitude and bias voltage,

$$A = \frac{a}{v^*} \qquad \text{and} \qquad V_{\text{b}} = \frac{v_{\text{b}}}{v^*}, \tag{3.2}$$

where v^* is defined in (2.7). The solution of (2.16) can be expanded as

$$V(X,T) = V_{\text{b}} + AV^{(0)}(X,T) + A^2V^{(1)}(X,T) + \mathcal{O}(A^3). \tag{3.3}$$

Thus

$$V_{\text{I}}(T) = V(0,T) = V_{\text{b}} + AV_{\text{I}}^{(0)}(T) + A^2V_{\text{I}}^{(1)}(T) + \mathcal{O}(A^3); \tag{3.4}$$

$I_{\text{I}}(T)$, $V_{\text{III}}(T)$ and $I_{\text{III}}(T)$ will be written in a similar way.

The function $H(V(X,T))$ is expanded as

$$\begin{aligned} H(V(X,T)) &= H(V_{\text{b}}) + H'(V_{\text{b}})(V(X,T) - V_{\text{b}}) + \frac{1}{2}H''(V_{\text{b}})(V(X,T) - V_{\text{b}})^2 + \mathcal{O}(A^3) \\ &= H(V_{\text{b}}) + G_{\text{n}}(V_{\text{b}})^{\frac{1}{2}}\Bigg(AV^{(0)}(X,T) \\ &\quad + A^2V^{(1)}(X,T) + \frac{1}{4}A^2\frac{G_{\text{n}}'(V_{\text{b}})}{G_{\text{n}}(V_{\text{b}})}(V^{(0)}(X,T))^2\Bigg) \\ &\quad + \mathcal{O}(A^3), \end{aligned}$$

and the characteristic curves and initial times functions as

$$T_\pm(X) = T_\pm^{(0)}(X) + AT_\pm^{(1)}(X) + \mathcal{O}(A^2) \tag{3.5}$$

and

$$\tau_\pm(S) = \tau_\pm^{(0)}(S) + A\tau_\pm^{(1)}(S) + \mathcal{O}(A^2). \tag{3.6}$$

Finally, the equality (2.26) is expanded as

$$\begin{aligned} &G_{\rm n}(V_{\rm b})^{\frac{1}{2}}\Big[AV_{\rm I}^{(0)}(\tau_\pm^{(0)}(S) + A\tau_\pm^{(1)}(S)) \\ &\qquad +A^2\left(V_{\rm I}^{(1)}(\tau_\pm^{(0)}(S)) + \frac{1}{4}\frac{G_{\rm n}'(V_{\rm b})}{G_{\rm n}(V_{\rm b})}V_{\rm I}^{(0)}(\tau_\pm^{(0)}(S))^2\right)\Big] \\ &\qquad \pm\left[AI_{\rm I}^{(0)}(\tau_\pm^{(0)}(S) + A\tau_\pm^{(1)}(S)) + A^2 I_{\rm I}^{(1)}(\tau_\pm^{(0)}(S))\right] \\ &= G_{\rm n}(V_{\rm b})^{\frac{1}{2}}\left[AV_{\rm III}^{(0)}(S) + A^2\left(V_{\rm III}^{(1)}(S) + \frac{1}{4}\frac{G_{\rm n}'(V_{\rm b})}{G_{\rm n}(V_{\rm b})}V_{\rm III}^{(0)}(S)^2\right)\right] \\ &\qquad \pm\left[AI_{\rm III}^{(0)}(S) + A^2 I_{\rm III}^{(1)}(S)\right] + \mathcal{O}(A^3). \end{aligned} \tag{3.7}$$

In section 3.1, we shall impose (3.7), keeping only terms proportional to A. The result is the classical transmission-reflection relation for finite-length linear media: perfect transmission is found at frequencies such that the length of region II is a multiple of half the wavelength of the input sinusoid. In section 3.2, we shall also keep terms proportional to A^2 in (3.7).

3.1. Lowest order. To order A, the characteristic curves (2.21) and (2.21) are straight lines:

$$T_\pm^{(0)}(X) = T_\pm^{(0)}(0) \pm \frac{X}{U_{\rm m}}, \tag{3.8}$$

where the normalized speed of propagation in region II is

$$U_{\rm m} = G_{\rm n}(V_{\rm b})^{-\frac{1}{2}}U = \frac{1}{c}\left(L_{\rm II}C_{\rm m}(V_{\rm b})\right)^{-\frac{1}{2}}. \tag{3.9}$$

Thus the initial times functions at order A are simply

$$\tau_\pm^{(0)}(S) = S \mp U_{\rm m}^{-1}. \tag{3.10}$$

To order A, the invariant functions along the characteristics are

$$\Gamma_\pm^{(0)}(X,T) = G_{\rm n}(V_{\rm b})^{\frac{1}{2}}V^{(0)}(X,T) \pm I^{(0)}(X,T) \tag{3.11}$$

and the condition (2.26) is

$$G_{\rm n}(V_{\rm b})^{\frac{1}{2}}V_{\rm I}^{(0)}\left(S \mp U_{\rm m}^{-1}\right) \pm I_{\rm I}^{(0)}\left(S \mp U_{\rm m}^{-1}\right) = G_{\rm n}(V_{\rm b})^{\frac{1}{2}}V_{\rm III}^{(0)}(S) \pm I_{\rm III}^{(0)}(S). \tag{3.12}$$

Notice that (3.8) and (3.12) apply to the following linear wave equation:

$$\frac{\partial}{\partial X}\begin{pmatrix} I^{(0)} \\ V^{(0)}\end{pmatrix} = -U^{-1}\begin{pmatrix} 0 & G(V_{\rm b}) \\ 1 & 0\end{pmatrix}\frac{\partial}{\partial T}\begin{pmatrix} I^{(0)} \\ V^{(0)}\end{pmatrix}. \tag{3.13}$$

The function $G(V_{\rm b})$ is piecewise constant, so the speed of propagation in region II depends on bias voltage.

With a continuous sinusoidal input, (3.1), the solution of (3.13) after a transient has the form of an incident and reflected wavetrain in region I and a transmitted wavetrain in region III. In region I, as $T \to \infty$, let

$$V^{(0)}(X,T) = \frac{1}{2}\left(\exp\left(\mathrm{i}\Omega\left(T - \frac{X}{U_{\rm I}}\right)\right) + \mathcal{R}^{(0)}(\Omega)\exp\left(\mathrm{i}\Omega\left(T + \frac{X}{U_{\rm I}}\right)\right) + \text{c.c.}\right)$$

(3.14) and

$$I^{(0)}(X,T) = \left(\frac{L_{\rm II}C_{\rm I}}{L_{\rm I}C_{\rm II}}\right)^{\frac{1}{2}}\frac{1}{2}\left(\exp\left(\mathrm{i}\Omega\left(T - \frac{X}{U_{\rm I}}\right)\right) - \mathcal{R}^{(0)}(\Omega)\exp\left(\mathrm{i}\Omega\left(T + \frac{X}{U_{\rm I}}\right)\right) + \text{c.c.}\right),$$

where c.c. indicates complex conjugate and we have used the freedom to set the origin of T. The normalized speed of propagation in region I is

$$U_{\rm I} = \frac{1}{c}\left(L_{\rm I}C_{\rm I}\right)^{-\frac{1}{2}}, \tag{3.15}$$

and the normalized signal frequency is

$$\Omega = 2\pi f\frac{l}{c}. \tag{3.16}$$

Similarly, in region III, let

$$V^{(0)}(X,T) = \frac{1}{2}\left(\mathcal{T}^{(0)}(\Omega)\exp\left(\mathrm{i}\Omega\left(T - \frac{(X-1)}{U_{\rm I}}\right)\right) + \text{c.c.}\right)$$

(3.17) and

$$I^{(0)}(X,T) = \left(\frac{L_{\rm II}C_{\rm I}}{L_{\rm I}C_{\rm II}}\right)^{\frac{1}{2}}\frac{1}{2}\left(\mathcal{T}^{(0)}(\Omega)\exp\left(\mathrm{i}\Omega\Big(T - \frac{(X-1)}{U_{\rm I}}\Big)\right) + \text{c.c.}\right).$$

Notice that the phase at $X = 1$ is absorbed into $\mathcal{T}^{(0)}(\Omega)$.

The lowest-order transmission and reflection coefficients $\mathcal{T}^{(0)}(\Omega)$ and $\mathcal{R}^{(0)}(\Omega)$ are found by imposing the condition (3.12). Using (3.10), (3.14), and (3.17) gives explicit expressions for the functions $V_{\rm I}^{(0)}(T)$, $I_{\rm I}^{(0)}(T)$, $V_{\rm III}^{(0)}(T)$, and $I_{\rm III}^{(0)}(T)$ introduced in (2.24) and (2.25). Thus (3.12) is

$$1 + \mathcal{R}^{(0)}(\Omega) \pm \beta^{\frac{1}{2}}\left(1 - \mathcal{R}^{(0)}(\Omega)\right) = \mathrm{e}^{\pm\mathrm{i}\Omega U_{\rm m}^{-1}}\mathcal{T}^{(0)}(\Omega)\left(1 \pm \beta^{\frac{1}{2}}\right), \tag{3.18}$$

where

$$\beta = \frac{L_{\rm II}C_{\rm I}}{L_{\rm I}C_{\rm m}(v_{\rm b})} = \frac{L_{\rm II}C_{\rm I}}{L_{\rm I}C_{\rm II}}\frac{1}{G_{\rm n}(V_{\rm b})}. \tag{3.19}$$

In the experimental system $\beta \ll 1$ because the impedance in region II is higher than that in regions I and III.

The pair of equations (3.18) can be solved to yield

$$\mathcal{R}^{(0)}(\Omega) = -\frac{(1-\beta)\sin(\frac{\Omega}{U_{\rm m}})}{(1+\beta)\sin(\frac{\Omega}{U_{\rm m}}) - 2\mathrm{i}\beta^{\frac{1}{2}}\cos(\frac{\Omega}{U_{\rm m}})}$$

$$= -\frac{(1-\beta^2)\sin^2(\frac{\Omega}{U_\mathrm{m}}) - 2\mathrm{i}\beta^{\frac{1}{2}}(1-\beta)\sin(\frac{\Omega}{U_\mathrm{m}})\cos(\frac{\Omega}{U_\mathrm{m}})}{(1+\beta)^2\sin^2(\frac{\Omega}{U_\mathrm{m}}) + 4\beta\cos^2(\frac{\Omega}{U_\mathrm{m}})}, \tag{3.20}$$

$$\mathcal{T}^{(0)}(\Omega) = -\frac{2\mathrm{i}\beta^{\frac{1}{2}}}{(1+\beta)\sin(\frac{\Omega}{U_\mathrm{m}}) - 2\mathrm{i}\beta^{\frac{1}{2}}\cos(\frac{\Omega}{U_\mathrm{m}})}$$

$$= -\frac{4\beta\cos(\frac{\Omega}{U_\mathrm{m}}) + 2\mathrm{i}\beta^{\frac{1}{2}}(1+\beta)\sin(\frac{\Omega}{U_\mathrm{m}})}{(1+\beta)^2\sin^2(\frac{\Omega}{U_\mathrm{m}}) + 4\beta\cos^2(\frac{\Omega}{U_\mathrm{m}})}\,. \tag{3.21}$$

These coefficients, also found in the quantum-mechanical problem of a potential step [17], satisfy the energy constraint $|\mathcal{R}^{(0)}(\Omega)|^2 + |\mathcal{T}^{(0)}(\Omega)|^2 = 1$. Perfect transmission, i.e., $|\mathcal{T}^{(0)}(\Omega)| = 1$, is found at resonant frequencies Ω satisfying

$$\sin\left(\frac{\Omega}{U_\mathrm{m}}\right) = 0. \tag{3.22}$$

3.2. Next order. In order to impose the equality (3.7) to order A^2 we write

$$\begin{aligned} &AV_\mathrm{I}^{(0)}(\tau_\pm^{(0)}(S) + A\tau_\pm^{(1)}(S)) \\ &= AV_\mathrm{I}^{(0)}(S \pm U_\mathrm{m}^{-1}) + A^2V_\mathrm{I}'(S \pm U_\mathrm{m}^{-1})\tau_\pm^{(1)}(S) + \mathcal{O}(A^3), \end{aligned} \tag{3.23}$$

where $V_\mathrm{I}'(T) = \frac{\mathrm{d}}{\mathrm{d}T}V_\mathrm{I}^{(0)}(T)$. The current $I_\mathrm{I}^{(0)}(\tau_\pm^{(0)}(S) + A\tau_\pm^{(1)}(S))$ is expanded in the same way. To find the solution at order A^2, we therefore need an explicit form for the initial times function to order A. The relationship that gives the initial times for the two characteristics that cross at $(1, S)$ is

$$\begin{aligned} S &= \tau_\pm(S) \pm \frac{1}{U}\int_0^1 (G_\mathrm{n}(V(X, T_\pm(X)))^{\frac{1}{2}}\,\mathrm{d}X \\ &= \tau_\pm^{(0)}(S) + A\tau_\pm^{(1)}(S) \\ &\quad \pm \frac{1}{U_\mathrm{m}}\int_0^1 \left(1 + \frac{1}{2}\frac{G_\mathrm{n}'(V_\mathrm{b})}{G_\mathrm{n}(V_\mathrm{b})}AV^{(0)}(X, T_\pm^{(0)}(X))\right)\mathrm{d}X + \cdots. \end{aligned} \tag{3.24}$$

Using (3.10) yields

$$\tau_\pm^{(1)}(S) = \mp\frac{1}{2}U_\mathrm{m}^{-1}\frac{G_\mathrm{n}'(V_\mathrm{b})}{G_\mathrm{n}(V_\mathrm{b})}\int_0^1 V^{(0)}(X, T_\pm^{(0)}(X))\mathrm{d}X. \tag{3.25}$$

In Appendix A, we construct the explicit form of $V^{(0)}(X, T_\pm^{(0)})$ and perform the integral in (3.25), leading to the expansion (A.8) of (3.7) to order A^2.

The only sinusoidal terms on the right-hand side of (A.8) have frequency 2Ω, so there is a transmitted second-harmonic signal with amplitude proportional to the square of the input amplitude [11]. We therefore let the nonconstant terms of order A^2 in the expansion (3.4) be

$$V_\mathrm{I}^{(1)}(T) = \frac{1}{2}\left[\mathcal{R}^{(1)}(\Omega)\,\mathrm{e}^{2\,\mathrm{i}\Omega T} + \mathrm{c.c.}\right], \quad I_\mathrm{I}^{(1)}(T) = -\left(\frac{L_\mathrm{II}C_\mathrm{I}}{L_\mathrm{I}C_\mathrm{II}}\right)^{\frac{1}{2}}\frac{1}{2}\left[\mathcal{R}^{(1)}(\Omega)\,\mathrm{e}^{2\,\mathrm{i}\Omega T} + \mathrm{c.c.}\right],$$

$$V_\mathrm{III}^{(1)}(T) = \frac{1}{2}\left[\mathcal{T}^{(1)}(\Omega)\,\mathrm{e}^{2\,\mathrm{i}\Omega T} + \mathrm{c.c.}\right], \quad I_\mathrm{III}^{(1)}(T) = \left(\frac{L_\mathrm{II}C_\mathrm{I}}{L_\mathrm{I}C_\mathrm{II}}\right)^{\frac{1}{2}}\frac{1}{2}\left[\mathcal{T}^{(1)}(\Omega)\,\mathrm{e}^{2\,\mathrm{i}\Omega T} + \mathrm{c.c.}\right].$$

Then (A.8) becomes

$$
\begin{aligned}
&(1\pm\beta^{\frac12})\mathcal{T}^{(1)}(\Omega)-(1\mp\beta^{\frac12})\mathcal{R}^{(1)}(\Omega)\mathrm{e}^{\mp 2\mathrm{i}\Omega U_{\mathrm{m}}^{-1}}\\
&=\frac18\,\frac{G_{\mathrm{n}}'(V_{\mathrm{b}})}{G_{\mathrm{n}}(V_{\mathrm{b}})}\mathcal{T}^{(0)}(\Omega)^2\Bigg[\left(\cos\frac{\Omega}{U_{\mathrm{m}}}-\mathrm{i}\beta^{\frac12}\sin\frac{\Omega}{U_{\mathrm{m}}}\right)^2\mathrm{e}^{\mp 2\mathrm{i}\Omega U_{\mathrm{m}}^{-1}}-1\\
&\mp\ \mathrm{i}\left((1\pm\beta^{\frac12})^2\frac{\Omega}{U_{\mathrm{m}}}+(1-\beta)\frac12\left(\sin\left(2\frac{\Omega}{U_{\mathrm{m}}}\right)\pm\mathrm{i}\left(1-\cos\left(2\frac{\Omega}{U_{\mathrm{m}}}\right)\right)\right)\right)\Bigg].
\end{aligned}
\tag{3.26}
$$

Explicit expressions for $\mathcal{T}^{(1)}(\Omega)$ and $\mathcal{R}^{(1)}(\Omega)$ can now be obtained by inverting the pair of complex equations (3.26). The result simplifies considerably at *resonant frequencies*, (3.22), when (3.26) reduces to

$$
(1\pm\beta^{\frac12})\mathcal{T}^{(1)}(\Omega)-(1\mp\beta^{\frac12})\mathcal{R}^{(1)}(\Omega)=\mp\,\mathrm{i}(1\pm\beta^{\frac12})^2\frac18\frac{G_{\mathrm{n}}'(V_{\mathrm{b}})}{G_{\mathrm{n}}(V_{\mathrm{b}})}\frac{\Omega}{U_{\mathrm{m}}}.
\tag{3.27}
$$

The solution of (3.27) is

$$
\mathcal{T}^{(1)}(\Omega)=-\frac{1}{16}\,\frac{\mathrm{i}}{\beta^{\frac12}}\frac{G_{\mathrm{n}}'(V_{\mathrm{b}})}{G_{\mathrm{n}}(V_{\mathrm{b}})}\frac{\Omega}{U_{\mathrm{m}}}(1+3\beta),
\tag{3.28}
$$

$$
\mathcal{R}^{(1)}(\Omega)=-\frac{1}{16}\,\frac{\mathrm{i}}{\beta^{\frac12}}\frac{G_{\mathrm{n}}'(V_{\mathrm{b}})}{G_{\mathrm{n}}(V_{\mathrm{b}})}\frac{\Omega}{U_{\mathrm{m}}}(1-\beta).
\tag{3.29}
$$

Note that $G'(0)=0$ and $G'(V)<0$ when $V\neq 0$. A relatively simple form is also found in the limit $\frac{\Omega}{U_{\mathrm{m}}}\gg 1$, where

$$
\begin{aligned}
&(1\pm\beta^{\frac12})\mathcal{T}^{(1)}(\Omega)-(1\mp\beta^{\frac12})\mathcal{R}^{(1)}(\Omega)\mathrm{e}^{\mp 2\mathrm{i}\Omega U_{\mathrm{m}}^{-1}}\\
&\qquad=\mp\mathcal{T}^{(0)}(\Omega)^2\mathrm{i}(1\pm\beta^{\frac12})^2\frac18\frac{G_{\mathrm{n}}'(V_{\mathrm{b}})}{G_{\mathrm{n}}(V_{\mathrm{b}})}\frac{\Omega}{U_{\mathrm{m}}}\,.
\end{aligned}
\tag{3.30}
$$

The solution of the system (3.30) is

$$
\begin{aligned}
\mathcal{T}^{(1)}(\Omega)&=-\frac{\mathrm{i}}{8}\,\mathcal{T}^{(0)}(\Omega)^2\frac{G_{\mathrm{n}}'(V_{\mathrm{b}})}{G_{\mathrm{n}}(V_{\mathrm{b}})}\frac{\Omega}{U_{\mathrm{m}}}\frac{(1+3\beta)\cos 2\frac{\Omega}{U_{\mathrm{m}}}+\mathrm{i}(3\beta^{\frac12}+\beta^{\frac32})\sin 2\frac{\Omega}{U_{\mathrm{m}}}}{2\beta^{\frac12}\cos 2\frac{\Omega}{U_{\mathrm{m}}}+\mathrm{i}(1+\beta)\sin 2\frac{\Omega}{U_{\mathrm{m}}}}\ ,\\
\mathcal{R}^{(1)}(\Omega)&=-\frac{\mathrm{i}}{8}\,\mathcal{T}^{(0)}(\Omega)^2\frac{G_{\mathrm{n}}'(V_{\mathrm{b}})}{G_{\mathrm{n}}(V_{\mathrm{b}})}\frac{\Omega}{U_{\mathrm{m}}}\frac{1-\beta}{2\beta^{\frac12}\cos 2\frac{\Omega}{U_{\mathrm{m}}}+\mathrm{i}(1+\beta)\sin 2\frac{\Omega}{U_{\mathrm{m}}}}\ .
\end{aligned}
\tag{3.31}
$$

At the nth resonance, $\frac{\Omega}{U_{\mathrm{m}}}=n\pi$, so (3.31) is a rather good approximation in typical experimental situations where $n\geq 4$.

The factors on the right-hand side of (3.31) have simple interpretations. The factor $\mathcal{T}^{(0)}(\Omega)^2$ means that the amplitude of the second harmonic is maximized when the frequency of the input signal is at resonance. The factor $G_{\mathrm{n}}'(V_{\mathrm{b}})/G_{\mathrm{n}}(V_{\mathrm{b}})$ gives the explicit dependence on bias voltage; because $G_{\mathrm{n}}'(0)=0$ there is no second-harmonic generation at order A^2 for zero bias. The factor $\Omega=2\pi fl/c$ means that the second-harmonic amplitude is proportional to the "effective length" of the waveguide, the number of wavelengths over which the nonlinearity has time to act. The final factor involving transcendental functions is complicated. However, it can be seen that the

dependence on $2\Omega/U_{\rm m}$ will produce secondary maxima in the second-harmonic amplitude at frequencies halfway between resonances of the input signal. Finally, we remark that the constant terms (zero-frequency "DC-component") of amplitude A^2, also generated by the quadratic component of nonlinearity, can always be absorbed into the bias voltage, and hence their effects are already accounted for by V_b-variations.

4. Comparison with numerical solutions. The response of a finite-length nonlinear device was modeled by dividing the space into three one-dimensional regions. In the central region (region II) the wave equation, two coupled PDEs for the charge and current fields (2.12), is nonlinear. The numerical runs described in this section used values of the capacitances and inductances, length of region II, and input frequencies, chosen to be similar to those found in the experiments described in section 5 below [23].

In the interior of each region, timestepping is performed according to the Lax–Wendroff method described in section B.1. At the boundaries between region I and II and between regions II and III, we impose continuity of voltage and current. In section B.2 we describe how this is made consistent with the different relations between charge and voltage functions that hold to the left and to the right of the boundary. In the numerics, we can examine the whole configuration of current and charge at any time. An example numerical configuration (a"snapshot" at one instant of time) is shown in Figure 4.1.

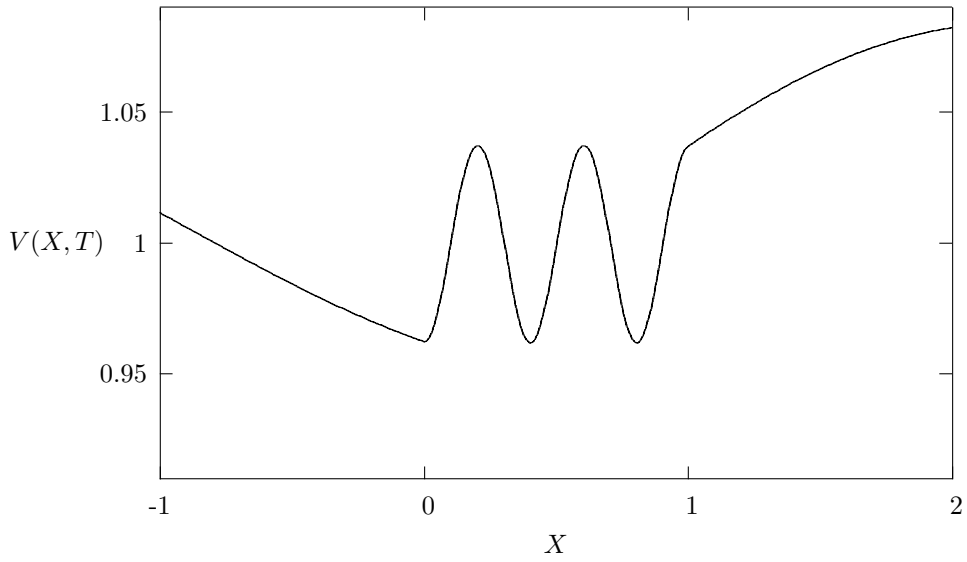


FIG. 4.1. *Numerical voltage field as a function of* X *for* $T = 1000$. *In regions* I *and* III *(*$X < 0$ *and* $X > 1$*), the wave equation is linear; in region* II *(*$0 \le X \le 1$*), the wave equation is nonlinear. A sinusoidal signal was input at* $f = 250$MHz *with* $a = 0.1$V. *The parameters are* $L_{\rm I}C_{\rm I}c^2 = 9$, $L_{\rm II}C_{\rm II}c^2 = 2500$, *and* $l = 6$cm, $v_{\rm b} = 1$V, *and* $v^* = 10$V. *The partial differential equations* (2.12) *were solved with* $\Delta X = 0.001$ *and* $\Delta T = 0.001$.

Regions I and III, semi-infinite in the PDE (2.12), are of finite length in the numerical scheme. In region I, which stretches from $X = -L$ to $X = 0$, there is a prescribed input signal; the update employed at the leftmost extremity for producing it is derived in section B.3. Fourier transforms of numerical time series at points in

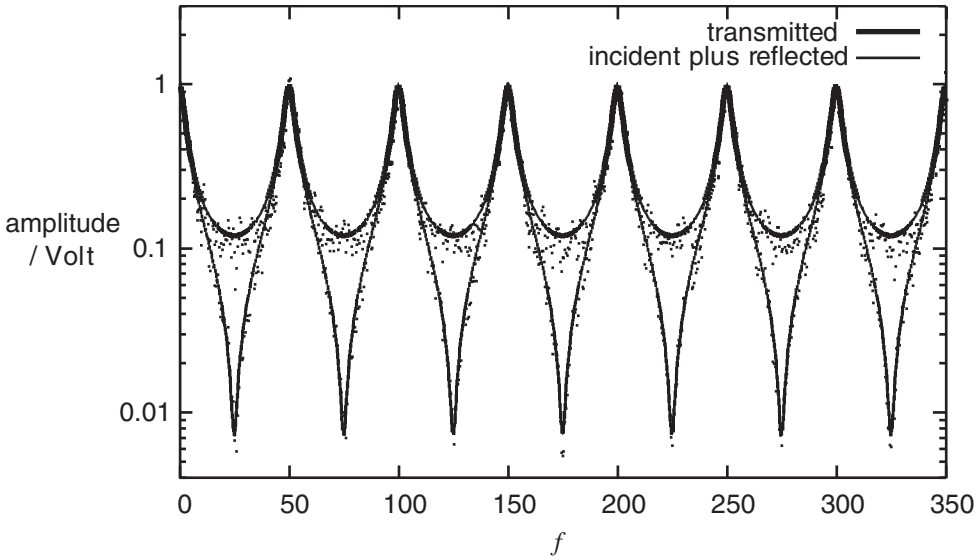


FIG. 4.2. *The solid lines are modulus of the transmission coefficient* $\mathcal{T}^{(0)}(\Omega)$ *and of* $1+\mathcal{R}^{(0)}(\Omega)$ *as a function of frequency at* (3.21). *The dots are the Fourier transforms of numerical time series with white noise input.*

regions I and III are used to evaluate the transmitted and reflected signals. Sinusoidal input signals may be accompanied by noise, either white or with constant power in a window of frequencies. The latter "band-limited" noise is generated either by filtering white noise or by explicitly constructing a signal as a sum of Fourier components with constant power in the frequency window.

Our analytical and numerical methods are not altered by changes in the form of the function $C(V)$, assuming that it is a positive even function with maximum at 0. In the numerical runs producing the results displayed in Figures 4.1–4.5, we used the following form:

$$C_{\rm m}(v) = C_{\rm II}\left(1+\left(\frac{v}{v^*}\right)^2\right)^{-1}, \tag{4.1}$$

giving the convenient explicit form $V(x,t) = \tan(Q(x,t))$ in region II and

$$G_{\rm n}(V) = \frac{1}{1+V^2}. \tag{4.2}$$

The parameter values used in the numerical runs reported in this section were as follows. The length of region II was taken to be $l = 6$cm and the characteristic voltage $v^* = 10$V. The L and C constants in the regions were given by $L_{\rm I}C_{\rm I}c^2 = 9$, $L_{\rm II}C_{\rm II}c^2 = 2500$. Thus the speed of light in regions I and III was one third of that in a vacuum, c. The corresponding speed in region II and the factor β are bias-dependent; at zero bias the speed was $c/50$ and $\beta = 0.0036$.

The results of one numerical experiment are shown in Figure 4.2. The input was broadband white noise with small amplitude, so that all frequencies have equal average amplitude in the input. The solid lines are the predictions obtained from (3.21). Assuming a flat input spectrum, we obtain the transmission coefficient $\mathcal{T}^{(0)}(\Omega)$ as the Fourier transform of the time series in region III, showing resonances when

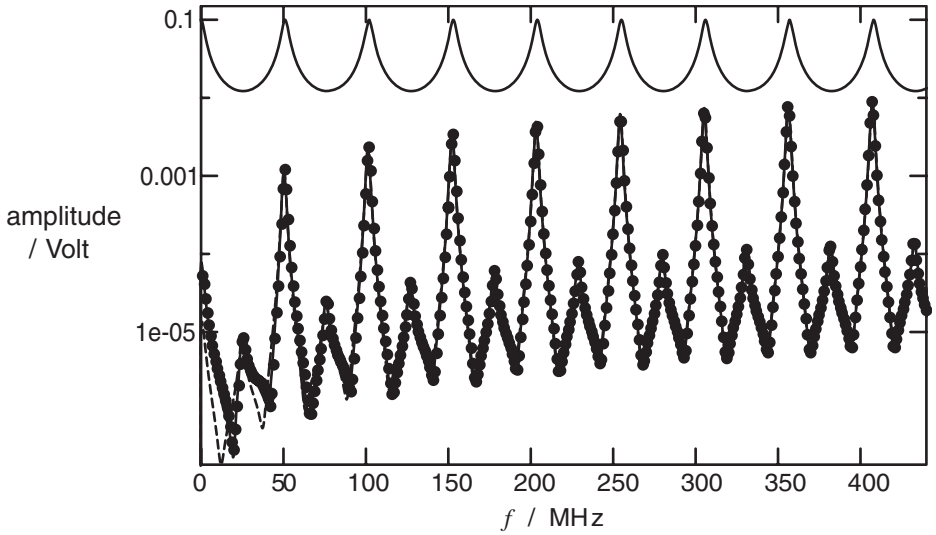


FIG. 4.3. *Transmission and second-harmonic generation as a function of input frequency. The transmitted amplitudes at f and $2f$ are shown on a logarithmic scale. Numerical results are shown as solid circles and analytical results as lines. In particular, the dashed line corresponds to the high frequency limit* (3.31)*, and, as seen, it becomes accurate past frequencies of about* 50MHz.

Ω/U_m was a multiple of π. (With $l = 6$cm, resonant frequencies are at multiples of 50MHz.) The numerical results, shown as dots, fluctuate about the theoretical line as expected. In region I, the time-dependent voltage is the sum of incident and reflected signals; the Fourier amplitude in this case, shown in the figure, is $1 + \mathcal{R}^{(0)}(\Omega)$. At resonant frequencies (3.22), there is perfect transmission and $\mathcal{R}^{(0)}(\Omega) = 0$. Halfway between resonant frequencies, at frequencies satisfying $\sin\left(\frac{\Omega}{U_{\rm m}}\right) = \pm 1$, we find minimum transmission

$$|\mathcal{T}^{(0)}(\Omega)|^2_{\rm min} = \frac{4\beta}{(1+\beta)^2} \tag{4.3}$$

and

$$\mathcal{R}^{(0)}(\Omega) = -\frac{1-\beta}{1+\beta} = -1 + 2\beta + \mathcal{O}(\beta^2); \tag{4.4}$$

at these frequencies $\mathcal{R}^{(0)}(\Omega) \simeq -1$, and the reflected and incident signals almost cancel each other out. This feature is clearly visible in the numerical results. No nonlinear effects are displayed in Figure 4.2, but the run serves as a stringent check on the numerical algorithm over several orders of magnitude in voltage.

The effect of the nonlinearity in generating an output at twice the input frequency is shown in Figure 4.3, which summarizes the results of a series of numerical runs, each with a single sinusoidal input. The input amplitude $a = 0.1$V was fixed, and all runs were performed with bias voltage $v_{\rm b} = 2.01$; the frequency f of the input signal was changed from run to run. At the chosen bias voltage, resonant frequencies are multiples of 51MHz. The figure shows the transmitted amplitude at frequencies f and $2f$, along with the analytical predictions. The amplitude at f is given by (3.21). The prediction at $2f$, obtained by inverting (3.26), is remarkably accurate over many orders of magnitude in the transmitted amplitude. The asymptotic formula for high

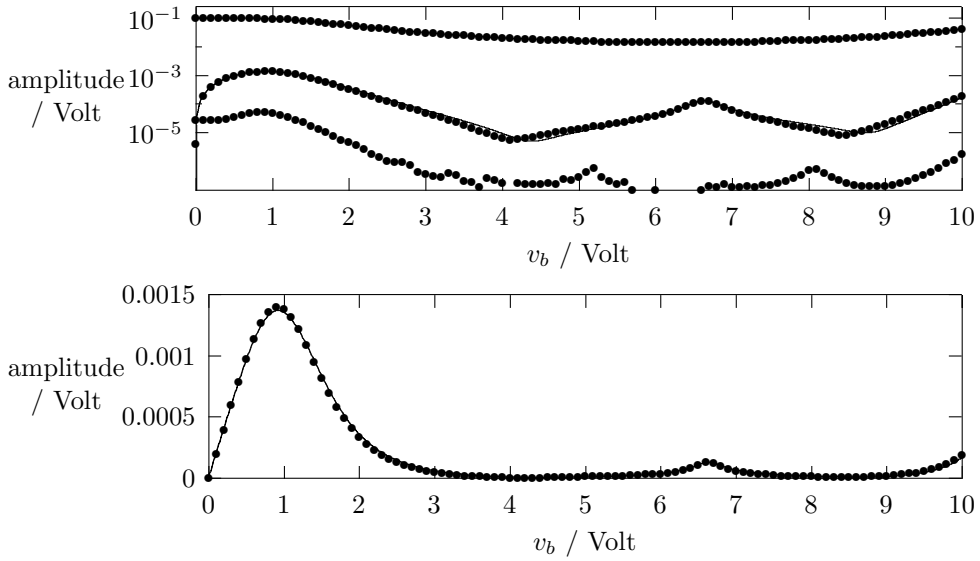


FIG. 4.4. *Transmission and harmonic generation as a function of bias voltage. In the upper graph the output amplitudes at f, $2f$, and $3f$ are shown on a logarithmic scale. In the lower graph the amplitude at $2f$ is shown on a linear scale. Numerical results are shown as solid circles. The solid lines use* (3.21) *and* (3.26). *The dotted line, using* (3.31), *differs visibly from* (3.26) *only at low frequencies. In each case $f = 150$MHz and $a = 0.1$V.*

frequency (3.31), shown in this figure as a dashed line, is indistinguishable from the full inversion of (3.26) and the numerical data past frequencies of about 50MHz.

The features of the second-harmonic amplitude predicted by (3.31) are clearly evident in the numerical results. The amplitude of the second harmonic is maximized when the frequency of the input signal is at resonance. Local maxima of the second-harmonic amplitude are also found at frequencies halfway between resonances of the input signal. The height of the maxima coinciding with the nth resonance is proportional to n because the second-harmonic amplitude is proportional to Ω.

A similar series of runs was used to produce Figure 4.4, this time with all parameters except bias voltage held fixed. The input frequency is chosen to be at resonance at zero bias. The top graph shows the transmitted amplitude at f, $2f$, and $3f$ on a logarithmic scale. (We have not attempted to calculate the $3f$ amplitude but conjecture that it is proportional to $C''(V_\mathrm{b})$, and hence nonzero at $V_\mathrm{b} = 0$.) In the lower graph we plot the transmitted second-harmonic ($2f$) amplitude and the approximation (3.31) on a linear scale. At zero bias voltage, it is zero at order A^2 because $C'(0) = 0$. The second-harmonic amplitude at first increases with bias voltage, and then decreases as the bias voltage is further increased, as the resonant frequency is shifted away from the input frequency. For sufficiently large bias voltage, the shift in resonant frequencies is such that the input frequency can once again be at resonance.

The convenient analytical forms (3.31), for the amplitudes and phases of the transmitted and reflected second harmonic at order A^2, give several intriguing possibilities if the input signal consists of two or more sinusoids. For example, in addition to a sinusoidal input at frequency f, another at frequency $2f$ can be chosen so that the total transmission at $2f$ is zero to order A^2 by cancellation of the second harmonic produced by the nonlinearity with the transmitted signal at $2f$. This cancellation

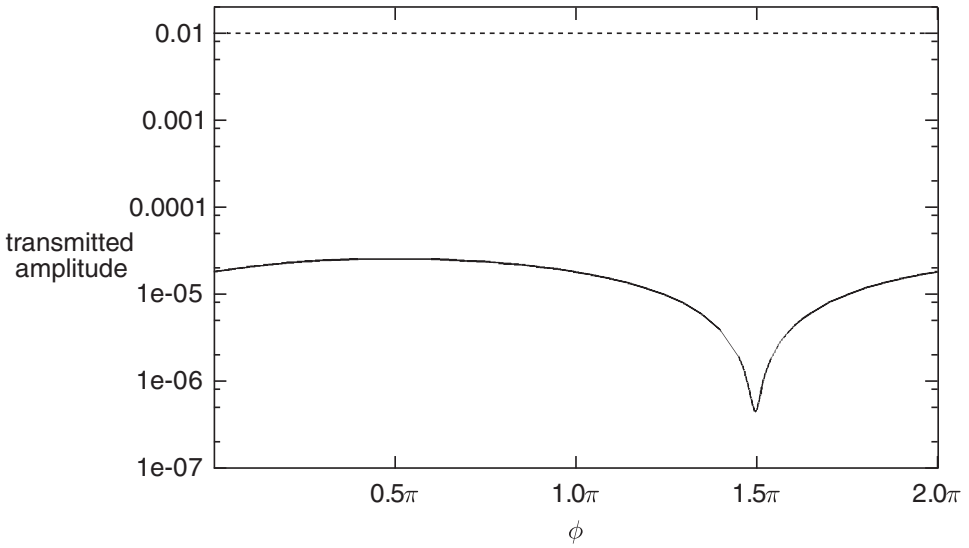


FIG. 4.5. *Transmitted amplitude at f (dashed) and $2f$ (solid) versus phase ϕ, logarithmic scale. The input signal is given by* (4.5)*, with the amplitude b chosen to be equal to that of the second harmonic generated by the nonlinearity from the input at frequency f with amplitude a. The parameters are $f = 51$MHz and $a = 0.01$.*

can be produced for any input frequency and will be explored further below in the case where f is a resonant frequency. Alternatively, for an input frequency slightly off resonance, it is possible to choose a combination such that the *reflected* signal is free of harmonics to order A^2.

In Figure 4.5 we show the transmitted amplitude at f and $2f$, as a function of the phase ϕ, when the input signal is given by

$$v_{\text{in}} = a\cos(2\pi f t) + b\cos(4\pi f t + \phi), \tag{4.5}$$

where f is chosen to be the lowest resonant frequency at a nonzero bias voltage. At lowest order, the transmitted signal is equal to the input signal, since $2f$ is also a resonant frequency. There is another source of transmitted power at $2f$, the generated second-harmonic signal with amplitude, given by (3.28), proportional to a^2/v^*. We choose b in (4.5) to be equal to that amplitude. (With the parameters in Figure 4.5, $|\mathcal{T}^{(1)}(\Omega)| \simeq 1.3$, so $b = |\mathcal{T}^{(1)}(\Omega)|a^2/v^* \simeq 0.000013$.) According to (3.28), the phase of the generated $2f$ signal is $\frac{1}{2}\pi$. Thus when $\phi = -\frac{1}{2}\pi$, the generated and directly transmitted outputs have equal amplitude and opposite phase; the result is to eliminate the total transmitted $2f$ signal to order A^2. (Also shown in Figure 4.5, as a dashed line, is the transmitted amplitude at f, which is independent of ϕ and indistinguishable from the input amplitude.)

The procedure of producing a pure transmitted signal with frequency f from a mixed input can in principle be carried to nth order, using an input signal with amplitudes and phases chosen to eliminate n multiples of f from the transmitted signal. The result would be a unique multifrequency input that gives a pure single frequency transmitted signal through a given device. This uniqueness property, and its dependence on the external bias field, might have encoding implications in signal transmission protocols.

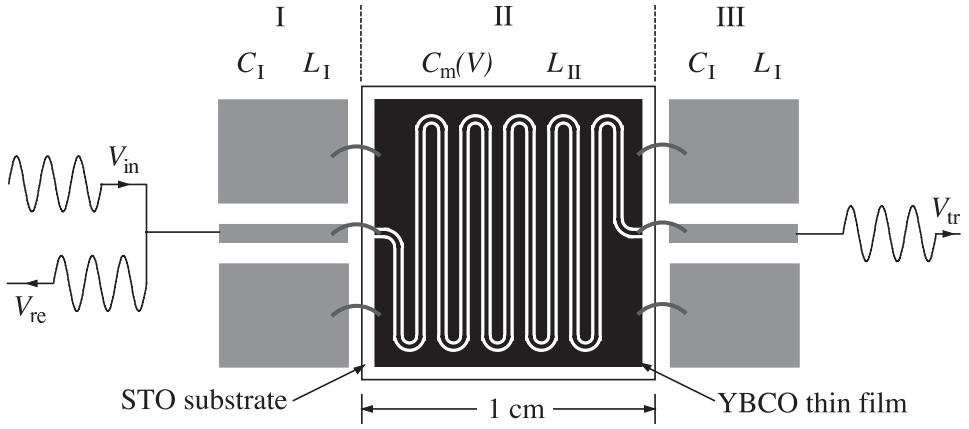


FIG. 5.1. *Device top view and schematic diagram. The input circuitry, device, and output circuitry are modeled as regions* I, II, *and* III. *The time-dependent voltage in region* I *is a sum of the incident and reflected sinusoidal signals,* $V_{\rm in}$ *and* $V_{\rm re}$. *In the central region, region* II, *the wave equation is nonlinear. Signals travel along the meandering waveguide patterned into a superconducting thin film. The transmitted signal in region* III, $V_{\rm tr}$, *is measured and analyzed.*

5. Comparison with an experimental system. In this section we report on experimental results. Microwave signals and noise were passed through a small device whose properties depend on operating temperature and applied voltage. A schematic diagram is shown in Figure 5.1. The devices are 1cm × 1cm × 0.5mm, consisting of 0.4μm-thick superconducting $YBa_2Cu_3O_7$ (YBCO) film on single-crystal substrates [22]. They are manufactured with two parallel meandering gaps (width 15μm) patterned into the superconducting thin film. The result is a waveguide with a narrow meandering centerline (length $l \simeq$ 8cm and width 20μm) and two groundplanes (the rest of the superconducting film).

A constant voltage difference, the "bias" voltage, is maintained between the centerline and the groundplane. Because the lateral dimensions of the waveguide are much smaller than the wavelengths of the input signals, wave propagation is effectively one-dimensional. In the superconducting state, below the transition temperature $T_{\rm c} \simeq 85K$, resistive losses of the YBCO film are negligible. Working temperatures were in the range 20K–60K. The source of the nonlinearity in the experiment is the substrate that the superconducting waveguide rests on, a single crystal of strontium titanate, $SrTiO_3$ (STO), with large permittivity. Due to the nonlinear dielectric properties of STO [19, 20], the shunt capacitance per unit length of the waveguide depends on temperature and on voltage. The resulting voltage-dependent differential capacitance produces behavior that combines resonance effects with harmonic generation and frequency mixing. More details can be found in [3], [10], and [11].

Our quantitative studies begin with a measurement of the differential capacitance as a function of temperature and voltage. We assume that the inductance per unit length, $L_{\rm II}$, is unaffected by changes in temperature and voltage. (The value $L_{\rm II} = 505.7$pH/mm is deduced from measurements of the capacitance and resonant frequencies of the devices [11].) Measurement of the differential capacitance is conveniently carried out using a small-amplitude broadband noise input at various temperatures and bias voltages. The dominant feature of the transmitted spectrum is a series of maxima at resonant frequencies. The resonant frequencies at bias voltage $v_{\rm b}$

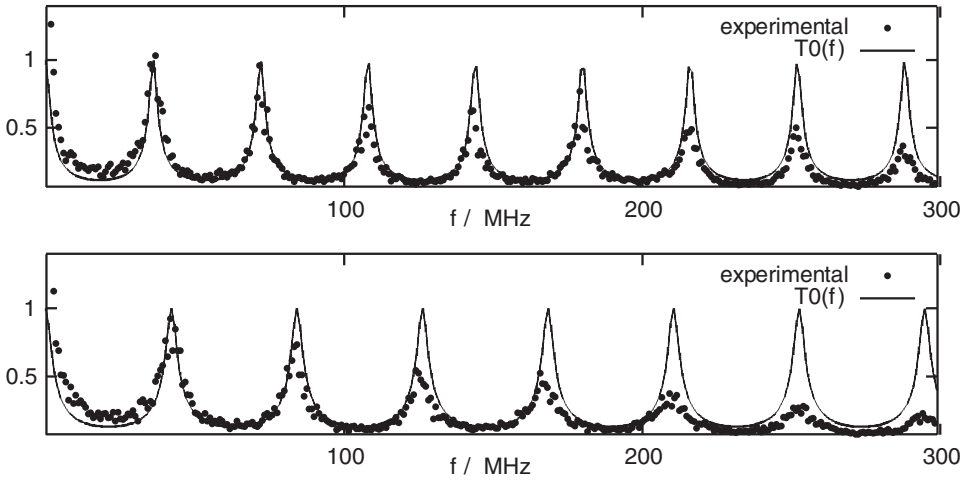


FIG. 5.2. *Experimental and theoretical results: Transmission versus frequency. The experimental results were obtained at operating temperature* 40K. *In the upper graph, bias voltage* $= 0$. *In the lower graph, bias voltage* $= 10$V. *The solid lines are the theoretical curve* (3.21) *for linear lossless transmission. The parameters are* $L_{\rm I}C_{\rm I}/L_{\rm II}C_{\rm II} = 9/2500$ *and* $l = 7.8$cm. *The positions of the resonance peaks are well predicted, but successive experimental peaks are of reduced height due to losses in the device.*

are $f_n(v_{\rm b})$, where

$$f_n(v_{\rm b}) = \frac{n}{2}\frac{u_{\rm m}(v_{\rm b})}{l}, \qquad n = 0, 1, 2, \ldots, \tag{5.1}$$

and the propagation speed for small-amplitude signals, $u_{\rm m}(v_{\rm b})$, is given by (2.14): $u_{\rm m}(v_{\rm b}) = (L_{\rm II}C_{\rm m}(v_{\rm b}))^{-\frac{1}{2}}$.

Experimental results for the transmitted signal as a function of input frequency are shown in Figure 5.2. The input in each case was broadband white noise and the operating temperature was 40K. Because $C_{\rm m}(v_{\rm b})$ is a decreasing function on $v_{\rm b}$, the distance between resonant peaks increases when a bias voltage is applied. The theoretical curve accurately gives the position of the resonant peaks in transmission but does not predict their decreasing height as a function of frequency. The reason is that the model PDEs (2.12) do not include losses whose effect is to reduce transmission by a factor that increases with frequency.

The decreasing height of resonant peaks does not prevent us from accurately determining their position. Using (2.14) determines the differential capacitance as a function of voltage; $C_{\rm m}(v_{\rm b})$ is related to the frequency f_n of the nth peak in the transmission versus frequency curve by

$$L_{\rm II}C_{\rm m}(v_{\rm b}) = \left(\frac{n}{2}\frac{1}{l}\right)^2 f_n^{-2}(v_{\rm b}). \tag{5.2}$$

The upper plot in Figure 5.3 is the position of the third resonance as a function of bias voltage at operating temperature 40K. The function used to obtain the fit shown as a solid line is

$$f(v_{\rm b}) = f(0) + c_1 v_{\rm b}\tanh(c_2 v_{\rm b}), \tag{5.3}$$

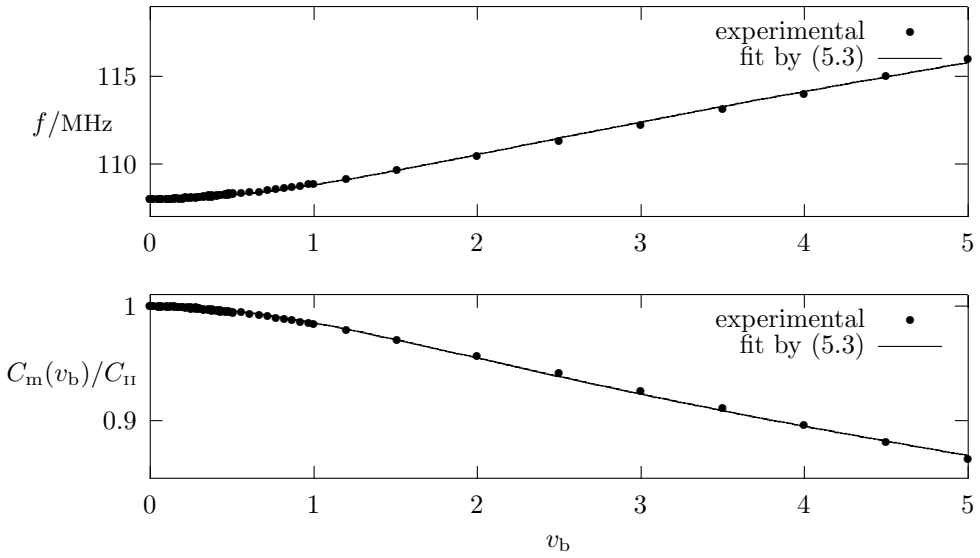


FIG. 5.3. *Experimental results at operating temperature* 40K *and best fit using* (5.3). *Upper graph: Frequency at the* $n = 3$ *peak as a function of bias voltage. Lower graph: Differential capacitance as a function of bias voltage divided by* $C_{\rm II} = C_{\rm m}(0)$.

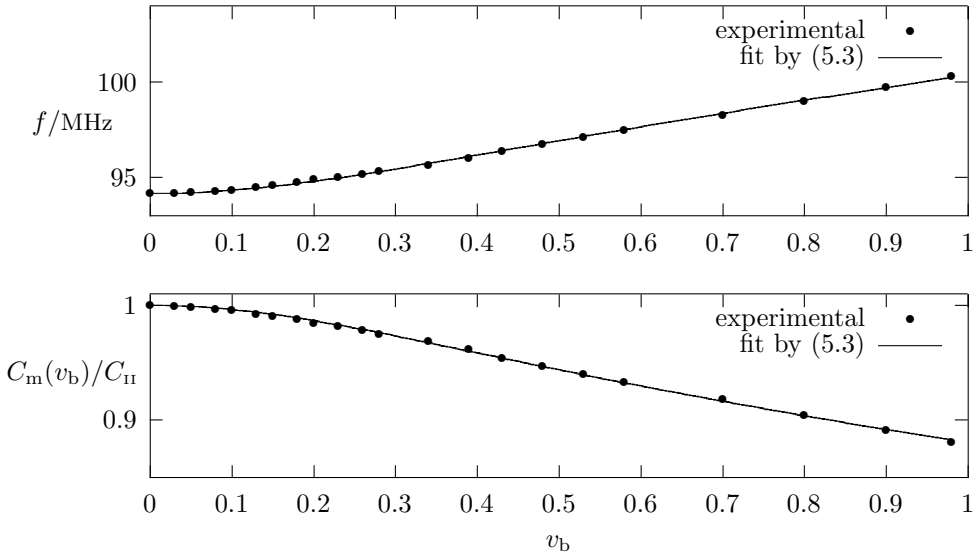


FIG. 5.4. *Experimental results at operating temperature* 20K *and best fit using* (5.3). *Upper graph: Frequency at the* $n = 4$ *peak as a function of bias voltage. Lower graph: Differential capacitance as a function of bias voltage divided by* $C_{\rm II} = C_{\rm m}(0)$.

where the two parameters c_1 and c_2 are adjusted once to obtain a fit through all the experimental points. The lower plot in Figure 5.3 is the ratio $C_{\rm m}(v_{\rm b})/C_{\rm II}$, deduced from the upper curve using (5.2). In Figure 5.4, similar results are shown for operating temperature 20K. At this lower temperature, differential capacitance is a more rapidly varying function of bias voltage. As the graphs in Figure 5.4 show, the functional

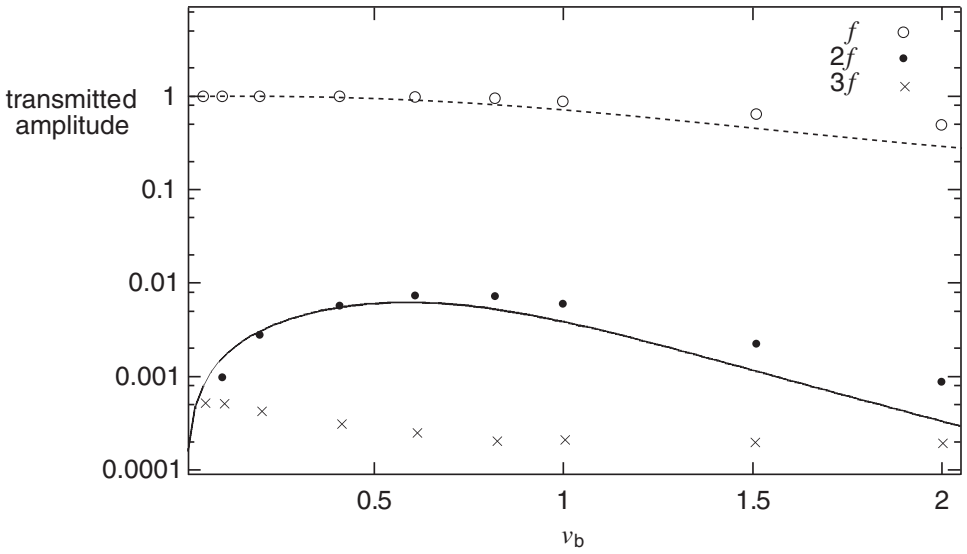


FIG. 5.5. *Experimental results and theoretical curves: Transmitted amplitudes at f (empty circles), $2f$ (filled circles), and $3f$ (crosses) versus bias. The input is of fixed amplitude and frequency $f = 215$MHz. The lines are the theoretical curves using asymptotic analysis of the lossless nonlinear PDE* (2.12). *The dashed line is the transmission coefficient from the linear equation. The solid line shows the order A^2 second-harmonic amplitude* (3.31).

form (5.3) appears quite suitable for the purpose of modeling the medium's nonlinear response within the range of voltage amplitudes we have considered.

Figure 5.5 summarizes experimental data taken at 40K using an input signal of fixed amplitude at $f = 215$MHz. The frequency is chosen to be at the sixth resonance peak for zero bias voltage. The upper part of the plot shows the transmitted amplitude at f, normalized to its value at zero bias. As bias voltage increases away from zero, there is a shift in the resonance frequency (compare Figure 5.2) so that the input frequency is no longer a resonant frequency and the transmitted power decreases. The decrease is less rapid in the experiment (empty circles) than in the theoretical curve (dotted line) because losses have the effect of widening (as well as reducing the height of) resonant peaks. Dielectric losses are also known to decrease with bias voltage [22, 24]. The solid line and filled circles show the transmitted amplitude at $2f$ as a function of bias. The reason for the initial rapid increase with bias is that the amplitude of the second-harmonic signal is proportional to $C'_{\rm m}(v_{\rm b})$ and $C'_{\rm m}(0) = 0$. The subsequent decline is due to the decrease of the transmitted power at frequency f (compare (3.31)). Also shown is the third-harmonic power, which has a nonzero value at zero bias voltage and then decreases slowly [11].

6. Conclusion. A few points emerge from our side-by-side modeling and experimental investigation of electromagnetic propagation in a class of nonlinear waveguides. First, our simple voltage-current model seems to capture most of the physical features of the experimental device, including the tunable nonlinearity via an externally applied electric field. We find quantitative agreement with most measurable quantities. Some effects we neglect, such as dissipation, do show up in the data, but at least for the experimental regimes we have focused on, they can be considered secondary.

Second, transmission of sinusoidal signals through a linear medium is maximized

when the length of the region is equal to a multiple of half the wavelength. This remains the dominant feature in transmission through the nonlinear region we studied, because the signal amplitude is small enough and the nonlinear region is short enough that nonlinear and dispersive effects have a subdominant role. The main effect of the nonlinearity is to produce a second-harmonic output, with amplitude that is well predicted by an expansion that assumes that the input signal is sufficiently weak. The explicit and, especially near resonance, relatively simple formulae we have derived using an asymptotic approach perform well against both numerical solutions of our model equation and experimental results, thus giving us confidence that the theoretical building blocks we have established can be used in other studies of this class of materials.

Third, the numerical simulations in which we mix the input signal with a second harmonic show that it is possible to achieve transparency (or total reflection) at second order in the input amplitude. In principle the expansion we have set up could be carried out to higher orders, thus eliminating all higher harmonics from the transmitted signal. Such transparency, and its dependence on the external bias field, might be exploited for technological applications such as signal transmission protocols.

The main difference between our system and optical systems [25] is the absence of frequency dispersion. Future work will apply and extend the tools we have developed in this paper to study arrays of similar devices, where the longer effective length will permit nonlinear pulses of permanent form (or "solitons") because of the dispersive effects that such configurations generate. Distributed (or spatially extended) nonlinear arrays have intrinsic broadband capability (at least several hundred GHz) in contrast to electronic circuits based on discrete elements and high-dimensional complexity, governed by PDEs or coupled ODEs, in contrast to low-dimensional systems whose dynamics are governed by ODEs. Thus, distributed nonlinear arrays have the potential to handle wider data bandwidth at a high level of security in communication.

Appendix A. The initial times function.

In this appendix we find the explicit lowest-order solution for the voltage in region II, evaluate the integral (3.25) that gives the initial times function, and explicitly expand (3.7) to order A^2.

Consider a point (T, X) in region II. It is the intersection of the two characteristics that intersect $X = 0$ at $(T_\pm(0), 0)$, where

$$T_\pm^{(0)}(0) = T \mp \frac{X}{U_{\rm m}}\,. \tag{A.1}$$

Now the solution is constructed using

$$V^{(0)}(X,T) = (2G_{\rm n}(V_{\rm b})^{\frac{1}{2}})^{-1}\left(\Gamma_+^{(0)}(X,T) + \Gamma_-^{(0)}(X,T)\right). \tag{A.2}$$

Using the property of the characteristics, $\Gamma_\pm(X,T) = \Gamma_\pm(0, T_\pm(0))$, and the explicit forms,

$$V_{\rm I}^{(0)}(T) = \frac{1}{2}\left((1+\mathcal{R}^{(0)}(\Omega))\mathrm{e}^{\mathrm{i}\Omega T} + \text{c.c.}\right),$$

$$I_{\rm I}^{(0)}(T) = \left(\frac{L_{\rm II}C_{\rm I}}{L_{\rm I}C_{\rm II}}\right)^{\frac{1}{2}}\frac{1}{2}\left((1-\mathcal{R}^{(0)}(\Omega))\mathrm{e}^{\mathrm{i}\Omega T} + \text{c.c.}\right),$$

$$V_{\rm III}^{(0)}(T) = \frac{1}{2}\left(\mathcal{T}^{(0)}(\Omega)\mathrm{e}^{\mathrm{i}\Omega T} + \text{c.c.}\right),$$

$$I^{(0)}_{\rm III}(T) = \left(\frac{L_{\rm II}C_{\rm I}}{L_{\rm I}C_{\rm II}}\right)^{\frac{1}{2}} \frac{1}{2}\left(\mathcal{T}^{(0)}(\Omega)\mathrm{e}^{\mathrm{i}\Omega T} + \text{c.c.}\right), \tag{A.3}$$

we obtain

$$\begin{aligned}
V^{(0)}(X,T) &= G_{\rm n}(V_{\rm b})^{-\frac{1}{2}}\frac{1}{2}\Bigg[G_{\rm n}(V_{\rm b})^{\frac{1}{2}}V_{\rm I}^{(0)}\left(T-\frac{X}{U_{\rm m}}\right) + I_{\rm I}^{(0)}\left(T-\frac{X}{U_{\rm m}}\right) \\
&\qquad + G_{\rm n}(V_{\rm b})^{\frac{1}{2}}V_{\rm I}^{(0)}\left(T+\frac{X}{U_{\rm m}}\right) - I_{\rm I}^{(0)}\left(T+\frac{X}{U_{\rm m}}\right)\Bigg] \\
&= \frac{1}{2}\Bigg[\exp\left(\mathrm{i}\Omega\left(T-\frac{X}{U_{\rm m}}\right)\right) + \mathcal{R}^{(0)}(\Omega)\exp\left(\mathrm{i}\Omega\left(T-\frac{X}{U_{\rm m}}\right)\right) \\
&\quad +\beta^{\frac{1}{2}}\left(\exp\left(\mathrm{i}\Omega\left(T-\frac{X}{U_{\rm m}}\right)\right) - \mathcal{R}^{(0)}(\Omega)\exp\left(\mathrm{i}\Omega\left(T-\frac{X}{U_{\rm m}}\right)\right)\right) \\
&\quad + \exp\left(\mathrm{i}\Omega\left(T+\frac{X}{U_{\rm m}}\right)\right) + \mathcal{R}^{(0)}(\Omega)\exp\left(\mathrm{i}\Omega\left(T+\frac{X}{U_{\rm m}}\right)\right) \\
&\quad -\beta^{\frac{1}{2}}\left(\exp\left(\mathrm{i}\Omega\left(T+\frac{X}{U_{\rm m}}\right)\right) - \mathcal{R}^{(0)}(\Omega)\exp\left(\mathrm{i}\Omega\left(T+\frac{X}{U_{\rm m}}\right)\right)\right) + \text{c.c.}\Bigg] \\
&= \frac{1}{2}\Bigg[\mathrm{e}^{\mathrm{i}\Omega T}\Bigg((1+\mathcal{R}^{(0)}(\Omega))\cos\left(\frac{\Omega}{U_{\rm m}}X\right) \\
&\quad - \mathrm{i}\beta^{\frac{1}{2}}(1-\mathcal{R}^{(0)}(\Omega))\sin\left(\frac{\Omega}{U_{\rm m}}X\right)\Bigg) + \text{c.c.}\Bigg].
\end{aligned}$$

Thus

$$\begin{aligned}
V^{(0)}(X,T) &= |1+\mathcal{R}^{(0)}(\Omega)|\cos\left(\frac{\Omega}{U_{\rm m}}X\right)\cos(\Omega T+\phi_1) \\
&\quad - \beta^{\frac{1}{2}}|1-\mathcal{R}^{(0)}(\Omega)|\sin\left(\frac{\Omega}{U_{\rm m}}X\right)\sin(\Omega T+\phi_2)\,,
\end{aligned} \tag{A.4}$$

where $\phi_1 = \arg(1+\mathcal{R}^{(0)}(\Omega))$ and $\phi_2 = \arg(1-\mathcal{R}^{(0)}(\Omega))$. In the case of resonance, $\Omega = U_{\rm m}\, n\pi$, $\mathcal{R}^{(0)}(\Omega) = 0$, and the solution in region II is simply

$$V(X,T) = V_{\rm b} + A\left(\cos(n\pi X)\cos(\Omega T) - \beta^{\frac{1}{2}}\sin(n\pi X)\sin(\Omega T)\right). \tag{A.5}$$

To evaluate the integral in (3.25), we need the quantity $V^{(0)}(X, T^{(0)}_{\pm}(X))$. Inserting

$$T^{(0)}_{\pm}(X) = S \mp \frac{1-X}{U_{\rm m}} \tag{A.6}$$

and using (3.18) gives

$$V^{(0)}(X,T^{(0)}_{\pm}(X)) = \frac{1}{4}\Bigg[\mathrm{e}^{\mathrm{i}\Omega(S\mp U_{\rm m}^{-1})}\Bigg(1+\mathcal{R}^{(0)}(\Omega) \pm \beta^{\frac{1}{2}}(1-\mathcal{R}^{(0)}(\Omega))$$

$$+\left(\cos\left(2\frac{\Omega}{U_{\rm m}}X\right)\pm\ {\rm i}\,\sin\left(2\frac{\Omega}{U_{\rm m}}X\right)\right)\left(1+\mathcal{R}^{(0)}(\Omega)\mp\beta^{\frac{1}{2}}(1-\mathcal{R}^{(0)}(\Omega))\right)\Bigg)+{\rm c.c.}\Bigg]$$

$$=\frac{1}{4}\left[{\rm e}^{{\rm i}\Omega S}\mathcal{T}^{(0)}(\Omega)\Big(1\pm\beta^{\frac{1}{2}}+(1\mp\beta^{\frac{1}{2}}){\rm e}^{\mp 2{\rm i}\frac{\Omega}{U_{\rm m}}}\Big(\cos\left(2\frac{\Omega}{U_{\rm m}}X\right)\pm{\rm i}\sin\left(2\frac{\Omega}{U_{\rm m}}X\right)\Big)\Big)+{\rm c.c.}\right].$$

Thus, the next-to-lowest order term in the initial times function is explicitly

$$\tau_{\pm}^{(1)}(S)=\mp U_{\rm m}^{-1}\frac{1}{8}\frac{G_{\rm n}'(V_{\rm b})}{G_{\rm n}(V_{\rm b})}\Bigg[{\rm e}^{{\rm i}\Omega S}\mathcal{T}^{(0)}(\Omega)\Bigg(1\pm\beta^{\frac{1}{2}}$$

$$+\frac{1}{2}\frac{U_{\rm m}}{\Omega}(1\mp\beta^{\frac{1}{2}}){\rm e}^{\mp 2{\rm i}\frac{\Omega}{U_{\rm m}}}\left(\sin\left(2\frac{\Omega}{U_{\rm m}}\right)\pm{\rm i}\mp{\rm i}\cos\left(2\frac{\Omega}{U_{\rm m}}\right)\right)\Bigg)+{\rm c.c.}\Bigg]. \tag{A.7}$$

The full expression for (3.7), up to order A^2 with (3.14) and (3.17), is

$$\begin{aligned}
&AG_{\rm n}(V_{\rm b})^{\frac{1}{2}}\left(1\pm\beta^{\frac{1}{2}}\right)\frac{1}{2}\left(\mathcal{T}^{(0)}(\Omega){\rm e}^{{\rm i}\Omega S}+{\rm c.c.}\right)\\
&+\ A^2G_{\rm n}(V_{\rm b})^{\frac{1}{2}}V_{\rm III}^{(1)}(S)\pm A^2I_{\rm III}^{(1)}(S)\\
&+\ A^2G_{\rm n}(V_{\rm b})^{\frac{1}{2}}\frac{1}{8}\frac{G_{\rm n}'(V_{\rm b})}{G_{\rm n}(V_{\rm b})}\left(|\mathcal{T}^{(0)}(\Omega)|^2+\frac{1}{2}(\mathcal{T}^{(0)}(\Omega)^2{\rm e}^{2{\rm i}\Omega S}+{\rm c.c.})\right)\\
&=AG_{\rm n}(V_{\rm b})^{\frac{1}{2}}\frac{1}{2}\left[{\rm e}^{{\rm i}\Omega(S\mp U_{\rm m}^{-1})}\left(1+\mathcal{R}^{(0)}(\Omega)\pm\beta^{\frac{1}{2}}(1-\mathcal{R}^{(0)}(\Omega))\right)+{\rm c.c.}\right]\\
&+\ A^2G_{\rm n}(V_{\rm b})^{\frac{1}{2}}V_{\rm I}^{(1)}(S\mp U_{\rm m}^{-1})\pm A^2I_{\rm I}^{(1)}(S\mp U_{\rm m}^{-1})\\
&+\ A^2G_{\rm n}(V_{\rm b})^{\frac{1}{2}}\frac{1}{8}\frac{G_{\rm n}'(V_{\rm b})}{G_{\rm n}(V_{\rm b})}\left(|1+\mathcal{R}^{(0)}(\Omega)|^2+\frac{1}{2}\left[(1+\mathcal{R}^{(0)}(\Omega))^2\,{\rm e}^{2{\rm i}\Omega(S\mp U_{\rm m}^{-1})}+{\rm c.c.}\right]\right)\\
&+\ A^2G_{\rm n}(V_{\rm b})^{\frac{1}{2}}\frac{1}{2}\Omega\left[{\rm i}{\rm e}^{{\rm i}\Omega(S\mp U_{\rm m}^{-1})}\left(1+\mathcal{R}^{(0)}(\Omega)\pm\beta^{\frac{1}{2}}(1-\mathcal{R}^{(0)}(\Omega))\right)+{\rm c.c.}\right]\tau_{\pm}^{(1)}(S).
\end{aligned}$$

To order A we regain (3.18). With (A.7), we obtain the following expression for (3.7) at order A^2:

$$\begin{aligned}
&V_{\rm III}^{(1)}(S)\pm G_{\rm n}(V_{\rm b})^{-\frac{1}{2}}I_{\rm III}^{(1)}(S)-V_{\rm I}^{(1)}(S\mp U_{\rm m}^{-1})\mp G_{\rm n}(V_{\rm b})^{-\frac{1}{2}}I_{\rm I}^{(1)}(S\mp U_{\rm m}^{-1})\\
=&\frac{1}{16}\frac{G_{\rm n}'(V_{\rm b})}{G_{\rm n}(V_{\rm b})}\Bigg[\Bigg(-|\mathcal{T}^{(0)}(\Omega)|^2-\mathcal{T}^{(0)}(\Omega)^2\,{\rm e}^{2{\rm i}\Omega S}+|\mathcal{T}^{(0)}(\Omega)|^2\left(\cos^2\frac{\Omega}{U_{\rm m}}+\beta\sin^2\frac{\Omega}{U_{\rm m}}\right)\\
&+\mathcal{T}^{(0)}(\Omega)^2\left(\cos^2\frac{\Omega}{U_{\rm m}}-{\rm i}\beta^{\frac{1}{2}}\sin^2\frac{\Omega}{U_{\rm m}}\right)^2{\rm e}^{2{\rm i}\Omega(S\mp U_{\rm m}^{-1})}+{\rm c.c.}\Bigg)\\
&\mp\frac{\Omega}{U_{\rm m}}(1\pm\beta^{\frac{1}{2}})\Bigg(\Bigg(1\pm\beta^{\frac{1}{2}}+\frac{1}{2}\frac{U_{\rm m}}{\Omega}\Big(\sin\left(2\frac{\Omega}{U_{\rm m}}\right)\\
&\pm{\rm i}\left(1-\cos\left(2\frac{\Omega}{U_{\rm m}}\right)\right)\Big)\Bigg){\rm i}\mathcal{T}^{(0)}(\Omega)^2{\rm e}^{2{\rm i}\Omega S}+{\rm c.c.}\Bigg)\Bigg].
\end{aligned} \tag{A.8}$$

Appendix B. Numerical techniques.

B.1. Lax–Wendroff finite-difference method. The fields Q and I are updated on a grid with spacing ΔX. In the first-order Lax–Wendroff method, the values at $T + \Delta T$ are obtained from those at T using [18]:

$$\begin{aligned} Q(X, T+\Delta T) = &\frac{1}{2}\left(Q(X+\Delta X, T) + Q(X, T)\right) \\ &-\frac{1}{2}U\frac{\Delta T}{\Delta X}\left(I(X+\Delta X, T) - I(X-\Delta X, T)\right), \end{aligned} \tag{B.1}$$

$$\begin{aligned} I(X, T+\Delta T) = &\frac{1}{2}\left(I(X+\Delta X, T) + I(X, T)\right) \\ &-\frac{1}{2}U\frac{\Delta T}{\Delta X}\left(\mathcal{V}(Q(X+\Delta X, T)) - \mathcal{V}(Q(X-\Delta X, T))\right). \end{aligned} \tag{B.2}$$

We use the following second-order adaptation [18]:

$$\begin{aligned} Q(X, T+\Delta T) = &Q(X, T) - \frac{1}{2}U\frac{\Delta T}{\Delta X}\left(I(X+\Delta X, T) - I(X-\Delta X, T)\right) \\ &+\frac{1}{2}\left(U\frac{\Delta T}{\Delta X}\right)^2\left(\mathcal{V}(Q(X+\Delta X, T)) + \mathcal{V}(Q(X-\Delta X, T))\right), \end{aligned} \tag{B.3}$$

$$\begin{aligned} I(X, T+\Delta T) = I(X, T) - U\frac{\Delta T}{\Delta X}\Bigg[&\mathcal{V}\Bigg(\frac{1}{2}(Q(X+\Delta X, T) + Q(X, T)) \\ &-\frac{1}{2}U\frac{\Delta T}{\Delta X}(I(X+\Delta X, T) - I(X, T))\Bigg) \\ -&\mathcal{V}\Bigg(\frac{1}{2}(Q(X, T) + Q(X-\Delta X, T)) \\ &-\frac{1}{2}U\frac{\Delta T}{\Delta X}(I(X, T) - I(X-\Delta X, T))\Bigg)\Bigg]. \end{aligned} \tag{B.4}$$

B.2. Discretized boundary conditions. At the interface between different media it is important to implement numerical boundary conditions that are of the same order of the scheme and do not produce artificial reflections [26]. Consider the boundary between regions I and II. Let $Q_1(n\Delta X, T)$ and $I_1(n\Delta X, T)$ be the values of the last point in region I; let $Q_2(0, T)$ and $I_2(0, T)$ be the values of the first point in region II. We cannot immediately implement the Lax–Wendroff scheme because the neighboring points to the boundary points are in different regions, where the equation of motion takes different forms: $\mathcal{V}(Q) = V_{\rm l}(Q)$ and $\mathcal{V}(Q) = V_{\rm n}(Q)$.

Because $Q_1(n\Delta x, T)$ and $Q_2(0, T)$ correspond to the same position,

$$I_1(n\Delta x, T) = I_2(0, T) \qquad \text{and} \quad V_{\rm l}(Q_1(n\Delta x, T)) = V_{\rm n}(Q_2(0, T)), \tag{B.5}$$

with $V_{\rm l}$ and $V_{\rm n}$ as defined in (2.11). As (B.5) is true for any T, we also have the equalities

$$\frac{\partial}{\partial T} I_1(n\Delta x, T) = \frac{\partial}{\partial T} I_2(0, T) \tag{B.6}$$

and

$$\frac{\partial}{\partial T}V_{\rm l}(Q_1(n\Delta x,T)) = \frac{\partial}{\partial T}V_{\rm n}(Q_2(0,T)). \tag{B.7}$$

Therefore, from the equation of motion (2.12),

$$\frac{\partial}{\partial X}V_{\rm l}(Q_1(n\Delta x,T)) = \frac{\partial}{\partial X}V_{\rm n}(Q_2(0,T)) \tag{B.8}$$

and

$$V_{\rm l}'(Q_1(n\Delta x,T))\frac{\partial}{\partial X}I_1(n\Delta x,T) = V_{\rm n}'(Q_2(0,T))\frac{\partial}{\partial X}I_2(0,T), \tag{B.9}$$

where

$$V_{\rm l}'(Q) = \frac{\partial}{\partial Q}V_{\rm l}(Q), \qquad V_{\rm n}'(Q) = \frac{\partial}{\partial Q}V_{\rm n}(Q). \tag{B.10}$$

The discrete approximation to the partial derivatives on the right-hand side of (B.9) is

$$\frac{\partial}{\partial X}I_1(n\Delta x,T) = (2\Delta x)^{-1}\left(3I_1(n\Delta x) - 4I_1((n-1)\Delta x) + I_1((n-2)\Delta x)\right), \tag{B.11}$$

$$\frac{\partial}{\partial X}I_2(0,T) = (2\Delta x)^{-1}\left(-3I_2(0) + 4I_2(\Delta x) - I_2(2\Delta x)\right); \tag{B.12}$$

the derivatives in (B.8) are evaluated similarly. The discretized version of (B.8)–(B.9) is therefore

$$\begin{aligned} V_{\rm l}(Q_1(n\Delta x,T)) &= V_{\rm n}(Q_2(0,T)) \\ &= \frac{1}{6}\Big(4V_{\rm l}(Q_1((n-1)\Delta x,T)) - V_{\rm l}(Q_1((n-2)\Delta x,T)) \\ &\quad + 4V_{\rm n}(Q_2(\Delta x,T)) - V_{\rm n}(Q_2(2\Delta x,T))\Big), \end{aligned}$$

$$\begin{aligned} I_1(n\Delta x,T) &= I_2(0,T) \\ &= \frac{1}{3}\left(1 + \frac{V_{\rm n}'(Q_2(0,T))}{V_{\rm l}'(Q_1(n\Delta x,T))}\right)^{-1}\left(4I_1((n-1)\Delta x) - I_1((n-2)\Delta x)\right. \\ &\quad \left. + \frac{V_{\rm n}'(Q_2(0,T))}{V_{\rm l}'(Q_1(n\Delta x,T))}(4I_2(\Delta x) - I_2(2\Delta x))\right). \end{aligned}$$

B.3. Input signal. At the left-hand extremity of region I, $X = -L$, the solution of the field equation (2.12) can be written as a superposition of two fields: a prescribed incident wave $Q_{\rm in}(X - cT)$, $I_{\rm in}(X - cT)$, and an (unknown) outgoing wave $Q_{\rm out}(X + cT)$, $I_{\rm out}(X + cT)$ ($c > 0$). Thus, the incoming field satisfies the unidirectional wave equation

$$\frac{\partial}{\partial T}\begin{pmatrix} Q_{\rm in} \\ I_{\rm in} \end{pmatrix} = -c\frac{\partial}{\partial X}\begin{pmatrix} Q_{\rm in} \\ I_{\rm in} \end{pmatrix}, \tag{B.13}$$

while the outgoing field satisfies

$$\frac{\partial}{\partial T}\begin{pmatrix} Q_{\text{out}} \\ I_{\text{out}} \end{pmatrix} = +c\frac{\partial}{\partial X}\begin{pmatrix} Q_{\text{out}} \\ I_{\text{out}} \end{pmatrix}. \tag{B.14}$$

Hence, for the total field $Q(X,T)$, $I(X,T)$,

$$\begin{aligned}\frac{\partial}{\partial T}\begin{pmatrix} Q \\ I \end{pmatrix} &= \frac{\partial}{\partial T}\begin{pmatrix} Q_{\text{in}} \\ I_{\text{in}} \end{pmatrix} + \frac{\partial}{\partial T}\begin{pmatrix} Q_{\text{out}} \\ I_{\text{out}} \end{pmatrix} \\ &= \frac{\partial}{\partial T}\begin{pmatrix} Q_{\text{in}} \\ I_{\text{in}} \end{pmatrix} + c\frac{\partial}{\partial X}\begin{pmatrix} Q_{\text{out}} \\ I_{\text{out}} \end{pmatrix} \\ &= \frac{\partial}{\partial T}\begin{pmatrix} Q_{\text{in}} \\ I_{\text{in}} \end{pmatrix} + c\frac{\partial}{\partial X}\begin{pmatrix} Q - Q_{\text{in}} \\ I - I_{\text{in}} \end{pmatrix}. \end{aligned} \tag{B.15}$$

Using (B.13), the governing equation of the total field in the left linear region can then be rewritten as

$$\frac{\partial}{\partial T}\begin{pmatrix} Q \\ I \end{pmatrix} = 2\frac{\partial}{\partial T}\begin{pmatrix} Q_{\text{in}} \\ I_{\text{in}} \end{pmatrix} + c\frac{\partial}{\partial X}\begin{pmatrix} Q \\ I \end{pmatrix}; \tag{B.16}$$

this is used with one-sided spatial derivatives to construct the field at $X = -L$ from the knowledge of the field at the previous time step.

For example, consider a sinusoidal input signal, as in (3.1). Then, using (3.2),

$$V_{\text{in}}(X,T) = A\cos(\Omega(T - X/U_{\text{I}})). \tag{B.17}$$

The incident Q and I fields are given by

$$Q_{\text{in}}(X,T) = \frac{C_{\text{I}}}{C_{\text{II}}}A\cos\left(\Omega\left(T - \frac{X}{U_{\text{I}}}\right)\right), \tag{B.18}$$

$$I_{\text{in}}(X,T) = \left(\frac{L_{\text{II}}C_{\text{I}}}{L_{\text{I}}C_{\text{II}}}\right)^{\frac{1}{2}} A\cos\left(\Omega\left(T - \frac{X}{U_{\text{I}}}\right)\right). \tag{B.19}$$

Thus the (lowest-order) increments at the left extremity of region I are

$$\begin{aligned} Q_1(-L, T+\Delta T) = Q_1(-L,T) + \Delta T\Big(&2\frac{\mathrm{d}}{\mathrm{d}t}Q_{\text{in}}(-L,T) \\ &- U_{\text{I}}(2\Delta x)^{-1}\big(3Q_1(0) - 4Q_1(-L+\Delta x) + Q_1(-L+2\Delta x)\big)\Big), \\ I_1(-L, T+\Delta T) = I_1(-L,T) + \Delta T\Big(&2\frac{\mathrm{d}}{\mathrm{d}t}I_{\text{in}}(-L,T) \\ &- U_{\text{I}}(2\Delta x)^{-1}\big(3I_1(0) - 4I_1(-L+\Delta x) + I_1(-L+2\Delta x)\big)\Big). \end{aligned}$$

Acknowledgments. GL thanks Theoretical Division T7 and the Superconductivity Technology Center, Los Alamos National Laboratory, for hosting visits during which this work was completed.

REFERENCES

[1] R. E. Collins, *Foundations for Microwave Engineering*, McGraw–Hill, New York, 1992.

[2] V. K. Varadan, D. K. Ghodgaonkar, V. V. Varadan, J. F. Kelly, and P. Glikerdas, *Ceramic phase shifters for electronically steerable antenna systems*, Microwave J., 35 (1992), pp. 116–127.

[3] A. T. Findikoglu, Q. X. Jia, D. W. Reagor, and X. D. Wu, *Tunable microwave mixing in nonlinear dielectric thin films of $SrTiO_3$ and $Sr_{0.5}Ba_{0.5}TiO_3$*, Electron. Lett., 31 (1995), pp. 1814–1815.

[4] F. W. VanKeuls, R. R. Romanofsky, D. Y. Bohman, M. D. Winters, F. A. Miranda, C. H. Mueller, R. E. Treece, T. V. Rivkin, and D. Galt, *$YBa_2Cu_3O_{7-\delta}$,Au)/$SrTiO_3$/$LaAlO_3$ thin film conductor ferroelectric coupled microstripline phase shifters for phased array applications*, Appl. Phys. Lett., 71 (1997), pp. 3075–3077.

[5] G. Subramanyam, F. VanKeuls, and F. A. Miranda, *A K-band tunable microstrip bandpass filter using a thin-film conductor/ferroelectric/dielectric multilayer configuration*, IEEE Microwave Guided Wave Lett., 8 (1998), pp. 78–80.

[6] A. B. Kozyrev, T. B. Samoilova, A. A. Golovkov, E. K. Hollmann, D. A. Kalinikos, V. E. Loginov, A. M. Prudan, O. I. Soldatenkov, D. Galt, C. H. Mueller, T. V. Rivkin, and G. A. Koepf, *Nonlinear behavior of thin film $SrTiO_3$ capacitors at microwave frequencies,* J. Appl. Phys., 84 (1998), pp. 3326–3332.

[7] H. Fuke, Y. Terashima, H. Kayano, and H. Yoshino, *Electrically tunable $YBa_2Cu_3O_y$ resonators using interdigital electrodes and dielectric film*, Phys. C, 336 (2000), pp. 80–84.

[8] G. Subramanyam, F. W. VanKeuls, and F. A. Miranda, *Effect of DC biasing on YBCO/STO/LAO tunable microstrip filters*, Integ. Ferroelec., 29 (2000), pp. 81–93.

[9] G. Subramanyam, F. W. VanKeuls, F. A. Miranda, R. R. Romanofsky, and J. D. Warner, *Design and development of ferroelectric tunable HTS microstrip filters for Ku- and K-band applications*, Mat. Chem. Phys., 79 (2003), pp. 147–150.

[10] A. T. Findikoglu, R. Camassa, G. Lythe, and Q. X. Jia, *New potential applications of nonlinear dielectrics: Microwave solitons and stochastic resonance*, Integ. Ferroelec., 22 (1998), pp. 259–268.

[11] A. T. Findikoglu, R. Camassa, G. Lythe, and Q. X. Jia, *Dielectric nonlinearity and stochastic effects in strontium titanate*, Appl. Phys. Lett., 80 (2002), pp. 3391–3393.

[12] A. T. Findikoglu, Q. X. Jia, I. H. Campbell, X. D. Wu, D. W. Reagor, C. B. Mombourquette, and D. Murray, *Electrically tunable coplanar transmission line resonators using $YBa_2Cu_3O_{7-x}$/$SrTiO_3$ bilayers*, Appl. Phys. Lett., 66 (1995), pp. 3674–3676.

[13] M. J. Lancaster, J. Powell, and A. Porch, *Thin-film ferroelectric microwave devices*, Superconductor Science and Technology, 11 (1998), pp. 1323–1334.

[14] Q. X. Jia, A. T. Findikoglu, D. W. Reagor, and P. Lu, *Improvement in performance of electrically tunable devices based on nonlinear dielectric $SrTiO_3$ using a homoepitaxial $LaAlO_3$ interlayer*, Appl. Phys. Lett., 73 (1998), pp. 897–899.

[15] A. T. Findikoglu, Q. X. Jia, X. D. Wu, and D. W. Reagor, *Paraelectric thin films for microwave applications*, Integ. Ferroelec., 15 (1997), pp. 163–171.

[16] P. Prasad and R. Ravindran, *Partial Differential Equations*, Wiley Eastern, New Delhi, India, 1985.

[17] A. Messiah, *Quantum Mechanics*, North–Holland, Amsterdam, 1962.

[18] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes–the Art of Scientific Computing*, Cambridge University Press, Cambridge, UK, 1986.

[19] R. C. Neville, C. A. Mead, and B. Hoeneise, *Permittivity of strontium titanate*, J. Appl. Phys., 43 (1972), pp. 2124–2131.

[20] O. G. Vendik, *Dielectric nonlinearity of the displacive ferroelectrics at UHF*, Ferroelectrics, 12 (1976), pp. 85–90.

[21] M. J. Dalberth, R. E. Stauber, J. C. Price, C. T. Rogers, and D. Galt, *Improved low frequency and microwave dielectric response in strontium titanate thin films grown by pulsed laser ablation*, Appl. Phys. Lett., 72 (1998), pp. 507–509.

[22] A. T. Findikoglu, D. W. Reagor, K. O. Rasmussen, A. R. Bishop, N. Gronbech-Jensen, Q. X. Jia, Y. Fan, C. Kwon, and L. A. Ostrovsky, *Electrodynamic properties of coplanar waveguides made from high-temperature superconducting $YBa_2Cu3O_{7-\delta}$ electrodes on nonlinear dielectric $SrTiO_3$ substrates*, J. Appl. Phys., 86 (1999), pp. 1558–1568.

[23] The FORTRAN program used to produce the numerical results discussed in this section can be viewed at http://www.maths.leeds.ac.uk/Applied/stochastic/scan10.f and a sample input file at http://www.maths.leeds.ac.uk/Applied/stochastic/scan.in.

[24] O. G. Vendik, L. T. Ter-Martirosyan, and S. P. Zubk, *Microwave losses in incipient ferroelectrics as functions of the temperature and the biasing field*, J. Appl. Phys., 84 (1998), pp. 993–998.

[25] P. Mandel, *Theoretical Problems in Cavity Nonlinear Optics*, Cambridge University Press, Cambridge, UK, 1997.

[26] J. M. Hyman, *A method of lines approach to the numerical solutions of conservation laws*, in Advances in Computer Methods for Partial Differential Equations III, R. Vichnevetsky and R. S. Stepleman, eds., IMACS, New Brunswick, NJ, 1979, pp. 313–321.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 29–56

(SEMI)CLASSICAL LIMIT OF THE HARTREE EQUATION WITH HARMONIC POTENTIAL*

RÉMI CARLES†, NORBERT J. MAUSER‡, AND HANS PETER STIMMING§

Abstract. Nonlinear Schrödinger equations (NLS) of the Hartree type occur in the modeling of quantum semiconductor devices. Their "semiclassical" limit of vanishing (scaled) Planck constant is both a mathematical challenge and practically relevant when coupling quantum models to classical models. With the aim of describing the semiclassical limit of the three-dimensional (3D) Schrödinger–Poisson system with an additional harmonic potential, we study some semiclassical limits of the Hartree equation with harmonic potential in space dimension $n \geq 2$. The harmonic potential is confining and causes focusing periodically in time. We prove asymptotics in several cases, showing different possible nonlinear phenomena according to the interplay of the size of the initial data and the power of the Hartree potential. In the case of the 3D Schrödinger–Poisson system with harmonic potential, we can give only a formal computation since the need for modified scattering operators for this long-range scattering case goes beyond current theory.

We also deal with the case of an additional "local" nonlinearity given by a power of the local density—a model that is relevant when incorporating the Pauli principle in the simplest model given by the "Schrödinger–Poisson-Xα equation." Further we discuss the connection of our WKB-based analysis to the Wigner function approach to semiclassical limits.

Key words. Schrödinger–Poisson, Hartree equation, semiclassical limit, harmonic potential

AMS subject classifications. 35B33, 35B40, 35C20, 35Q40, 81Q20, 81S30

DOI. 10.1137/040609732

1. Introduction.

Nonlinear Schrödinger equations (NLS) are important both for many different applications and as a source of rich mathematical theory, with several hard challenges still open. The NLS in the most common meaning contains a "local" nonlinearity given by a power of the local density, in particular the (de)focusing "cubic" NLS which arises, e.g., in nonlinear optics or for Bose–Einstein condensates. In one dimension this NLS is an integrable system, and the "semiclassical limit" ("high wave number limit") can be performed by methods of inverse scattering (see, e.g., [20] and [22] for results on the defocusing and focusing cases). A class of NLS with a "nonlocal" nonlinearity that we call "Hartree type" occurs in the modeling of quantum semiconductor devices. Their "semiclassical" limit of vanishing (scaled) Planck constant is both a mathematical challenge and practically relevant when coupling quantum models to classical models.

Incorporating the Pauli principle for fermions in the simplest possible model yields the case of a Hartree equation with an additional "local" nonlinearity given by a power of the local density, the "Schrödinger–Poisson-Xα equation" (see [25]).

*Received by the editors June 8, 2004; accepted for publication (in revised form) March 25, 2005; published electronically October 3, 2005. The authors acknowledge support by the Austrian START award project (FWF, contract Y-137-TEC) of N.J.M. and by the Wissenschaftskolleg (doctoral school) "Differential Equations" (FWF, contract W8) as well as the European network HYKE, funded by the EC as contract HPRN-CT-2002-00282.

http://www.siam.org/journals/siap/66-1/60973.html

†MAB, Université Bordeaux 1, 351 cours de la Libération, F 33 405 Talence cedex, France (Remi.Carles@math.cnrs.fr).

‡Wolfgang Pauli Institute, c/o Inst. f. Math., Universität Wien, Nordbergstr. 15, A 1090 Wien, Austria (mauser@courant.nyu.edu).

§Wolfgang Pauli Institute, Wien, Austria, and ENS Lyon, 46 allée d'Italie, 69364 Lyon CEDEX, France (hans.peter.stimming@univie.ac.at).

In this paper we deal with the "semiclassical limit" of nonlinear Schrödinger equations of Hartree type, with a harmonic potential and a "weak" nonlinearity which is a convolution of the density with a more or less singular potential.

In three space dimensions, for the case where we convolute with the Newtonian potential $1/|x|$, the Hartree equation is the Schrödinger–Poisson system with harmonic potential:

$$\left\{\begin{array}{r} i\varepsilon\partial_t \mathrm{u}^\varepsilon + \frac{1}{2}\varepsilon^2\Delta \mathrm{u}^\varepsilon = \frac{|x|^2}{2}\mathrm{u}^\varepsilon + V(x)\mathrm{u}^\varepsilon\,, \\ \Delta V = |\mathrm{u}^\varepsilon|^2\,, \\ \mathrm{u}^\varepsilon_{|t=0} = \mathrm{u}^\varepsilon_0\,, \end{array}\right. \tag{1.1}$$

with $x \in \mathbb{R}^3$.

This equation typically arises if we consider the quantum mechanical time evolution of electrons in the mean field approximation of the many body effects, modeled by the Poisson equation, with a confinement modeled by the quadratic potential of the harmonic oscillator.

The limit $\varepsilon \to 0$ in such a quantum model corresponds to a "classical limit" of vanishing Planck constant $\hbar = \varepsilon \to 0$. We adopt the term "semiclassical limit" for what should properly be called "classical limit" (see the discussion in [31]), the term "semiclassical" being actually more appropriate for the situation of the homogenization limit from a Schrödinger equation with periodic potential (see, e.g., [2]).

The problem of the mathematically rigorous "classical limit" of the Schrödinger–Poisson system is highly nontrivial. First results of weak limits $\varepsilon \to 0$ to the Vlasov–Poisson system were given in [23] and [24] using Wigner transform techniques for the "mixed state case," where additional strong assumptions on the initial data can be imposed (which are necessary to guarantee a uniform L^2 bound on the Wigner function). In [31] this assumption was removed for the 1D case, and the classical limit for the "pure state" case was performed, where the notorious problem of nonuniqueness of the Vlasov–Poisson system with measure valued initial data reappeared. For an overview of this kind of "semiclassical limits" of Hartree equations see [26]. For an introduction to Wigner transforms and their comparison to WKB methods for the linear case see [11] and [29].

Up to a constant, (1.1) is equivalent to the Hartree equation

$$i\varepsilon\partial_t \mathrm{u}^\varepsilon + \frac{1}{2}\varepsilon^2\Delta \mathrm{u}^\varepsilon = \frac{|x|^2}{2}\mathrm{u}^\varepsilon + \left(|x|^{-1} * |\mathrm{u}^\varepsilon|^2\right)\mathrm{u}^\varepsilon, \quad \mathrm{u}^\varepsilon_{|t=0} = \mathrm{u}^\varepsilon_0\,. \tag{1.2}$$

We restrict our attention to small data cases with $\mathrm{u}^\varepsilon_0 = \varepsilon^{\alpha/2} f$, where f is independent of ε and $\alpha \geq 1$.

Notice that we can allow for more general data with initial plane oscillations,

$$\mathbf{u}^\varepsilon_{|t=0} = \varepsilon^{\alpha/2} f(x) e^{i\frac{x\cdot\xi_0}{\varepsilon}} \quad \text{for} \quad \xi_0 \in \mathbb{R}^3, \tag{1.3}$$

since the change of variables given in [6],

$$\mathbf{u}^\varepsilon(t,x) = \mathrm{u}^\varepsilon(t, x - \xi_0 \sin t) e^{i\left(x - \frac{\xi_0}{2}\sin t\right).\xi_0 \cos t/\varepsilon}\,, \tag{1.4}$$

yields the solution of (1.2). This change of variable could also be used in (1.6) below, and hence our results also hold for the more general ε-dependent class of data (1.3).

Note that "small data" can be equivalently written as "small nonlinearity," since with the change of the unknown $u^\varepsilon = \varepsilon^{-\alpha/2}\mathtt{u}^\varepsilon$, (1.2) becomes

$$i\varepsilon\partial_t u^\varepsilon + \frac{1}{2}\varepsilon^2\Delta u^\varepsilon = \frac{|x|^2}{2}u^\varepsilon + \varepsilon^\alpha\left(|x|^{-1} * |u^\varepsilon|^2\right)u^\varepsilon, \quad u^\varepsilon_{|t=0} = f\,. \tag{1.5}$$

We will consider the more general "semiclassical Hartree equation"

$$i\varepsilon\partial_t u^\varepsilon + \frac{1}{2}\varepsilon^2\Delta u^\varepsilon = \frac{|x|^2}{2}u^\varepsilon + \varepsilon^\alpha\left(|x|^{-\gamma} * |u^\varepsilon|^2\right)u^\varepsilon, \quad u^\varepsilon_{|t=0} = f\,, \tag{1.6}$$

with $\gamma > 0$, $\alpha \geq 1$, and $x \in \mathbb{R}^n$, where the space dimension $n \geq 2$ may be different from 3.

The first point to notice is that in the linear case, the harmonic potential causes focusing at the origin (resp., at $(-1)^k\xi_0$ in the case (1.4)) at times $t = \pi/2 + k\pi$ for any $k \in \mathbb{N}$. The solution $u^\varepsilon_{\rm free}$ of the linear equation

$$i\varepsilon\partial_t u^\varepsilon_{\rm free} + \frac{1}{2}\varepsilon^2\Delta u^\varepsilon_{\rm free} = \frac{|x|^2}{2}u^\varepsilon_{\rm free}, \quad u^\varepsilon_{\rm free|t=0} = f\,, \tag{1.7}$$

is initially of size $\mathcal{O}(1)$. At time $t = \pi/2$, the solution focuses at the origin and is of order $\mathcal{O}(\varepsilon^{-n/2})$; it is of order $\mathcal{O}(1)$ for $t = \pi$, and so on (for a more precise analysis, see [6]). This phenomenon is easy to read from Mehler's formula (see, e.g., [10, 18]): for $0 < t < \pi$, we have

$$u^\varepsilon_{\rm free}(t,x) = \frac{e^{-in\frac{\pi}{4}}}{(2\pi\varepsilon\sin t)^{n/2}}\int_{\mathbb{R}^n} e^{\frac{i}{\varepsilon\sin t}\left(\frac{x^2+y^2}{2}\cos t - x\cdot y\right)} f(y)dy\,. \tag{1.8}$$

Essentially, one can apply a stationary phase formula for $t \in]0,\pi/2[\,\cup\,]\pi/2,\pi[$ ($u^\varepsilon_{\rm free}$ is $\mathcal{O}(1)$), while it is not possible at $t = \pi/2$ ($u^\varepsilon_{\rm free}$ is $\mathcal{O}(\varepsilon^{-n/2})$). Following the same approach as in [3], we get the following distinctions.

	$\alpha > \gamma$	$\alpha = \gamma$
$\alpha > 1$	linear WKB, linear focus	linear WKB, nonlinear focus
$\alpha = 1$	nonlinear WKB, linear focus	nonlinear WKB, nonlinear focus

The expression "linear WKB" means that the nonlinear Hartree interaction term is negligible away from the focus (when the WKB approximation is valid); "linear focus" means that the nonlinearity is negligible near the focus; the WKB regime (resp., the focus) is "nonlinear" when the Hartree term has a leading order influence away from (resp., in the neighborhood of) the focus, in the limit $\varepsilon \to 0$. This terminology follows [19].

We did not obtain a rigorous description of the case $\alpha = \gamma = 1$, which corresponds to the Schrödinger–Poisson system (1.1) when $n = 3$. This problem seems out of reach for the methods currently available in this field. On the other hand, we study rigorously the other three cases in an exhaustive way.

In section 3, we prove that the Hartree term has no influence at leading order when $\alpha > \gamma = 1$. Back to (1.2), this shows that initial data of size $\varepsilon^{\alpha/2}$ with $\alpha > 1$ yield a linearizable solution. The expected critical size is $\sqrt{\varepsilon}$; this heuristic is reinforced by the next three sections.

In section 4, we study the case $\alpha = 1 > \gamma$. We prove that the nonlinear term must be taken into account to describe the solution u^ε. It is so through a slowly oscillating phase term. On the other hand, no nonlinear effect occurs at leading order near the focus.

In section 5, we show that when $\alpha = \gamma > 1$, nonlinear effects occur at leading order at the foci, while they are negligible elsewhere. This phenomenon is the same as in [6] for the nonlinear Schrödinger equation; each focus crossing is described in terms of the scattering operator associated to the Hartree equation

$$i\partial_t \psi + \frac{1}{2}\Delta\psi = \left(|x|^{-\gamma} * |\psi|^2\right)\psi\,. \tag{1.9}$$

In section 6, we perform a formal computation suggested by the results of sections 4 and 5. This can be seen as further evidence that nonlinear effects are always relevant in the case $\alpha = \gamma = 1$, along with a precise idea of the nature of these nonlinear effects, which we expect to be true. We add a brief discussion of the case of an additional local nonlinearity in the equation and some remarks on the Wigner measures in view of the ill-posedness results of [5].

This program is very similar to the one achieved in [3]. We want to underscore at least two important differences. First, we have to adapt the notion of oscillatory integral to incorporate the presence of the harmonic potential (see section 4.1). Second, the power-like nonlinearity treated in [3] is replaced by a Hartree-type nonlinearity. This yields different and less technical proofs (we do not use Strichartz estimates in sections 3 and 4) and makes a more complete description of the above table possible; the case "nonlinear WKB, linear focus" was treated very partially in [3] due to the lack of regularity of the map $z \mapsto |z|^{2\sigma}z$ for small $\sigma > 0$. This technical difficulty does not occur in the present case, and the main result of section 4 (Proposition 4.1) is proved with no restriction.

The content of this article is as explained above, plus a paragraph dedicated to a quick review of the facts we will need about the Cauchy problem (1.6) (see section 2).

We will use the following notation throughout this paper.

Notation. If $(a^\varepsilon)_{\varepsilon\in]0,1]}$ and $(b^\varepsilon)_{\varepsilon\in]0,1]}$ are two families of numbers, we write

$$a^\varepsilon \lesssim b^\varepsilon$$

if there exists C independent of $\varepsilon \in]0,1]$ such that for any $\varepsilon \in]0,1]$, $a^\varepsilon \leq Cb^\varepsilon$.

2. The Cauchy problem. Before studying semiclassical limits, we recall some known facts about the initial value problem (1.6). We will always assume that the initial datum f is in the space Σ defined by

$$\Sigma := \left\{\phi \in H^1(\mathbb{R}^n)\ ;\ \|\phi\|_\Sigma := \|\phi\|_{L^2} + \|x\phi\|_{L^2} + \|\nabla\phi\|_{L^2} < +\infty\right\}.$$

This space is natural in the case of Schrödinger equations with harmonic potential, since Σ is the domain of $\sqrt{-\Delta + |x|^2}$ (see, for instance, [27]). Local existence results for (1.6) follow, for instance, from Strichartz inequalities (one can do without these inequalities; see [27]). Global existence results then stem from conservation laws (see (2.3) below). From Mehler's formula (1.8), Strichartz type estimates are available for

$$e^{-i\frac{t}{2\varepsilon}(-\varepsilon^2\Delta + x^2)} =: \mathcal{U}^\varepsilon(t)\,.$$

DEFINITION 2.1. *Let $n \geq 2$. A pair (q,r) is* admissible *if $2 \leq r < \frac{2n}{n-2}$ (resp., $2 \leq r < \infty$ if $n = 2$) and*

$$\frac{2}{q} = \delta(r) \equiv n\left(\frac{1}{2} - \frac{1}{r}\right).$$

Following [6], we have the following scaled Strichartz inequalities.

PROPOSITION 2.2. *Let I be a* finite *time interval.*

(1) *For any admissible pair (q,r), there exists $C_r(I)$ such that*

$$\varepsilon^{\frac{1}{q}} \left\|\mathcal{U}^\varepsilon(t)\phi\right\|_{L^q(I;L^r)} \leq C_r(I)\|\phi\|_{L^2}\,. \tag{2.1}$$

(2) *For any admissible pairs (q_1,r_1) and (q_2,r_2), there exists $C_{r_1,r_2}(I)$ such that*

$$\varepsilon^{\frac{1}{q_1}+\frac{1}{q_2}} \left\|\int_{I\cap\{s\leq t\}} \mathcal{U}^\varepsilon(t-s)F(s)ds\right\|_{L^{q_1}(I;L^{r_1})} \leq C_{r_1,r_2}(I)\left\|F\right\|_{L^{q_2'}(I;L^{r_2'})}\,. \tag{2.2}$$

The above constants are independent of ε.

The main result of this section follows from [7, 12]. Denote

$$Y(I) = \{\phi \in C(I,\Sigma)\ ;\ \phi,\, |x|\phi,\, \nabla_x\phi \in L^q_{\rm loc}(I,L^r_x)\ \forall (q,r) \text{ admissible}\}\,.$$

PROPOSITION 2.3. *Fix $\varepsilon \in]0,1]$ and let $f \in \Sigma$. Then (1.6) has a unique solution $u^\varepsilon \in Y(\mathbb{R})$. Moreover, the following quantities are independent of time:*

$$\begin{aligned} &\textit{Mass: } \|u^\varepsilon(t)\|_{L^2}\,,\\ &\textit{Energy: } \frac{1}{2}\|\varepsilon\nabla_x u^\varepsilon(t)\|_{L^2}^2 + \frac{1}{2}\|xu^\varepsilon(t)\|_{L^2}^2 + \varepsilon\int_{\mathbb{R}^n}\left(|x|^{-\gamma} * |u^\varepsilon|^2\right)|u^\varepsilon(t,x)|^2dx. \end{aligned} \tag{2.3}$$

It was noticed in [6] that this result can be retrieved very simply thanks to the following lemma, which we will use to prove asymptotics.

LEMMA 2.4 (see [6]). *Define the operators*

$$J^\varepsilon(t) = \frac{x}{\varepsilon}\sin t - i\cos t\nabla_x, \qquad H^\varepsilon(t) = x\cos t + i\varepsilon\sin t\nabla_x\,. \tag{2.4}$$

J^ε and H^ε satisfy the following properties.

- *They are Heisenberg observables:*

$$J^\varepsilon(t) = -i\,\mathcal{U}^\varepsilon(t)\nabla_x\mathcal{U}^\varepsilon(-t), \qquad H^\varepsilon(t) = \mathcal{U}^\varepsilon(t)x\,\mathcal{U}^\varepsilon(-t)\,. \tag{2.5}$$

- *The commutation relation:*

$$\left[J^\varepsilon(t), i\varepsilon\partial_t + \frac{\varepsilon^2}{2}\Delta - \frac{|x|^2}{2}\right] = \left[H^\varepsilon(t), i\varepsilon\partial_t + \frac{\varepsilon^2}{2}\Delta - \frac{|x|^2}{2}\right] = 0\,. \tag{2.6}$$

- *Denote $M^\varepsilon(t) = e^{-i\frac{x^2}{2\varepsilon}\tan t}$ and $Q^\varepsilon(t) = e^{i\frac{x^2}{2\varepsilon}\cot t}$; then*

$$J^\varepsilon(t) = -i\cos t\, M^\varepsilon(t)\nabla_x M^\varepsilon(-t), \qquad H^\varepsilon(t) = i\varepsilon\sin t\, Q^\varepsilon(t)\nabla_x Q^\varepsilon(-t)\,. \tag{2.7}$$

- *The modified Sobolev inequalities. Let $2 \leq r \leq \frac{2n}{n-2}$ ($2 \leq r < \infty$ if $n = 2$); there exists C_r independent of ε such that, for any $\phi \in \Sigma$,*

$$\begin{aligned} \|\phi\|_{L^r} &\leq C_r|\cos t|^{-\delta(r)}\|\phi\|_{L^2}^{1-\delta(r)}\|J^\varepsilon(t)\phi)\|_{L^2}^{\delta(r)}\,,\\ \|\phi\|_{L^r} &\leq C_r|\varepsilon\sin t|^{-\delta(r)}\|\phi\|_{L^2}^{1-\delta(r)}\|H^\varepsilon(t)\phi\|_{L^2}^{\delta(r)}\,. \end{aligned} \tag{2.8}$$

- *Action on nonlinear Hartree term: for $\phi = \phi(t,x)$,*

$$J^\varepsilon(t)\left(\left(|x|^{-\gamma} * |\phi|^2\right)\phi\right) = \left(|x|^{-\gamma} * |\phi|^2\right) J^\varepsilon(t)\phi + 2\,\mathrm{Re}\left(|x|^{-\gamma} * \left(\overline{\phi} J^\varepsilon(t)\phi\right)\right)\phi\,. \tag{2.9}$$

The same holds for $H^\varepsilon(t)$.

Remark 2.5. Property (2.6) follows from (2.5), which is the way J^ε and H^ε appear in the linear theory (see, e.g., [30, p. 108]). Property (2.8) is a consequence of Gagliardo–Nirenberg inequalities and (2.7). Finally, (2.9) stems from (2.7).

3. "Very weak nonlinearity" case. In this section, we study the semiclassical limit of u^ε when $\gamma = 1$ *and* $\alpha > 1$, which is equivalent to "very small" data in our context (cf. (1.2)). This case includes the 3D Schrödinger–Poisson equation with "very small data." We prove that the Hartree term plays no role at leading order.

PROPOSITION 3.1. *Let $f \in \Sigma$, $n \geq 2$, and assume $\alpha > \gamma = 1$. Then for any $T > 0$,*

$$\|u^\varepsilon - u^\varepsilon_{\mathrm{free}}\|_{L^\infty([0,T];L^2)} = \mathcal{O}\left(\varepsilon^{\alpha-1}\ln\frac{1}{\varepsilon}\right) \quad \text{as } \varepsilon \to 0\,,$$

and for any $\delta > 0$ ($\alpha - 1 - \delta > 0$),

$$\|A^\varepsilon(t)\,(u^\varepsilon - u^\varepsilon_{\mathrm{free}})\|_{L^\infty([0,T];L^2)} = \mathcal{O}\left(\varepsilon^{\alpha-1-\delta}\right) \quad \text{as } \varepsilon \to 0\,,$$

where A^ε is either of the operators J^ε or H^ε, and $u^\varepsilon_{\mathrm{free}}$ is the solution of (1.7).

Remark 3.2. Using modified Sobolev inequalities (2.8), we can deduce L^p estimates for $u^\varepsilon - u^\varepsilon_{\mathrm{free}}$ for $2 \leq p \leq 2n/(n-2)$ ($2 \leq p < \infty$ if $n = 2$) from the above result.

Remark 3.3. We could probably get the logarithmic estimate for the second part of the statement as well, using Strichartz estimates. The proof given below is not technically involved and suffices for our purpose; we do not seek sharp results.

Proof. Denote $w^\varepsilon = u^\varepsilon - u^\varepsilon_{\mathrm{free}}$. It solves the initial value problem

$$i\varepsilon\partial_t w^\varepsilon + \frac{1}{2}\varepsilon^2\Delta w^\varepsilon = \frac{|x|^2}{2}w^\varepsilon + \varepsilon^\alpha\left(|x|^{-1} * |u^\varepsilon|^2\right)u^\varepsilon, \qquad w^\varepsilon_{|t=0} = 0\,.$$

Standard energy estimates for Schrödinger equations yield

$$\varepsilon\partial_t\|w^\varepsilon(t)\|_{L^2} \lesssim \varepsilon^\alpha\left\|\left(|x|^{-1} * |u^\varepsilon|^2\right)u^\varepsilon\right\|_{L^2}\,. \tag{3.1}$$

From Hölder's inequality, we have

$$\left\|\left(|x|^{-1} * |u^\varepsilon|^2\right)u^\varepsilon\right\|_{L^2} \leq \left\||x|^{-1} * |u^\varepsilon|^2\right\|_{L^r}\|u^\varepsilon\|_{L^k} \quad \text{for } \frac{1}{r} + \frac{1}{k} = \frac{1}{2}\,. \tag{3.2}$$

From the Hardy–Littlewood–Sobolev inequality,

$$\left\||x|^{-1} * |u^\varepsilon(t)|^2\right\|_{L^r} \lesssim \|u^\varepsilon(t)\|^2_{L^p} \quad \text{for } 1 < r, \frac{p}{2} < \infty \text{ and } 1 + \frac{1}{r} = \frac{2}{p} + \frac{1}{n}\,. \tag{3.3}$$

Therefore, (3.1) yields

$$\varepsilon\partial_t\|w^\varepsilon(t)\|_{L^2} \lesssim \varepsilon^\alpha\|u^\varepsilon(t)\|^2_{L^p}\|u^\varepsilon(t)\|_{L^k}\,, \tag{3.4}$$

where p and k satisfy the properties stated in (3.2) and (3.3). For $k = 2$, $r = \infty$, and $p = 2n/(n-1)$, the algebraic identities stated in (3.2) and (3.3) are satisfied. Now

since the conditions $1 < r < \infty$ and $1 < p/2 < \infty$ are open, a continuity argument shows that we can find p and k satisfying all the properties stated in (3.2) and (3.3). Notice that they imply the relation $2\delta(p) + \delta(k) = 1$ and hence $\delta(p)$, $\delta(k) < 1$; this allows us to use weighted Gagliardo–Nirenberg inequalities.

We have $w^\varepsilon_{|t=0} = 0$, and from Proposition 2.3, $w^\varepsilon \in C(\mathbb{R}_+; \Sigma)$. Therefore, there exists $t^\varepsilon > 0$ such that

$$\|J^\varepsilon(t) w^\varepsilon\|_{L^2} \leq 1 \tag{3.5}$$

for $0 \leq t \leq t^\varepsilon$. The argument of the proof then follows [28] (see also [6]). Recall that from (2.5), $\|J^\varepsilon(t) u^\varepsilon_{\rm free}\|_{L^2} = \|\nabla f\|_{L^2}$.

Because of (2.6), $J^\varepsilon u^\varepsilon_{\rm free}$ solves the linear Schrödinger equation with harmonic potential, and $\|J^\varepsilon(t) u^\varepsilon_{\rm free}\|_{L^2} \equiv \|\nabla f\|_{L^2}$. So long as (3.5) holds, we have, from (2.8),

$$\|u^\varepsilon(t)\|_{L^p} \leq \frac{C_0}{|\cos t|^{\delta(p)}}, \qquad \|u^\varepsilon(t)\|_{L^k} \leq \frac{C_0}{|\cos t|^{\delta(k)}}$$

for some C_0 independent of ε and t. Then (3.4) yields

$$\varepsilon \partial_t \|w^\varepsilon(t)\|_{L^2} \lesssim \frac{\varepsilon^\alpha}{|\cos t|^{2\delta(p)+\delta(k)}} = \frac{\varepsilon^\alpha}{|\cos t|}\,.$$

Integration in time on $[0,t]$ yields, so long as (3.5) holds,

$$\|w^\varepsilon\|_{L^\infty([0,t];L^2)} \lesssim \varepsilon^{\alpha-1} \int_0^t \frac{d\tau}{|\cos\tau|}\,.$$

For $t < \pi/2$, we get, so long as (3.5) holds,

$$\|w^\varepsilon\|_{L^\infty([0,t];L^2)} \lesssim \varepsilon^{\alpha-1} \left|\ln\left(\frac{\pi}{2} - t\right)\right|\,.$$

From (2.6), $J^\varepsilon(t) w^\varepsilon$ solves

$$i\varepsilon\partial_t J^\varepsilon w^\varepsilon + \frac{1}{2}\varepsilon^2 \Delta J^\varepsilon w^\varepsilon = \frac{|x|^2}{2} J^\varepsilon w^\varepsilon + \varepsilon^\alpha J^\varepsilon \left(\left(|x|^{-1} * |u^\varepsilon|^2\right) u^\varepsilon\right), \qquad J^\varepsilon w^\varepsilon_{|t=0} = 0\,.$$

By use of (2.9), the energy estimate for $J^\varepsilon w^\varepsilon$ yields

$$\begin{aligned}\varepsilon\partial_t \|J^\varepsilon(t) w^\varepsilon\|_{L^2} &\lesssim \varepsilon^\alpha \left(\left\|\left(|x|^{-1} * |u^\varepsilon|^2\right) J^\varepsilon(t) u^\varepsilon\right\|_{L^2} + \left\||x|^{-1} * (\overline{u^\varepsilon} J^\varepsilon u^\varepsilon)\cdot u^\varepsilon\right\|_{L^2}\right)\\ &\lesssim \varepsilon^\alpha \left(\left\||x|^{-1} * |u^\varepsilon|^2\right\|_{L^\infty} \|J^\varepsilon(t) u^\varepsilon\|_{L^2} + \left\||x|^{-1} * (\overline{u^\varepsilon} J^\varepsilon u^\varepsilon)\cdot u^\varepsilon\right\|_{L^2}\right).\end{aligned}$$

For the first term of the right-hand side, use the easy estimate

$$\left\||x|^{-1} * f\right\| \lesssim \|f\|_{L^{(n^-)'}} + \|f\|_{L^{(n^+)'}},$$

where n^- (resp., n^+) stands for $n - \eta$ (resp., $n + \eta$) for any small $\eta > 0$. We have

$$\left\||x|^{-1} * |u^\varepsilon|^2\right\|_{L^\infty} \lesssim \|u^\varepsilon(t)\|^2_{L^{\kappa^-}} + \|u^\varepsilon(t)\|^2_{L^{\kappa^+}}\,, \quad \text{with} \quad \kappa = \frac{2n}{n-1}\,.$$

It is at this stage that we lose the logarithmic rate (we cannot use Hardy–Littlewood–Sobolev inequality when an exponent is infinite): using Strichartz estimates (see section 5), we believe that we could recover that rate, with a more technically involved proof.

For the second term, we proceed as in the beginning of the proof. From Hölder's inequality,

$$\left\| \left(|x|^{-1} * \overline{u^\varepsilon} J^\varepsilon u^\varepsilon \right) \cdot u^\varepsilon \right\|_{L^2} \leq \left\| |x|^{-1} * (\overline{u^\varepsilon} J^\varepsilon u^\varepsilon) \right\|_{L^r} \|u^\varepsilon\|_{L^\sigma} , \text{ with } \frac{1}{r} + \frac{1}{\sigma} = \frac{1}{2} \cdot \tag{3.6}$$

From the Hardy–Littlewood–Sobolev inequality, this is estimated, up to a constant, by

$$\|\overline{u^\varepsilon} J^\varepsilon u^\varepsilon\|_{L^p} \|u^\varepsilon\|_{L^\sigma} , \text{ with } 1 + \frac{1}{r} = \frac{1}{p} + \frac{1}{n} \text{ for } 1 < r, p < \infty . \tag{3.7}$$

Use of Hölder's inequality again yields an estimate by

$$\|u^\varepsilon\|_{L^k} \|J^\varepsilon u^\varepsilon\|_{L^2} \|u^\varepsilon\|_{L^\sigma} , \text{ with } \frac{1}{p} = \frac{1}{2} + \frac{1}{k} \cdot \tag{3.8}$$

Take $r = n$, $\sigma = 2n/(n-2)$, $k = 2$, and $p = 1$; the algebraic identities from (3.6), (3.7), and (3.8) are satisfied, but not the bound $p > 1$. Decreasing σ slightly increases p (take σ large but finite when $n = 2$), so we can find indices satisfying (3.6), (3.7), and (3.8) by a continuity argument. Note that they satisfy $\delta(k) + \delta(\sigma) = 1$, and each term is positive.

Gathering all these estimates together we get the energy estimate

$$\varepsilon \partial_t \|J^\varepsilon(t) w^\varepsilon\|_{L^2} \lesssim \varepsilon^\alpha \left(\|u^\varepsilon(t)\|^2_{L^{\kappa^-}} + \|u^\varepsilon(t)\|^2_{L^{\kappa^+}} + \|u^\varepsilon\|_{L^k} \|u^\varepsilon\|_{L^\sigma} \right) \|J^\varepsilon(t) u^\varepsilon\|_{L^2} .$$

So long as (3.5) holds, we deduce from (2.8) that

$$\begin{aligned} \varepsilon \partial_t \|J^\varepsilon(t) w^\varepsilon\|_{L^2} &\lesssim \varepsilon^\alpha \left(\frac{1}{|\cos t|^{2\delta(\kappa^-)}} + \frac{1}{|\cos t|^{2\delta(\kappa^+)}} + \frac{1}{|\cos t|^{\delta(k)+\delta(\sigma)}} \right) \\ &\lesssim \varepsilon^\alpha \left(\frac{1}{|\cos t|^{2\delta(\kappa^+)}} + \frac{1}{|\cos t|} \right) \lesssim \frac{\varepsilon^\alpha}{|\cos t|^{1^+}} . \end{aligned}$$

Integrate this, so long as (3.5) holds:

$$\|J^\varepsilon w^\varepsilon\|_{L^\infty([0,t];L^2)} \lesssim \varepsilon^{\alpha-1} \left(\frac{\pi}{2} - t \right)^{0^-} .$$

Fix $\delta, \Lambda > 0$. So long as (3.5) holds, we infer, for $t \leq \pi/2 - \Lambda\varepsilon$,

$$\|J^\varepsilon w^\varepsilon\|_{L^\infty([0,t];L^2)} \lesssim \varepsilon^{\alpha-1} (\Lambda\varepsilon)^{-\delta} .$$

Therefore, there exists $\varepsilon_\Lambda > 0$ such that, for $0 < \varepsilon \leq \varepsilon_\Lambda$, (3.5) holds up to time $\pi/2 - \Lambda\varepsilon$, with the estimates

$$\|w^\varepsilon\|_{L^\infty([0,\pi/2-\Lambda\varepsilon];L^2)} \lesssim \varepsilon^{\alpha-1} \ln \frac{1}{\varepsilon}, \qquad \|J^\varepsilon w^\varepsilon\|_{L^\infty([0,\pi/2-\Lambda\varepsilon];L^2)} \lesssim \varepsilon^{\alpha-1-\delta} . \tag{3.9}$$

An estimate similar to that of $J^\varepsilon w^\varepsilon$ then follows for $H^\varepsilon w^\varepsilon$, since from the conservation laws (2.3), $\|H^\varepsilon(t) u^\varepsilon\|_{L^2} \lesssim \|f\|_\Sigma$.

Denote $I^\varepsilon_\Lambda = [\pi/2 - \Lambda\varepsilon, \pi/2 + \Lambda\varepsilon]$. Mimicking the above computations, we have

$$\|w^\varepsilon\|_{L^\infty(I^\varepsilon_\Lambda;L^2)} \lesssim \left\| w^\varepsilon \left(\frac{\pi}{2} - \Lambda\varepsilon \right) \right\|_{L^2} + \varepsilon^{\alpha-1} \int_{I^\varepsilon_\Lambda} \|u^\varepsilon(\tau)\|^2_{L^p} \|u^\varepsilon(\tau)\|_{L^k} d\tau ,$$

where p and k satisfy (3.2) and (3.3). Recall that they satisfy $2\delta(p)+\delta(k)=1$. Using the conservations of mass and energy (2.3), along with Gagliardo–Nirenberg inequalities, we have, for any t,

$$\|u^\varepsilon(t)\|_{L^p}\lesssim \varepsilon^{-\delta(p)}, \qquad \|u^\varepsilon(t)\|_{L^k}\lesssim \varepsilon^{-\delta(k)}\,.$$

We deduce

$$\|w^\varepsilon\|_{L^\infty(I^\varepsilon_\Lambda;L^2)}\lesssim \left\|w^\varepsilon\left(\frac{\pi}{2}-\Lambda\varepsilon\right)\right\|_{L^2}+\varepsilon^{\alpha-1}\varepsilon^{-2\delta(p)-\delta(k)}\left|I^\varepsilon_\Lambda\right|\lesssim \varepsilon^{\alpha-1}\ln\frac{1}{\varepsilon}+\Lambda\varepsilon^{\alpha-1}\,.$$

The same method yields, since (2.3) shows that $\|H^\varepsilon(t)u^\varepsilon\|_{L^2}\lesssim\|f\|_\Sigma$,

$$\|H^\varepsilon w^\varepsilon\|_{L^\infty(I^\varepsilon_\Lambda;L^2)}\lesssim \varepsilon^{\alpha-1-\delta}\quad \text{for any } \delta>0\,.$$

To treat the case of $J^\varepsilon w^\varepsilon$, introduce

$$z_\varepsilon(t)=\sup_{\frac{\pi}{2}-\Lambda\varepsilon\le\tau\le t}\|J^\varepsilon(\tau)w^\varepsilon\|_{L^2}\,.$$

Proceeding as above, we have

$$\begin{aligned} z_\varepsilon(t)&\lesssim \left\|J^\varepsilon\left(\frac{\pi}{2}-\Lambda\varepsilon\right)w^\varepsilon\right\|_{L^2}+\varepsilon^{\alpha-1}\int_{\frac{\pi}{2}-\Lambda\varepsilon}^t\left\|J^\varepsilon(\tau)\left(|x|^{-1}*|u^\varepsilon|^2u^\varepsilon\right)\right\|_{L^2}d\tau\\ &\lesssim \varepsilon^{\alpha-1^+}+\varepsilon^{\alpha-1}\int_{\frac{\pi}{2}-\Lambda\varepsilon}^t\varepsilon^{-1^+}\left(z_\varepsilon(\tau)+\|J^\varepsilon(\tau)u^\varepsilon_{\rm free}\|_{L^2}\right)d\tau\,. \end{aligned} \tag{3.10}$$

We can then apply the Gronwall lemma (recall that $\|J^\varepsilon(\tau)u^\varepsilon_{\rm free}\|_{L^2}\equiv\|\nabla f\|_{L^2}$):

$$z_\varepsilon(t)\lesssim\varepsilon^{\alpha-1^+}.$$

Gathering this information we get, for any $\delta>0$,

$$\begin{aligned} \|w^\varepsilon\|_{L^\infty(I^\varepsilon_\Lambda;L^2)}&\lesssim\varepsilon^{\alpha-1}\ln\frac{1}{\varepsilon}\,,\\ \|J^\varepsilon w^\varepsilon\|_{L^\infty(I^\varepsilon_\Lambda;L^2)}+\|H^\varepsilon w^\varepsilon\|_{L^\infty(I^\varepsilon_\Lambda;L^2)}&\lesssim\varepsilon^{\alpha-1-\delta}\,. \end{aligned}$$

For $t\in[\pi/2+\varepsilon,\pi]$, we can use the same proof as for $t\in[0,\pi/2-\varepsilon]$ to obtain

$$\begin{aligned} \|w^\varepsilon\|_{L^\infty([0,\pi];L^2)}&\lesssim\varepsilon^{\alpha-1}\ln\frac{1}{\varepsilon}\,,\\ \|J^\varepsilon w^\varepsilon\|_{L^\infty([0,\pi];L^2)}+\|H^\varepsilon w^\varepsilon\|_{L^\infty([0,\pi];L^2)}&\lesssim\varepsilon^{\alpha-1-\delta}\,. \end{aligned}$$

Repetition of the same argument a finite number of times covers any given time interval $[0,T]$ and completes the proof of Proposition 3.1. □

4. Nonlinear propagation and linear focus. In this paragraph, we assume $\alpha=1$ and $\gamma<1$. We define

$$g(t,x)=-\left(|x|^{-\gamma}*|f|^2\right)(x)\int_0^t\frac{d\tau}{|\cos\tau|^\gamma}\,. \tag{4.1}$$

This function is well defined for any t, since $\gamma<1$. We will see later how this function appears.

PROPOSITION 4.1. *Let $n \geq 2$, $f \in \Sigma$, and assume $\gamma < \alpha = 1$. Let A^ε be one of the operators Id, J^ε, or H^ε.*

- *For $0 \leq t < \pi/2$, the following asymptotic relation holds:*

$$\sup_{0\leq\tau\leq t} \left\| A^\varepsilon(\tau) \left(u^\varepsilon(\tau,x) - \frac{1}{(\cos\tau)^{n/2}} f\left(\frac{x}{\cos\tau}\right) e^{-i\frac{x^2}{2\varepsilon}\tan\tau + ig\left(\tau,\frac{x}{\cos\tau}\right)} \right) \right\|_{L^2_x} \underset{\varepsilon\to 0}{\longrightarrow} 0\,.$$

- *For $\pi/2 < t \leq \pi$,*

$$\sup_{t\leq\tau\leq\pi} \left\| A^\varepsilon(\tau) \left(u^\varepsilon(\tau,x) - \frac{e^{-in\frac{\pi}{2}}}{(\cos\tau)^{n/2}} f\left(\frac{x}{\cos\tau}\right) e^{-i\frac{x^2}{2\varepsilon}\tan\tau + ig\left(\tau,\frac{x}{\cos\tau}\right)} \right) \right\|_{L^2_x} \underset{\varepsilon\to 0}{\longrightarrow} 0\,.$$

- *For $t = \pi/2$,*

$$\left\| B^\varepsilon \left(u^\varepsilon\left(\frac{\pi}{2}\right) - \frac{1}{\varepsilon^{n/2}} \mathcal{F}\left(f e^{ig\left(\frac{\pi}{2}\right)} \right)\left(\frac{\cdot}{\varepsilon}\right) \right) \right\|_{L^2} \underset{\varepsilon\to 0}{\longrightarrow} 0\,,$$

where B^ε is one of the operators Id, $\frac{x}{\varepsilon}$, or $\varepsilon\nabla_x$, and the Fourier transform is defined by

$$\mathcal{F}\phi(\xi) = \widehat{\phi(\xi)} = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} e^{-ix\cdot\xi}\phi(x)dx\,. \tag{4.2}$$

Remark 4.2. We can also prove estimates for arbitrarily large time intervals, with the same proof as below.

Remark 4.3. The difference between the asymptotic behavior before and after the focus is measured only by the Maslov index, through the phase shift $e^{-in\pi/2}$; no nonlinear phenomenon occurs at leading order near the focus. On the other hand, nonlinear effects are relevant outside the focus, as shown by the presence of g.

4.1. Oscillatory integrals. The main tool for proving Proposition 4.1 is the same as in linear cases ([9]; see also [21, 3] for applications in nonlinear settings); we represent the solution u^ε as an oscillatory integral. Recall that $u^\varepsilon \in C(\mathbb{R};\Sigma)$ and that $e^{-i\frac{t}{2\varepsilon}(-\varepsilon^2\Delta + x^2)} = \mathcal{U}^\varepsilon(t)$ is a unitary group on L^2. Define a^ε by

$$a^\varepsilon(t,x) = \mathcal{U}^\varepsilon(-t)u^\varepsilon(t,x)\,. \tag{4.3}$$

We first seek a limit as $\varepsilon \to 0$ for a^ε before the focus. This is suggested by a formal computation as in [4] and by the following lemma.

LEMMA 4.4. *For $t \in [0,\pi/2[\,\cup\,]\pi/2,\pi]$, define $\mathtt{V}^\varepsilon$ by*

$$\mathtt{V}^\varepsilon(t)\phi(x) = \begin{cases} \dfrac{1}{(\cos t)^{n/2}}\phi\left(\dfrac{x}{\cos t}\right) e^{-i\frac{x^2}{2\varepsilon}\tan t} & \text{if } 0 \leq t < \pi/2\,, \\ \dfrac{e^{-in\pi/2}}{|\cos t|^{n/2}}\phi\left(\dfrac{x}{\cos t}\right) e^{-i\frac{x^2}{2\varepsilon}\tan t} & \text{if } \pi/2 < t \leq \pi\,. \end{cases} \tag{4.4}$$

For any $\phi \in H^1(\mathbb{R}^n)$, any $\theta \in]0,1/2]$, and any $t \in [0,\pi/2[\,\cup\,]\pi/2,\pi]$,

$$\|\mathcal{U}^\varepsilon(t)\phi - \mathtt{V}^\varepsilon(t)\phi\|_{L^2} \leq 2|\varepsilon\tan t|^\theta \|\phi\|_{H^1}\,.$$

Proof. Notice that from Mehler's formula (1.8), we can write, for $0 < t < \pi$,

$$\mathcal{U}^\varepsilon(t) = \mathcal{M}^\varepsilon_t \mathtt{D}^\varepsilon_t \mathcal{F} \mathcal{M}^\varepsilon_t, \quad \text{where} \quad \mathcal{M}^\varepsilon_t(x) = e^{-i\frac{x^2}{2\varepsilon\tan t}}\,,\ \mathtt{D}^\varepsilon_t\phi(x) = \frac{1}{(i\varepsilon\sin t)^{n/2}}\phi\left(\frac{x}{\sin t}\right),$$

and the Fourier transform is defined by (4.2). We infer

$$\|\mathcal{U}^\varepsilon(t)\phi - \mathtt{V}^\varepsilon(t)\phi\|_{L^2} = \left\| \frac{1}{(2i\pi\tan t)^{n/2}} \int e^{i\frac{|x-y|^2}{2\varepsilon\tan t}} f(y)dy - f(x) \right\|_{L^2}.$$

From the Parseval formula,

$$\frac{1}{(2i\pi\tan t)^{n/2}} \int e^{i\frac{|x-y|^2}{2\varepsilon\tan t}} f(y)dy = \frac{1}{(2\pi)^{n/2}} \int e^{-i\varepsilon\tan t\frac{\xi^2}{2}+ix\cdot\xi} \mathcal{F}f(\xi)d\xi\,;$$

therefore

$$\begin{aligned}\|\mathcal{U}^\varepsilon(t)\phi - \mathtt{V}^\varepsilon(t)\phi\|_{L^2} &= \frac{1}{(2\pi)^{n/2}} \left\| \int \left(e^{-i\varepsilon\tan t\frac{\xi^2}{2}} - 1 \right) e^{ix\cdot\xi} \mathcal{F}f(\xi)d\xi \right\|_{L^2} \\ &= \left\| \left(e^{-i\varepsilon\tan t\frac{\xi^2}{2}} - 1 \right) \mathcal{F}f(\xi) \right\|_{L^2}\end{aligned}$$

from the Plancherel formula. The lemma then follows from the estimate $|e^{is} - 1| \leq 2|s|^\theta$ for $0 \leq \theta \leq 1/2$. □

From Duhamel's principle, we have

$$u^\varepsilon(t) = \mathcal{U}^\varepsilon(t)f - i\int_0^t \mathcal{U}^\varepsilon(t-s) \left(\left(|x|^{-\gamma} * |u^\varepsilon|^2 \right) u^\varepsilon \right)(s)ds\,.$$

Using (4.3), we deduce

$$\partial_t a^\varepsilon(t) = -i\mathcal{U}^\varepsilon(-t) \left(\left(|x|^{-\gamma} * |u^\varepsilon|^2 \right) u^\varepsilon \right)(t)\,. \tag{4.5}$$

Now the formal computation begins. Assume $a^\varepsilon \to a$ as $\varepsilon \to 0$, in some suitable sense. Then $u^\varepsilon(t) \sim \mathcal{U}^\varepsilon(t)a(t)$, and from Lemma 4.4,

$$u^\varepsilon(t,x) \underset{\varepsilon\to 0}{\sim} \frac{1}{(\cos t)^{n/2}} a\left(t, \frac{x}{\cos t}\right) e^{-i\frac{x^2}{2\varepsilon}\tan t} \quad \text{for} \quad 0 \leq t < \pi/2\,.$$

Plugging this into (4.5) and using Lemma 4.4 again (with $\mathcal{U}^\varepsilon(-t)$ instead of $\mathcal{U}^\varepsilon(t)$, the result still holds), we find

$$\partial_t a(t,x) = \frac{-i}{|\cos t|^\gamma} \left(|x|^{-\gamma} * |a|^2 \right) a(t,x)\,.$$

Recall that $a_{|t=0} = u^\varepsilon_{|t=0} = f$ and notice that from the above ordinary differential equation, $\partial_t|a|^2 = 0$; we have $a(t,x) = f(x)e^{ig(t,x)}$, where

$$\partial_t g(t,x) = \frac{-1}{|\cos t|^\gamma} \left(|x|^{-\gamma} * |f|^2 \right)(x), \qquad g_{|t=0} = 0\,.$$

Integration of this equation yields the definition of $g(t,x)$ given in (4.1).

Proposition 4.1 stems from the more precise following proposition, Lemma 4.4, and a density argument. In view of a rigorous justification, denote

$$b^\varepsilon(t,x) = a^\varepsilon(t,x)e^{-ig(t,x)} = e^{-ig(t,x)}\mathcal{U}^\varepsilon(-t)u^\varepsilon(t,x)\,. \tag{4.6}$$

Proposition 4.5. *Let $f \in \Sigma \cap H^2(\mathbb{R}^n)$. Fix $\delta > 0$. There exists C_δ such that*

$$\sup_{0\leq t\leq\pi} \|b^\varepsilon(t) - f\|_\Sigma \leq \int_0^\pi \|\partial_t b^\varepsilon(t)\|_\Sigma dt \leq C_\delta \varepsilon^{1-\gamma-\delta}\,.$$

The first inequality is trivial. We prove the second one in three steps:

(i) On $[0, \pi/2 - \Lambda\varepsilon]$ for any $\Lambda > 0$, with a constant depending on δ and Λ;
(ii) On $[\pi/2 - \Lambda\varepsilon, \pi/2 + \Lambda\varepsilon]$, with a constant depending on δ and Λ;
(iii) On $[\pi/2 + \Lambda\varepsilon, \pi]$, with a constant depending on δ and Λ.

As in section 3, the parameter $\Lambda > 0$ is arbitrary, while it has to be large in the case $\alpha = \gamma > 1$ (see section 5 and [6]). This situation is typical for a case where the focus is "linear."

4.2. Asymptotic behavior before the focus. Fix $\Lambda, \delta > 0$. We prove that there exists $C_{\Lambda,\delta}$ such that

$$\int_0^{\frac{\pi}{2}-\Lambda\varepsilon} \|\partial_t b^\varepsilon(t)\|_\Sigma dt \leq C_{\Lambda,\delta}\varepsilon^{1-\gamma-\delta}\,. \tag{4.7}$$

Denote

$$y_\varepsilon(t) = \int_0^t \|\partial_t b^\varepsilon(\tau)\|_{H^1} d\tau\,.$$

From (4.5) and the definition (4.6),

$$\begin{aligned}\|\partial_t b^\varepsilon(t)\|_{L^2} &= \left\| \mathcal{U}^\varepsilon(-t)\left(\left(|x|^{-\gamma} * |u^\varepsilon|^2\right) u^\varepsilon\right)(t) - \frac{1}{|\cos t|^\gamma}\left(|x|^{-\gamma} * |f|^2\right) a^\varepsilon(t)\right\|_{L^2} \\ &= \left\| \left(|x|^{-\gamma} * |u^\varepsilon|^2\right) u^\varepsilon(t) - \frac{1}{|\cos t|^\gamma}\mathcal{U}^\varepsilon(t)\left(\left(|x|^{-\gamma} * |f|^2\right) a^\varepsilon\right)(t)\right\|_{L^2}\,. \end{aligned} \tag{4.8}$$

Lemma 4.4 suggests that we can replace $\mathcal{U}^\varepsilon$ with $\mathtt{V}^\varepsilon$ in the last expression, up to a controllable error. Before going into further details, we prove two lemmas which will be of constant use in the proof of Proposition 4.5.

LEMMA 4.6. *Assume $\gamma < 1$, and let $0 < \delta < 2(1-\gamma)$. There exist p and q with*

$$2\delta(2p') = \gamma + \frac{\delta}{2}\big(< 1\big)\ ,\ p < \frac{n}{\gamma}, \qquad \delta(2q') = \frac{\gamma+1}{2} + \frac{\delta}{4}\big(<1\big),\ q < \frac{n}{\gamma+1}\,,$$

and such that there exists C such that for any $\phi \in C_c^\infty(\mathbb{R}^n)$,

$$\begin{aligned}\left\||x|^{-\gamma} * \phi\right\|_{L^\infty} &\leq C\left(\|\phi\|_{L^1} + \|\phi\|_{L^{p'}}\right), \\ \left\|\nabla\left(|x|^{-\gamma} * \phi\right)\right\|_{L^\infty} &\leq C\left(\|\phi\|_{L^1} + \|\phi\|_{L^{q'}}\right).\end{aligned}$$

Proof. We have $2\delta(2p') = \gamma$ when $p = n/\gamma$, and $\delta(2q') = \frac{\gamma+1}{2}$ when $q = n/(\gamma+1)$. Therefore $p < n/\gamma$ and $q < n/(\gamma+1)$ if $2\delta(2p') = \gamma + \delta/2$ and $\delta(2q') = \frac{\gamma+1}{2} + \frac{\delta}{4}$.

Let $\chi \in C_c^\infty(\mathbb{R}_+, [0,1])$ with $\chi \equiv 1$ on $[0,1]$. We have

$$\begin{aligned}\||x|^{-\gamma} * \phi\|_{L^\infty} &\leq \left\|\left(\chi|x|^{-\gamma}\right) * \phi\right\|_{L^\infty} + \left\|\left((1-\chi)|x|^{-\gamma}\right) * \phi\right\|_{L^\infty} \\ &\leq \left\|\chi|x|^{-\gamma}\right\|_{L^p}\|\phi\|_{L^{p'}} + \|(1-\chi)|x|^{-\gamma}\|_{L^\infty}\|\phi\|_{L^1} \\ &\leq C\left(\|\phi\|_{L^{p'}} + \|\phi\|_{L^1}\right),\end{aligned}$$

where we have used $x \mapsto |x|^{-\gamma} \in L^p_{\mathrm{loc}}(\mathbb{R}^n)$ because $p < n/\gamma$. The other estimate is similar, since $\nabla|x|^{-\gamma} = \mathcal{O}(|x|^{-\gamma-1})$. ∎

LEMMA 4.7. *Let $\gamma < 1$ and $f \in \Sigma \cap H^2(\mathbb{R}^n)$. Recall that g is defined by* (4.1). *We have*

$$|x|^{-\gamma} * |f|^2 \in W^{2,\infty};\ g \in L^\infty_{\mathrm{loc}}(\mathbb{R}; W^{2,\infty});\ fe^{ig}, \left(|x|^{-\gamma} * |f|^2\right) fe^{ig} \in L^\infty_{\mathrm{loc}}(\mathbb{R}; H^2)\,.$$

Proof. From Lemma 4.6 and Sobolev embeddings,

$$\begin{aligned}
\left\| |x|^{-\gamma} * |f|^2 \right\|_{L^\infty} &\lesssim \|f\|_{L^2}^2 + \|f\|_{L^{2p'}}^2 \lesssim \|f\|_{H^1}^2 \,, \\
\left\| \nabla |x|^{-\gamma} * |f|^2 \right\|_{L^\infty} &\lesssim \|f\|_{L^2}^2 + \|f\|_{L^{2q'}}^2 \lesssim \|f\|_{H^1}^2 \,, \\
\left\| \nabla^2 |x|^{-\gamma} * |f|^2 \right\|_{L^\infty} &\lesssim \left\| \nabla |x|^{-\gamma} * \left(\nabla |f|^2 \right) \right\|_{L^\infty} \\
&\lesssim \|f\|_{L^2} \|\nabla f\|_{L^2} + \|f\|_{L^{2q'}} \|\nabla f\|_{L^{2q'}} \lesssim \|f\|_{H^2}^2 \,.
\end{aligned}$$

Since $t \mapsto |\cos t|^{-\gamma} \in L^1_{\rm loc}(\mathbb{R})$, we infer that $g \in L^\infty_{\rm loc}(\mathbb{R}; W^{2,\infty})$. The last two properties follow easily. ∎

We can now replace $\mathcal{U}^\varepsilon$ with $\mathtt{V}^\varepsilon$ in (4.8), up to the following error. From Lemmas 4.4, 4.6, and 4.7,

$$\begin{aligned}
\left\| \left(\mathcal{U}^\varepsilon(t) - \mathtt{V}^\varepsilon(t) \right) \left(\left(|x|^{-\gamma} * |f|^2 \right) a^\varepsilon \right)(t) \right\|_{L^2} &\lesssim |\varepsilon \tan t|^\theta \left\| \left(|x|^{-\gamma} * |f|^2 \right) a^\varepsilon(t) \right\|_{H^1} \\
&\lesssim |\varepsilon \tan t|^\theta \left\| |x|^{-\gamma} * |f|^2 \right\|_{W^{1,\infty}} \left\| a^\varepsilon(t) \right\|_{H^1} \\
&\lesssim |\varepsilon \tan t|^\theta \left(\left\| a^\varepsilon(t) \right\|_{L^2} + \left\| \nabla_x a^\varepsilon(t) \right\|_{L^2} \right) \\
&\lesssim |\varepsilon \tan t|^\theta \left(\|f\|_{L^2} + \left\| \nabla_x \left(b^\varepsilon e^{ig} \right) \right\|_{L^2} \right) \\
&\lesssim |\varepsilon \tan t|^\theta \left(1 + \left\| \nabla_x b^\varepsilon(t) \right\|_{L^2} \right) \\
&\lesssim |\varepsilon \tan t|^\theta \left(1 + \left\| \nabla_x \left(b^\varepsilon(t) - f \right) \right\|_{L^2} \right) \\
&\lesssim |\varepsilon \tan t|^\theta \left(1 + \int_0^t \left\| \partial_t b^\varepsilon(\tau) \right\|_{H^1} d\tau \right)
\end{aligned}$$

for $0 < \theta \le 1/2$ to be fixed later. Plugging this estimate into (4.8) we find

$$\begin{aligned}
\left\| \partial_t b^\varepsilon(t) \right\|_{L^2} \lesssim & \left\| \left(|x|^{-\gamma} * |u^\varepsilon|^2 \right) u^\varepsilon(t) - \frac{1}{|\cos t|^\gamma} \mathtt{V}^\varepsilon(t) \left(\left(|x|^{-\gamma} * |f|^2 \right) a^\varepsilon \right)(t) \right\|_{L^2} \\
& + \frac{|\varepsilon \tan t|^\theta}{|\cos t|^\gamma} \left(1 + y_\varepsilon(t) \right) .
\end{aligned} \tag{4.9}$$

We check that

$$\frac{1}{|\cos t|^\gamma} \mathtt{V}^\varepsilon(t) \left(\left(|x|^{-\gamma} * |f|^2 \right) \phi \right) = \left(|x|^{-\gamma} * |\mathtt{V}^\varepsilon(t) f|^2 \right) \mathtt{V}^\varepsilon(t) \phi \,. \tag{4.10}$$

Since we expect $\mathtt{V}^\varepsilon(t) a^\varepsilon(t)$ to be close to $\mathcal{U}^\varepsilon(t) a^\varepsilon(t) = u^\varepsilon(t)$ as $\varepsilon \to 0$, we estimate the difference

$$\begin{aligned}
& \left\| \left(|x|^{-\gamma} * |\mathtt{V}^\varepsilon(t) f|^2 \right) \left(\mathtt{V}^\varepsilon(t) a^\varepsilon(t) - \mathcal{U}^\varepsilon(t) a^\varepsilon(t) \right) \right\|_{L^2} \\
& \qquad \lesssim \left\| |x|^{-\gamma} * |\mathtt{V}^\varepsilon(t) f|^2 \right\|_{L^\infty} \left\| \left(\mathtt{V}^\varepsilon(t) - \mathcal{U}^\varepsilon(t) \right) \left(b^\varepsilon e^{ig} \right) \right\|_{L^2} \\
& \qquad \lesssim \left(\|\mathtt{V}^\varepsilon(t) f\|_{L^2}^2 + \|\mathtt{V}^\varepsilon(t) f\|_{L^{2p'}}^2 \right) (\varepsilon \tan t)^\theta \left\| b^\varepsilon(t) e^{ig(t)} \right\|_{H^1} \\
& \qquad \lesssim \left(1 + |\cos t|^{-2\delta(2p')} \right) |\varepsilon \tan t|^\theta \left(\|b^\varepsilon(t) - f\|_{H^1} + \|f\|_{H^1} \right)
\end{aligned}$$

using the modified Sobolev inequality (2.8). Since $2\delta(2p') = \frac{n}{p} > \gamma$, we infer from (4.9) that

$$\begin{aligned}
\left\| \partial_t b^\varepsilon(t) \right\|_{L^2} \lesssim \; & \frac{|\varepsilon \tan t|^\theta}{|\cos t|^{2\delta(2p')}} \left(1 + y_\varepsilon(t) \right) \\
& + \left\| \left(|x|^{-\gamma} * \left(|u^\varepsilon(t)|^2 - |\mathtt{V}^\varepsilon(t) f|^2 \right) \right) u^\varepsilon(t) \right\|_{L^2} .
\end{aligned} \tag{4.11}$$

From Lemma 4.6, the last term is estimated, up to a constant, by

$$\begin{aligned}
&\left\||u^\varepsilon(t)|^2-|\mathtt{V}^\varepsilon(t)f|^2\right\|_{L^1}+\left\||u^\varepsilon(t)|^2-|\mathtt{V}^\varepsilon(t)f|^2\right\|_{L^{p'}}\\
&\lesssim \left\|u^\varepsilon(t)-\mathtt{V}^\varepsilon(t)\left(fe^{ig(t)}\right)\right\|_{L^2}\left(\|u^\varepsilon(t)\|_{L^2}+\left\|\mathtt{V}^\varepsilon(t)\left(fe^{ig(t)}\right)\right\|_{L^2}\right)\\
&\quad+\left\|u^\varepsilon(t)-\mathtt{V}^\varepsilon(t)\left(fe^{ig(t)}\right)\right\|_{L^{2p'}}\left(\|u^\varepsilon(t)\|_{L^{2p'}}+\left\|\mathtt{V}^\varepsilon(t)\left(fe^{ig(t)}\right)\right\|_{L^{2p'}}\right).
\end{aligned}\tag{4.12}$$

For the first term of the right-hand side, we have, since $\mathcal{U}^\varepsilon$ is unitary on L^2,

$$\begin{aligned}
\left\|\mathcal{U}^\varepsilon(t)\left(b^\varepsilon e^{ig}\right)-\mathtt{V}^\varepsilon(t)\left(fe^{ig}\right)\right\|_{L^2}&\lesssim \|b^\varepsilon(t)-f\|_{L^2}+\left\|(\mathcal{U}^\varepsilon(t)-\mathtt{V}^\varepsilon(t))\left(fe^{ig(t)}\right)\right\|_{L^2}\\
&\lesssim y_\varepsilon(t)+|\varepsilon\tan t|^\theta\left\|fe^{ig(t)}\right\|_{H^1}.
\end{aligned}$$

In addition, notice that $\|u^\varepsilon(t)\|_{L^2}=\|\mathtt{V}^\varepsilon(t)f\|_{L^2}=\|f\|_{L^2}$. The second term is estimated thanks to the modified Gagliardo–Nirenberg inequality (2.8):

$$\begin{aligned}
\left\|u^\varepsilon(t)-\mathtt{V}^\varepsilon(t)\left(fe^{ig(t)}\right)\right\|_{L^{2p'}}&\lesssim|\cos t|^{-\delta(2p')}\left\|u^\varepsilon(t)-\mathtt{V}^\varepsilon(t)\left(fe^{ig(t)}\right)\right\|_{L^2}^{1-\delta(2p')}\\
&\quad\times\left\|J^\varepsilon(t)\left(u^\varepsilon(t)-\mathtt{V}^\varepsilon(t)\left(fe^{ig(t)}\right)\right)\right\|_{L^2}^{\delta(2p')}.
\end{aligned}$$

The first L^2-norm was estimated just above. For the second one, notice that

$$J^\varepsilon(t)\mathcal{U}^\varepsilon(t)=-i\,\mathcal{U}^\varepsilon(t)\nabla_x,\qquad J^\varepsilon(t)\mathtt{V}^\varepsilon(t)=-i\,\mathtt{V}^\varepsilon(t)\nabla_x;$$

therefore

$$\begin{aligned}
\left\|J^\varepsilon(t)\Big(u^\varepsilon(t)-\mathtt{V}^\varepsilon(t)\Big(fe^{ig(t)}\Big)\Big)\right\|_{L^2}&\lesssim\left\|\mathcal{U}^\varepsilon(t)\nabla\left(b^\varepsilon(t)e^{ig(t)}\right)-\mathtt{V}^\varepsilon(t)\nabla\left(fe^{ig(t)}\right)\right\|_{L^2}\\
&\lesssim\left\|\nabla\left(b^\varepsilon(t)e^{ig(t)}-fe^{ig(t)}\right)\right\|_{L^2}+\left\|(\mathcal{U}^\varepsilon(t)-\mathtt{V}^\varepsilon(t))\,\nabla\left(fe^{ig(t)}\right)\right\|_{L^2}\\
&\lesssim y_\varepsilon(t)+|\varepsilon\tan t|^\theta,
\end{aligned}$$

where we have used Lemmas 4.4 and 4.7. We infer that

$$\left\|u^\varepsilon(t)-\mathtt{V}^\varepsilon(t)\left(fe^{ig(t)}\right)\right\|_{L^{2p'}}\lesssim|\cos t|^{-\delta(2p')}\left(y_\varepsilon(t)+|\varepsilon\tan t|^\theta\right).$$

We have explicitly

$$\left\|\mathtt{V}^\varepsilon(t)\left(fe^{ig(t)}\right)\right\|_{L^{2p'}}=|\cos t|^{-\delta(2p')}\|f\|_{L^{2p'}}\lesssim|\cos t|^{-\delta(2p')}.$$

Proceeding as above, we have

$$\|u^\varepsilon(t)\|_{L^{2p'}}\lesssim|\cos t|^{-\delta(2p')}\|u^\varepsilon\|_{L^2}^{1-\delta(2p')}\|J^\varepsilon(t)u^\varepsilon\|_{L^2}^{\delta(2p')},$$

with $\|J^\varepsilon(t)u^\varepsilon\|_{L^2}\lesssim\|b^\varepsilon(t)-f\|_{H^1}+\|f\|_{H^1}$. These estimates will eventually lead to an inequality of the form $y_\varepsilon'(t)\le a(t)y_\varepsilon(t)+b(t)y_\varepsilon(t)^\kappa+c(t)$ for some $\kappa>1$. To avoid that situation, we proceed as in section 3; there exists $t^\varepsilon>0$ such that

$$\|b^\varepsilon(t)\|_{H^1}\le 2\|f\|_{H^1}\tag{4.13}$$

for $t \in [0, t^\varepsilon]$. So long as (4.13) holds, we have from the above estimates

$$\|\partial_t b^\varepsilon(t)\|_{L^2} \lesssim |\cos t|^{-2\delta(2p')} \left(y_\varepsilon(t) + |\varepsilon \tan t|^\theta\right). \tag{4.14}$$

To prove that (4.13) holds up to time $\pi/2 - \Lambda\varepsilon$ for $0 < \varepsilon \leq \varepsilon_\Lambda$ along with the error estimate (4.7), we estimate the L^2-norm of $\nabla_x \partial_t b^\varepsilon$. From (4.5) and (4.6),

$$\begin{aligned}\nabla_x \partial_t b^\varepsilon(t) &= -i\nabla_x g(t) \partial_t b^\varepsilon(t) \\ &-ie^{-ig(t)} \nabla_x \left(\mathcal{U}^\varepsilon(-t) \left(\left(|x|^{-\gamma} * |u^\varepsilon|^2 \right) u^\varepsilon \right)(t) - \frac{1}{|\cos t|^\gamma} \left(|x|^{-\gamma} * |f|^2 \right) a^\varepsilon(t) \right).\end{aligned}$$

The first term is controlled thanks to Lemma 4.7 and (4.14). For the other term, we notice that since $\mathcal{U}^\varepsilon$ is unitary on L^2, from (2.5) its L^2-norm is equal to

$$\left\| J^\varepsilon(t) \left(\left(|x|^{-\gamma} * |u^\varepsilon|^2 \right) u^\varepsilon \right)(t) + \frac{i}{|\cos t|^\gamma} \mathcal{U}^\varepsilon(t) \nabla_x \left(\left(|x|^{-\gamma} * |f|^2 \right) a^\varepsilon(t) \right) \right\|_{L^2}.$$

We proceed as before: we first replace $\mathcal{U}^\varepsilon$ with $\mathtt{V}^\varepsilon$ in the last term, up to an error of $|\cos t|^{-\gamma}$ times:

$$\begin{aligned}&\left\| (\mathcal{U}^\varepsilon(t) - \mathtt{V}^\varepsilon(t)) \nabla_x \left(\left(|x|^{-\gamma} * |f|^2 \right) a^\varepsilon \right) \right\|_{L^2} \lesssim \\ &\qquad \lesssim \left\| (\mathcal{U}^\varepsilon(t) - \mathtt{V}^\varepsilon(t)) \nabla_x \left(\left(|x|^{-\gamma} * |f|^2 \right) (b^\varepsilon - f) e^{ig} \right) \right\|_{L^2} \\ &\qquad\quad + \left\| (\mathcal{U}^\varepsilon(t) - \mathtt{V}^\varepsilon(t)) \nabla_x \left(\left(|x|^{-\gamma} * |f|^2 \right) f e^{ig} \right) \right\|_{L^2}.\end{aligned}$$

For the first term, we do not use Lemma 4.4, but roughly the fact that $\mathcal{U}^\varepsilon$ and $\mathtt{V}^\varepsilon$ are unitary on L^2. It is not larger than

$$2 \left\| \nabla_x \left(\left(|x|^{-\gamma} * |f|^2 \right) (b^\varepsilon - f) e^{ig} \right) \right\|_{L^2} \lesssim \|b^\varepsilon(t) - f\|_{H^1},$$

from Lemma 4.7. The second term is controlled thanks to Lemmas 4.4 and 4.7:

$$\left\| (\mathcal{U}^\varepsilon(t) - \mathtt{V}^\varepsilon(t)) \nabla_x \left(\left(|x|^{-\gamma} * |f|^2 \right) f e^{ig} \right) \right\|_{L^2} \lesssim |\varepsilon \tan t|^\theta.$$

We now have, so long as (4.13) holds,

$$\begin{aligned}&\|\partial_t b^\varepsilon(t)\|_{H^1} \lesssim |\cos t|^{-2\delta(2p')} \left(y_\varepsilon(t) + |\varepsilon \tan t|^\theta \right) \\ &+ \left\| J^\varepsilon(t) \left(\left(|x|^{-\gamma} * |u^\varepsilon|^2 \right) u^\varepsilon \right) + \frac{i}{|\cos t|^\gamma} \mathtt{V}^\varepsilon(t) \nabla_x \left(\left(|x|^{-\gamma} * |f|^2 \right) a^\varepsilon(t) \right) \right\|_{L^2}.\end{aligned} \tag{4.15}$$

Using the identity $J^\varepsilon(t) \mathtt{V}^\varepsilon(t) = -i\, \mathtt{V}^\varepsilon(t) \nabla_x$ and (4.10), we have to estimate

$$\begin{aligned}&\left\| J^\varepsilon(t) \left(\left(\left(|x|^{-\gamma} * |u^\varepsilon|^2 \right) u^\varepsilon \right) - \frac{1}{|\cos t|^\gamma} \mathtt{V}^\varepsilon(t) \left(\left(|x|^{-\gamma} * |f|^2 \right) a^\varepsilon(t) \right) \right) \right\|_{L^2} \\ &= \left\| J^\varepsilon(t) \left(\left(|x|^{-\gamma} * |u^\varepsilon|^2 \right) u^\varepsilon - \left(|x|^{-\gamma} * |\mathtt{V}^\varepsilon(t) f|^2 \right) \mathtt{V}^\varepsilon(t) a^\varepsilon \right) \right\|_{L^2} \\ &\lesssim \left\| \left(|x|^{-\gamma} * |u^\varepsilon|^2 \right) J^\varepsilon(t) u^\varepsilon - \left(|x|^{-\gamma} * |\mathtt{V}^\varepsilon(t) f|^2 \right) J^\varepsilon(t) \mathtt{V}^\varepsilon(t) a^\varepsilon \right\|_{L^2} \\ &\quad + |\cos t| \left\| \nabla_x \left(|x|^{-\gamma} * |u^\varepsilon|^2 \right) u^\varepsilon - \nabla_x \left(|x|^{-\gamma} * |\mathtt{V}^\varepsilon(t) f|^2 \right) \mathtt{V}^\varepsilon(t) a^\varepsilon \right\|_{L^2}.\end{aligned} \tag{4.16}$$

We replace $\mathtt{V}^\varepsilon$ with $\mathcal{U}^\varepsilon$ in the first term of the right-hand side, up to the error

$$\begin{aligned}&\left\| \left(|x|^{-\gamma} * |\mathtt{V}^\varepsilon(t) f|^2 \right) J^\varepsilon(t) \left(\mathtt{V}^\varepsilon(t) - \mathcal{U}^\varepsilon(t) \right) a^\varepsilon \right\|_{L^2} \\ &\qquad \lesssim \left(\|\mathtt{V}^\varepsilon(t) f\|_{L^2}^2 + \|\mathtt{V}^\varepsilon(t) f\|_{L^{2p'}}^2 \right) \|(\mathtt{V}^\varepsilon(t) - \mathcal{U}^\varepsilon(t)) \nabla_x a^\varepsilon\|_{L^2} \\ &\qquad \lesssim |\cos t|^{-2\delta(2p')} \left\| (\mathtt{V}^\varepsilon(t) - \mathcal{U}^\varepsilon(t)) \nabla_x \left((b^\varepsilon - f) e^{ig} \right) \right\|_{L^2} \\ &\qquad\quad + |\cos t|^{-2\delta(2p')} \left\| (\mathtt{V}^\varepsilon(t) - \mathcal{U}^\varepsilon(t)) \nabla_x \left(f e^{ig} \right) \right\|_{L^2} \\ &\qquad \lesssim |\cos t|^{-2\delta(2p')} \left(\|b^\varepsilon(t) - f\|_{H^1} + |\varepsilon \tan t|^\theta \right),\end{aligned}$$

from the above computation. Therefore, the first term of the right-hand side of (4.16) is estimated by

$$|\cos t|^{-2\delta(2p')}\left(y_\varepsilon(t)+|\varepsilon\tan t|^\theta\right)+\left\|\left(|x|^{-\gamma}*\left(|u^\varepsilon|^2-|\mathtt{V}^\varepsilon(t)f|^2\right)\right)J^\varepsilon(t)u^\varepsilon\right\|_{L^2}\,.$$

So long as (4.13) holds, $\|J^\varepsilon(t)u^\varepsilon\|_{L^2}\lesssim 1$, and the last term is estimated by

$$\left\||x|^{-\gamma}*\left(|u^\varepsilon|^2-|\mathtt{V}^\varepsilon(t)f|^2\right)\right\|_{L^\infty}\,,$$

which already appeared above and was estimated in (4.12). We are left with the second term of the right-hand side of (4.16). Using Lemma 4.6 with q instead of p now, we see

$$\begin{aligned}\left\|\nabla_x\left(|x|^{-\gamma}*|\mathtt{V}^\varepsilon(t)f|^2\right)\left(\mathtt{V}^\varepsilon(t)-\mathcal{U}^\varepsilon(t)\right)a^\varepsilon\right\|_{L^2}&\lesssim\\ &\lesssim\left(\|\mathtt{V}^\varepsilon(t)f\|_{L^2}^2+\|\mathtt{V}^\varepsilon(t)f\|_{L^{2q'}}^2\right)\|(\mathtt{V}^\varepsilon(t)-\mathcal{U}^\varepsilon(t))\,a^\varepsilon\|_{L^2}\\ &\lesssim|\cos t|^{-2\delta(2q')}\left(y_\varepsilon(t)+|\varepsilon\tan t|^\theta\right)\,.\end{aligned}$$

The final term to estimate is

$$\begin{aligned}\left\|\nabla_x\left(|x|^{-\gamma}*\left(|u^\varepsilon|^2-|\mathtt{V}^\varepsilon(t)f|^2\right)\right)u^\varepsilon\right\|_{L^2}\lesssim&\left\||u^\varepsilon|^2-|\mathtt{V}^\varepsilon(t)f|^2\right\|_{L^1}\\ &+\left\||u^\varepsilon|^2-|\mathtt{V}^\varepsilon(t)f|^2\right\|_{L^{q'}}\,.\end{aligned}$$

The right-hand side was already estimated in (4.12) with p instead of q. We finally have, so long as (4.13) holds,

$$y'(t)\lesssim\left(|\cos t|^{-2\delta(2p')}+|\cos t|^{1-2\delta(2p')}\right)\left(y_\varepsilon(t)+|\varepsilon\tan t|^\theta\right).$$

Now recall that given $\delta>0$, $\delta(2p')$ and $\delta(2q')$ are explicit; hence

$$y'(t)\lesssim|\cos t|^{-\gamma-\frac{\delta}{2}}\left(y_\varepsilon(t)+|\varepsilon\tan t|^\theta\right).$$

It is now time to fix θ. In view of (4.7), it is natural to take $\theta=1-\gamma-\delta$. This yields, so long as (4.13) holds,

$$y_\varepsilon'(t)\lesssim|\cos t|^{-\gamma-\frac{\delta}{2}}\left(y_\varepsilon(t)+|\varepsilon\tan t|^{1-\gamma-\delta}\right)\lesssim|\cos t|^{-\gamma-\frac{\delta}{2}}y_\varepsilon(t)+\frac{\varepsilon^{1-\gamma-\delta}}{|\cos t|^{1-\frac{\delta}{2}}}\,. \tag{4.17}$$

The maps $t\mapsto|\cos t|^{-\gamma-\frac{\delta}{2}}$ and $t\mapsto|\cos t|^{-1+\frac{\delta}{2}}$ are locally integrable (we can assume $\gamma+\delta/2<1-\delta/2$; otherwise (4.7) is of no interest). From the Gronwall lemma, so long as (4.13) holds, we infer

$$y_\varepsilon(t)\lesssim\varepsilon^{1-\gamma-\delta}\,. \tag{4.18}$$

Therefore, there exists $\varepsilon_\Lambda>0$ such that for $0<\varepsilon\leq\varepsilon_\Lambda$, (4.13) holds up to time $\pi/2-\Lambda\varepsilon$, with (4.18). The estimate for $x\partial_t b^\varepsilon$ then is easy, and we leave out this part; this proves (4.7).

Remark 4.8. One might believe that we could deduce Proposition 4.5 in one shot from (4.17) and wonder why we split the proof into three steps. The reason is that we cannot apply Lemma 4.4 (which was used to get (4.17)) near $t=\pi/2$. On the other hand, we will see below that computations near $t=\pi/2$ are far simpler.

4.3. Near the focus and beyond. Keep $\Lambda, \delta > 0$ fixed. We prove that there exists $C_{\Lambda,\delta}$ such that

$$\int_{\frac{\pi}{2}-\Lambda\varepsilon}^{\frac{\pi}{2}+\Lambda\varepsilon} \|\partial_t b^\varepsilon(t)\|_\Sigma dt \leq C_{\Lambda,\delta}\varepsilon^{1-\gamma-\delta}\,. \tag{4.19}$$

A rough estimate in (4.8) yields

$$\begin{aligned}\|\partial_t b^\varepsilon(t)\|_{L^2} &\lesssim \left\|\left(|x|^{-\gamma} * |u^\varepsilon|^2\right) u^\varepsilon(t)\right\|_{L^2} + \frac{1}{|\cos t|^\gamma}\left\|\left(|x|^{-\gamma} * |f|^2\right) a^\varepsilon(t)\right\|_{L^2} \\ &\lesssim \left(\|u^\varepsilon(t)\|_{L^2}^2 + \|u^\varepsilon(t)\|_{L^{2p'}}^2\right)\|u^\varepsilon(t)\|_{L^2} + \frac{1}{|\cos t|^\gamma}\|u^\varepsilon(t)\|_{L^2}\,.\end{aligned} \tag{4.20}$$

The conservation of mass yields $\|u^\varepsilon(t)\|_{L^2} = \|f\|_{L^2}$. The conservations of mass and energy (2.3) yield, along with Gagliardo–Nirenberg inequalities,

$$\|u^\varepsilon(t)\|_{L^{2p'}} \lesssim \varepsilon^{-\delta(2p')}\,.$$

Using this estimate (which is sharp near the focus and only near the focus) and integrating (4.20), we get

$$\begin{aligned}\int_{\frac{\pi}{2}-\Lambda\varepsilon}^{\frac{\pi}{2}+\Lambda\varepsilon} \|\partial_t b^\varepsilon(t)\|_{L^2} dt &\lesssim \Lambda\varepsilon^{1-2\delta(2p')} + \int_{\frac{\pi}{2}-\Lambda\varepsilon}^{\frac{\pi}{2}+\Lambda\varepsilon} \frac{dt}{|\cos t|^\gamma} \\ &\lesssim \varepsilon^{1-\gamma-\frac{\delta}{2}} + \varepsilon^{1-\gamma}\,.\end{aligned}$$

The term $\|x\partial_t b^\varepsilon(t)\|_{L^2}$ is estimated the same way, since the conservation of energy yields an a priori bound for $H^\varepsilon u^\varepsilon$. For $\|\nabla_x \partial_t b^\varepsilon(t)\|_{L^2}$, we proceed as in section 3, (3.10) to get an estimate from the Gronwall lemma; the details are left to the reader.

Finally, one can prove that there exists $C_{\Lambda,\delta}$ such that

$$\int_{\frac{\pi}{2}+\Lambda\varepsilon}^{\pi} \|\partial_t b^\varepsilon(t)\|_\Sigma dt \leq C_{\Lambda,\delta}\varepsilon^{1-\gamma-\delta}$$

by mimicking the computations performed in section 4.2, and the proof of Proposition 4.5 is complete.

5. Linear propagation and nonlinear focus. We now consider the case where $\alpha = \gamma > 1$ in (1.6). Our results are similar to those of [6]. Before stating the main result, we recall some points of the scattering theory for (1.9).

PROPOSITION 5.1 (see [13, 17]). *Assume $\psi_- \in \Sigma$ and $1 < \gamma < \min(4, n)$. If $\gamma > 4/3$ or if $\|\psi_-\|_\Sigma$ is sufficiently small, then we have the following:*

- *There exists a unique $\psi \in C(\mathbb{R}_t, \Sigma)$ which is solution to (1.9), such that*

$$\lim_{t\to-\infty} \|\psi_- - \mathtt{U}(-t)\psi(t)\|_\Sigma = 0\,, \quad \text{where } \mathtt{U}(t) = e^{i\frac{t}{2}\Delta}\,.$$

- *There exists a unique $\psi_+ \in \Sigma$ such that*

$$\lim_{t\to+\infty} \|\psi_+ - \mathtt{U}(-t)\psi(t)\|_\Sigma = 0\,.$$

The scattering operator is defined as the map $S\colon \psi_- \mapsto \psi_+$.

Our main result in this section follows.

PROPOSITION 5.2. *Suppose* $n \geq 2$. *Let* $f \in \Sigma$, $1 < \gamma = \alpha < \min(4, n)$, *and* $k \in \mathbb{N}$. *Assume either* $\gamma > 4/3$ *or* $\|f\|_\Sigma$ *is sufficiently small. Then the asymptotic behavior of* u^ε *for* $\pi/2 + (k-1)\pi < a \leq b < \pi/2 + k\pi$ *is given by*

$$\sup_{a\leq t\leq b} \left\| A^\varepsilon(t) \left(u^\varepsilon(t,x) - \frac{e^{-ink\frac{\pi}{2}}}{|\cos t|^{n/2}} \left(\mathcal{F}\circ S^k \circ \mathcal{F}^{-1}\right) f\left(\frac{x}{\cos t}\right) e^{-i\frac{x^2}{2\varepsilon}\tan t} \right) \right\|_{L^2_x} \underset{\varepsilon\to 0}{\longrightarrow} 0\,,$$

where A^ε *is one of the operators* Id, J^ε, *or* H^ε, *and* S^k *denotes the* k*th iterate of* S *(which is well defined under our assumptions on* f*). At the foci,*

$$\left\| B^\varepsilon \left(u^\varepsilon\left(\frac{\pi}{2} + k\pi\right) - \frac{e^{-ink\frac{\pi}{2}}}{\varepsilon^{n/2}} \left(\mathcal{F}\circ S^k\right) f\left(\frac{\cdot}{\varepsilon}\right) \right) \right\|_{L^2} \underset{\varepsilon\to 0}{\longrightarrow} 0\,,$$

where B^ε *is one of the operators* Id, $\frac{x}{\varepsilon}$, *or* $\varepsilon\nabla_x$.

With Lemma 4.4 in mind, this shows that nonlinear effects are negligible away from foci, while they have an influence at leading order near the foci: each caustic crossing is described in average by the nonlinear scattering operator S (the phase shift $e^{-ink\frac{\pi}{2}}$ is the Maslov index, present in the linear case [9]).

The proof of Proposition 5.2 is very similar to the one in [6], which relies on (scaled) Strichartz estimates. We will refrain from repeating everything in detail and limit ourselves to proving the main technical proposition and presenting an outline for the rest of the proof. One main difference compared to the problem in [6] is the action of the operators $J^\varepsilon(t)$, $H^\varepsilon(t)$ on the Hartree nonlinearity as described by (2.9).

We start by reformulating (1.6) by Duhamel's formula:

$$\begin{aligned} u^\varepsilon(t) = \mathcal{U}^\varepsilon(t-t_0)u_0^\varepsilon - i\varepsilon^{\gamma-1}\int_{t_0}^t \mathcal{U}^\varepsilon(t-s)F^\varepsilon(u^\varepsilon)(s)ds \\ - i\varepsilon^{-1}\int_{t_0}^t \mathcal{U}^\varepsilon(t-s)h^\varepsilon(s)ds. \end{aligned} \tag{5.1}$$

This equation generalizes (1.6) to the case of an additional source term and a general nonlinear term F^ε. The main technical result which is used throughout the proof of Proposition 5.2 is as follows.

PROPOSITION 5.3. *Let* $t_1 > t_0$, *with* $|t_1 - t_0| \leq \pi$. *Let* $q, r, s, k \in [1, \infty]$ *be such that*

$$\begin{cases} \text{(a)} \quad \dfrac{1}{r'} = \dfrac{1}{r} + \dfrac{2}{s} + \dfrac{\gamma}{n} - 1 \quad \textit{and} \quad s < \dfrac{2n}{n-\gamma}, \\ \text{(b)} \quad \dfrac{1}{q'} = \dfrac{1}{q} + \dfrac{2}{k}, \\ \text{(c)} \quad (q,r) \textit{ is an admissible pair}, \\ \text{(d)} \quad 0 < \dfrac{1}{k} < \delta(s) < 1\,. \end{cases} \tag{5.2}$$

Assume that there exists a constant C *independent of* t *and* ε *such that for* $t_0 \leq t \leq t_1$,

$$\|F^\varepsilon(u^\varepsilon)(t)\|_{L^{r'}_x} \leq \frac{C}{(|\cos t| + \varepsilon)^{2\delta(s)}} \|u^\varepsilon(t)\|_{L^r_x}, \tag{5.3}$$

and define

$$A^\varepsilon(t_0,t_1) := \left(\int_{t_0}^{t_1} \frac{dt}{(|\cos t| + \varepsilon)^{k\delta(s)}} \right)^{2/k}.$$

Then there exists C^* independent of ε, t_0, and t_1 such that for any admissible pair (ρ,σ),

$$\begin{aligned} \|u^\varepsilon\|_{L^q(t_0,t_1;L^r)} \le{}& C^*\varepsilon^{-1/q}\|u_0^\varepsilon\|_{L^2} + C_{q,\rho}\varepsilon^{-1-\frac{1}{q}-\frac{1}{\rho}}\|h^\varepsilon\|_{L^{\rho'}(t_0,t_1;L^{\sigma'})} \\ &+ C^*\varepsilon^{2\left(\delta(s)-\frac{1}{k}\right)}A^\varepsilon(t_0,t_1)\|u^\varepsilon\|_{L^q(t_0,t_1;L^r)}. \end{aligned} \tag{5.4}$$

Mostly the following corollary is applied.

COROLLARY 5.4. *Suppose the assumptions of Proposition* 5.3 *are satisfied. Assume moreover that $C^*\varepsilon^{2\left(\delta(s)-\frac{1}{k}\right)}A^\varepsilon(t_0,t_1) \le 1/2$, which holds in either of the following two cases:*

- *$0 \le t_0 \le t_1 \le \frac{\pi}{2} - \Lambda\varepsilon$, with $\Lambda \ge \Lambda_0$ sufficiently large,*
- *$t_0, t_1 \in [\frac{\pi}{2} - \Lambda\varepsilon, \frac{\pi}{2} + \Lambda\varepsilon]$, with $\frac{t_1-t_0}{\varepsilon} \le \eta$ sufficiently small.*

Then

$$\|u^\varepsilon\|_{L^\infty(t_0,t_1;L^2)} \le C\|u_0^\varepsilon\|_{L^2} + C_{q,\rho}\varepsilon^{-1-\frac{1}{\rho}}\|h^\varepsilon\|_{L^{\rho'}(t_0,t_1;L^{\sigma'})}. \tag{5.5}$$

To prove Proposition 5.3, we first prove the following algebraic lemma.

LEMMA 5.5. *Let $n \ge 2$, and assume $1 < \gamma < \min(4,n)$. Then there exist $q,r,s,k \in [1,\infty]$ satisfying the conditions* (5.2).

Proof. Note that (a) is equivalent to demanding $\gamma/2 = \delta(r)+\delta(s)$ and $\gamma/2 > \delta(s)$.

Case $\gamma \le 2$. Suppose $\gamma/2 = \delta(s)$. Then by the first half of (a), $\delta(r) = 0$, and $(q,r) = (\infty,2)$ by (c). With $k = 2$, (b) and (d) are satisfied. Now choose s such that $1/2 < \delta(s) < \gamma/2$, but close enough to $\gamma/2$ for (5.2) still to be valid by continuity (for example, $\delta(s) = \frac{1}{2} + \frac{1}{2}(\frac{\gamma}{2} - \frac{1}{2})$). Then (5.2) is satisfied.

Case $\gamma > 2$. In this case take s such that $\delta(s) = 1$, e.g., $s = \frac{2n}{n-2}$. Up to a continuity argument as in the previous case, $\delta(s) < 1$, and (5.2) is satisfied. ∎

Proof of Proposition 5.3. Application of the (scaled) Strichartz estimates (Proposition 2.2) to (5.1) yields

$$\begin{aligned} \|u^\varepsilon\|_{L^q(t_0,t_1;L^r)} \le{}& C\varepsilon^{-1/q}\|u_0^\varepsilon\|_{L^2} + C_{q,\rho}\varepsilon^{-1-\frac{1}{q}-\frac{1}{\rho}}\|h^\varepsilon\|_{L^{\rho'}(t_0,t_1;L^{\sigma'})} \\ &+ C\varepsilon^{\gamma-1-\frac{2}{q}}\|F^\varepsilon(u^\varepsilon)\|_{L^{q'}(t_0,t_1;L^{r'})}. \end{aligned}$$

Then by the assumptions on $F^\varepsilon(u^\varepsilon)$, after an application of Hölder inequality in time, the statement follows. ∎

Proof of Corollary 5.4. The additional assumption implies that the last term in (5.4) can be absorbed by the left-hand side, and we get

$$\|u^\varepsilon\|_{L^q(t_0,t_1;L^r)} \le C\varepsilon^{-1/q}\|u_0^\varepsilon\|_{L^2} + C\varepsilon^{-1-\frac{1}{q}-\frac{1}{\rho}}\|h^\varepsilon\|_{L^{\rho'}(t_0,t_1;L^{\sigma'})}. \tag{5.6}$$

Another application of Strichartz estimates to (5.1), with indices $(\infty,2)$ on the left and (ρ,σ), respectively (q,r), on the right, yields

$$\begin{aligned} \|u^\varepsilon\|_{L^\infty(t_0,t_1;L^2)} \le{}& C\|u_0^\varepsilon\|_{L^2} + C\varepsilon^{-1-\frac{1}{\rho}}\|h^\varepsilon\|_{L^{\rho'}(t_0,t_1;L^{\sigma'})} \\ &+ C\varepsilon^{\gamma-1-\frac{1}{q}}\|F^\varepsilon(u^\varepsilon)\|_{L^{q'}(t_0,t_1;L^{r'})}. \end{aligned}$$

As before,

$$\begin{aligned}\varepsilon^{\gamma-1-\frac{1}{q}}\|F^\varepsilon(u^\varepsilon)\|_{L^{q'}(t_0,t_1;L^{r'})} &\leq C\varepsilon^{\frac{1}{q}}\varepsilon^{2\left(\delta(s)-\frac{1}{k}\right)}A^\varepsilon(t_0,t_1)\|u^\varepsilon\|_{L^q(t_0,t_1;L^r)}\\ &\leq C\varepsilon^{\frac{1}{q}}\|u^\varepsilon\|_{L^q(t_0,t_1;L^r)},\end{aligned}$$

and the statement now follows from (5.6). □

The proof of Proposition 5.2 consists of three parts: the propagation before the focus, the matching between the two regimes, and proof that near the focus, the harmonic potential is negligible. In all parts the main tool used to derive the major statements is Proposition 5.3. Since the proof is very similar to the one in [6], we do not repeat everything in detail but give a detailed proof only for the first part to show how the methods of [6] are applied.

We now show the proof for the propagation before the focus, that is, the approximation of $u^\varepsilon(t)$ by $u^\varepsilon_{\rm free}(t)$ for $0\leq t\leq \frac{\pi}{2}-\Lambda\varepsilon$, in the limit $\Lambda\to+\infty$. We prove that

$$\limsup_{\varepsilon\to 0}\sup_{0\leq t\leq\frac{\pi}{2}-\Lambda\varepsilon}\left\|A^\varepsilon(t)\left(u^\varepsilon(t,x)-u^\varepsilon_{\rm free}(t,x)\right)\right\|_{L^2_x}\underset{\Lambda\to+\infty}{\longrightarrow}0\,,$$

with $A^\varepsilon(t)$ being one of the operators Id, J^ε, or H^ε.

Define the remainder $w^\varepsilon=u^\varepsilon-u^\varepsilon_{\rm free}$. It solves

$$\begin{cases} i\varepsilon\partial_t w^\varepsilon+\dfrac{1}{2}\varepsilon^2\Delta w^\varepsilon=V(x)w^\varepsilon+\varepsilon^\gamma\left(|x|^{-\gamma}*|u^\varepsilon|^2\right)u^\varepsilon\,,\\ w^\varepsilon_{|t=0}=0\,.\end{cases}$$

From Duhamel's principle, this can be written as

$$w^\varepsilon(t)=\mathcal{U}^\varepsilon(t)r^\varepsilon-i\varepsilon^{\gamma-1}\int_0^t\mathcal{U}^\varepsilon(t-s)\left(|x|^{-\gamma}*|u^\varepsilon|^2\right)u^\varepsilon(s)ds. \tag{5.7}$$

Since $u^\varepsilon_{\rm free}$ solves the linear equation (1.7), so does $J^\varepsilon(t)u^\varepsilon_{\rm free}$ from (2.6), and

$$\|u^\varepsilon_{\rm free}(t)\|_{L^2}=\|f\|_{L^2},\qquad \|J^\varepsilon(t)u^\varepsilon_{\rm free}\|_{L^2}=\|\nabla f\|_{L^2}.$$

From the Sobolev inequality (2.8),

$$\|u^\varepsilon_{\rm free}(t)\|_{L^s}\leq\frac{C}{|\cos t|^{\delta(s)}}\|f\|_{L^2}^{1-\delta(s)}\|\nabla f\|_{L^2}^{\delta(s)}$$

for any $s\in[2,\frac{2n}{n-2}[$. Therefore there exists C_0 such that

$$\|u^\varepsilon_{\rm free}(t)\|_{L^s}\leq\frac{C_0}{|\cos t|^{\delta(s)}}\ . \tag{5.8}$$

From Proposition 2.3, for fixed $\varepsilon>0$, $u^\varepsilon\in C(\mathbb{R},\Sigma)$, and the same obviously holds for $u^\varepsilon_{\rm free}$. Therefore, $w^\varepsilon\in C(\mathbb{R},\Sigma)$, and there exists $t^\varepsilon>0$ such that

$$\|w^\varepsilon(t)\|_{L^s}\leq\frac{C_0}{|\cos t|^{\delta(s)}} \tag{5.9}$$

for any $t\in[0,t^\varepsilon]$. So long as (5.9) holds, we have

$$\|u^\varepsilon(t)\|_{L^s}\leq\frac{2C_0}{|\cos t|^{\delta(s)}},$$

and we can apply Proposition 5.3.

Take $h^\varepsilon = \varepsilon^\gamma \left(|x|^{-\gamma} * |u^\varepsilon|^2\right) u^\varepsilon_{\rm free}$ and $F^\varepsilon(w^\varepsilon) = \left(|x|^{-\gamma} * |u^\varepsilon|^2\right) w^\varepsilon$ and let $q, k, r, s \in [1, \infty]$ satisfy the assumptions of Proposition 5.3. Now by Hölder's inequality,

$$\|F^\varepsilon(w^\varepsilon)(t)\|_{L^{r'}} \leq \left\||x|^{-\gamma} * |u^\varepsilon(t)|^2\right\|_{L^\beta} \|w^\varepsilon(t)\|_{L^r},$$

with β such that $\frac{1}{r'} = \frac{1}{r} + \frac{1}{\beta}$. By the Hardy–Littlewood–Sobolev inequality and the above estimate,

$$\begin{aligned}\|F^\varepsilon(w^\varepsilon)(t)\|_{L^{r'}} &\lesssim \|u^\varepsilon(t)\|^2_{L^s} \|w^\varepsilon(t)\|_{L^r} \\ &\lesssim \frac{(2C_0)^2}{|\cos t|^{2\delta(s)}} \|w^\varepsilon(t)\|_{L^r}.\end{aligned}$$

Note that the second statement of (5.2)(a) ensures that $s, \beta \in (1, \infty)$ so the Hardy–Littlewood–Sobolev inequality is applicable here. Assume (5.9) holds for $0 \leq t \leq T^\varepsilon$. If $0 \leq t \leq T^\varepsilon \leq \frac{\pi}{2} - \Lambda\varepsilon$, then $\varepsilon \lesssim \cos t$, and the above estimate shows that F^ε satisfies assumption (5.3).

From Corollary 5.4, if Λ is sufficiently large, we get for $0 \leq t \leq T^\varepsilon \leq \frac{\pi}{2} - \Lambda\varepsilon$

$$\|w^\varepsilon\|_{L^\infty(0,T;L^2)} \leq C_\sigma \varepsilon^{\gamma - 1 - \frac{1}{\rho}} \left\|(|x|^{-\gamma} * |u^\varepsilon|^2) u^\varepsilon_{\rm free}\right\|_{L^{\rho'}(0,T;L^{\sigma'})}$$

for any admissible (ρ, σ). Now take $(\rho, \sigma) = (q, r)$ and proceed as above in space and apply Hölder inequality in time:

$$\left\|\left(|x|^{-\gamma} * |u^\varepsilon|^2\right) u^\varepsilon_{\rm free}\right\|_{L^{q'}(0,T;L^{r'})} \leq C_{\gamma,n} \|u^\varepsilon\|^2_{L^k(0,T;L^s)} \|u^\varepsilon_{\rm free}\|_{L^q(0,T;L^r)}.$$

The first term of the right-hand side is estimated through (5.8) and (5.9):

$$\|u^\varepsilon\|^2_{L^k(0,T;L^s)} \leq \frac{C}{\left(\frac{\pi}{2} - T\right)^{2(\delta(s) - 1/k)}}.$$

The last term is estimated the same way, for (5.8) still holds when replacing s with r:

$$\|u^\varepsilon_{\rm free}\|_{L^q(0,T;L^r)} \leq \frac{C}{\left(\frac{\pi}{2} - T\right)^{\delta(r) - 1/q}}.$$

We infer

$$\left\|\left(|x|^{-\gamma} * |u^\varepsilon|^2\right) u^\varepsilon_{\rm free}\right\|_{L^{q'}(0,T;L^{r'})} \leq \frac{C}{\left(\frac{\pi}{2} - T\right)^{\gamma - 1 - \frac{1}{q}}};$$

thus

$$\|w^\varepsilon\|_{L^\infty(0,T;L^2)} \leq C \left(\frac{\varepsilon}{\frac{\pi}{2} - T}\right)^{\gamma - 1 - \frac{1}{q}}. \tag{5.10}$$

Now apply the operator J^ε to (5.7). This yields, since J^ε and $\mathcal{U}^\varepsilon$ commute,

$$J^\varepsilon(t) w^\varepsilon = \mathcal{U}^\varepsilon(t) J^\varepsilon(0) r^\varepsilon - i\varepsilon^{\gamma - 1} \int_0^t \mathcal{U}^\varepsilon(t - s) J^\varepsilon(s) \left(\left(|x|^{-\gamma} * |u^\varepsilon|^2\right) u^\varepsilon\right)(s) ds.$$

The action of J^ε on the nonlinear term is described by (2.9). In order to apply Proposition 5.3 as before, we now take

$$h^\varepsilon = \varepsilon^{\gamma-1}\left(|x|^{-\gamma} * |u^\varepsilon|^2\right) J^\varepsilon(t) u^\varepsilon_{\rm free} + \varepsilon^{\gamma-1}\ 2{\rm Re}\Big(|x|^{-\gamma} * (\overline{u^\varepsilon} J^\varepsilon(t) u^\varepsilon_{\rm free})\Big) u^\varepsilon$$

and

$$F^\varepsilon(w^\varepsilon) = \left(|x|^{-\gamma} * |u^\varepsilon|^2\right) J^\varepsilon(t) w^\varepsilon + \ 2{\rm Re}\Big(|x|^{-\gamma} * (\overline{u^\varepsilon} J^\varepsilon(t) w^\varepsilon)\Big) u^\varepsilon\,. \tag{5.11}$$

The first term on the right-hand side of (5.11) leads to an equation which is very similar to (5.7), with w^ε replaced by $J^\varepsilon w^\varepsilon$, and is treated by the same computations as above. For the second term, we estimate by Hölder, by the Hardy–Littlewood–Sobolev inequality, and then again by Hölder:

$$\begin{aligned}\left\|2{\rm Re}\Big(|x|^{-\gamma} * (\overline{u^\varepsilon} J^\varepsilon(t) w^\varepsilon)\Big) u^\varepsilon\right\|_{L^{r'}} &\lesssim \||x|^{-\gamma} * (\overline{u^\varepsilon} J^\varepsilon(t) w^\varepsilon)\,\|_{L^{\beta_1}} \|u^\varepsilon(t)\|_{L^s} \\ &\lesssim \|u^\varepsilon(t)\|_{L^s} \|J^\varepsilon(t) w^\varepsilon(t)\|_{L^r} \|u^\varepsilon(t)\|_{L^s},\end{aligned}$$

with r, s as stated in (5.2) and $\frac{1}{r'} = \frac{1}{\beta_1} + \frac{1}{s}$. Here the condition to use the Hardy–Littlewood–Sobolev inequality is $\gamma > \delta(r) + \delta(s)$, which is always satisfied by (5.2). By applying (5.9) now we continue to estimate:

$$\leq \frac{(2C_0)^2}{(|\cos t|)^{2\delta(s)}} \|J^\varepsilon(t) w^\varepsilon(t)\|_{L^r}\,.$$

Then we apply, as before, Proposition 5.3 and estimate the term for h^ε as above, to obtain

$$\|J^\varepsilon w^\varepsilon\|_{L^\infty(0,T;L^2)} \leq C\|\nabla r^\varepsilon\|_{L^2} + C\left(\frac{\varepsilon}{\frac{\pi}{2} - T}\right)^{\gamma-1-\frac{1}{q}}. \tag{5.12}$$

Combining (5.10) and (5.12) along with (2.8), we see that

$$\forall t \in [0,T], \quad \|w^\varepsilon(t)\|_{L^s} \leq \frac{C}{|\cos t|^{\delta(s)}}\left(\|r^\varepsilon\|_{H^1} + \left(\frac{\varepsilon}{\frac{\pi}{2} - t}\right)^{\gamma-1-\frac{1}{q}}\right).$$

Therefore, choosing ε sufficiently small and Λ sufficiently large, we deduce that we can take $T = \frac{\pi}{2} - \Lambda\varepsilon$. With the result of Lemma 4.4 on the limit of $u^\varepsilon_{\rm free}$, this yields Proposition 5.2, away from the focus, for $A^\varepsilon = Id$ and J^ε. The case $A^\varepsilon = H^\varepsilon$ on this time interval is now straightforward.

The remaining parts of the proof for Proposition 5.2 are done as in [6] with the method changed as in the part shown above. It remains to show that the approximations in the two different regimes match at $t_* = \frac{\pi}{2} - \Lambda\varepsilon$ and that the influence of the harmonic potential is small near the focus so that the propagation there is given by

$$v^\varepsilon(t,x) = \frac{1}{\varepsilon^{n/2}}\ \psi\left(\frac{t - \frac{\pi}{2}}{\varepsilon}, \frac{x}{\varepsilon}\right), \tag{5.13}$$

where ψ is the solution of (1.9) subject to the following initial condition at $t = -\infty$:

$$\mathtt{U}(-t)\psi(t)\big|_{t=-\infty} = e^{i\frac{n\pi}{4}}\widehat{f}.$$

This solution exists according to Proposition 5.1.

Then the following asymptotic relation is proven:

$$\limsup_{\varepsilon\to 0} \sup_{\frac{\pi}{2}-\Lambda\varepsilon\le t\le\frac{\pi}{2}+\Lambda\varepsilon} \left\| A^\varepsilon(t)\left(u^\varepsilon(t,x)-v^\varepsilon(t,x)\right)\right\|_{L^2_x} \underset{\Lambda\to+\infty}{\longrightarrow} 0\,,$$

with $A^\varepsilon(t)$ being one of the operators Id, J^ε, or H^ε. Since these parts are quite similar to the treatment in [6], we do not repeat them.

After the crossing of the first focus, the solution is again propagated linearly, and at subsequent focusing points this process is iterated.

6. Formal computations and discussion.

6.1. The case $\alpha = \gamma = 1$ (in 3D: Schrödinger–Poisson). We saw in section 5 that when $\alpha = \gamma > 1$, the nonlinear term in (1.6) has a leading order influence near the foci and only in these regions. On the other hand, if $\alpha = 1$ and $\gamma < 1$, section 4 shows that the Hartree term cannot be neglected away from the foci. These two cases suggest that when $\alpha = \gamma = 1$, the nonlinear influence is relevant everywhere. The aim of this final section is to give convincing arguments that this is the case.

For the influence near the foci, we need the scattering theory for (1.9) at $\gamma = 1$. In this long-range scattering case, modified scattering operators are needed instead of the ones described in Proposition 5.1. Hayashi and Naumkin [16] obtained an asymptotic completeness result for $n \ge 2$ with smoothness assumptions which are applicable to our situation. On the other hand, they could not obtain wave operators. Ginibre and Velo [14, 15] obtained modified wave operators for (1.9) with $\gamma = 1$ using Gevrey spaces by a technically involved method. A drawback of both results is that they include a loss in regularity.

To show how the long-range scattering theory fits into our framework we report (a particular case of) the result of Hayashi and Naumkin [16].

PROPOSITION 6.1 (see [16]). *Assume $n = 3$, $\varphi \in \Sigma$, and $\delta = \|\varphi\|_\Sigma$ is sufficiently small. Let $\psi \in C(\mathbb{R}, \Sigma)$ be the solution of (1.9) with $\psi_{t=0} = \varphi$. Then there exists a unique function $\psi_+ \in H^{\sigma,0} \cap H^{0,\sigma}$, $\frac{1}{2} < \sigma < 1$, such that*

$$\left\|\psi(t) - \exp\left(i\left(|x|^{-1} * |\widehat{\psi_+}|^2\right)\left(\frac{x}{t}\right)\log|t|\right)\mathtt{U}(t)\psi_+\right\|_{L^2} \underset{t\to+\infty}{\longrightarrow} 0\,,$$

where $H^{\alpha,\beta} = \{\phi \in \mathcal{S}' \mid \|(1+|x|^2)^{\beta/2}(1+\Delta)^{\alpha/2}\phi\|_{L^2} < \infty\}$.

To summarize very roughly, the results in [14, 15] consist of showing that given some ψ_+ (or ψ_- for an asymptotic behavior for $t \to -\infty$), one can find ψ solving (1.9) such that the above asymptotic relation holds.

Analogously to the treatment of long-range scattering in [4], one can now define $g^\varepsilon(t,x) := \left(|x|^{-1} * |f|^2\right)(x)\log(\frac{\cos t}{\varepsilon})$ (compare with (4.1)) and add the phase $g^\varepsilon\big|_{t=0}$ to the initial data in (1.6). This yields

$$u^\varepsilon\big|_{t=0} = f(x)e^{-i\left(|x|^{-1}*|f|^2\right)(x)\log\varepsilon}\,.$$

Using the modified scattering operators from the results of [14, 15] we get, at least formally, for $0 \le t < \pi/2$,

$$u^\varepsilon(t,x) \sim \frac{1}{(\cos t)^{3/2}} f\left(\frac{x}{\cos t}\right) e^{-i\frac{x^2}{2\varepsilon}\tan t + ig^\varepsilon\left(t,\frac{x}{\cos t}\right)} \quad \text{as } \varepsilon \to 0\,.$$

This asymptotic relation also stems from the same computations as those performed in section 4.1. Notice that the matching for $|t-\frac{\pi}{2}| = \mathcal{O}(\varepsilon)$ is similar to the one in [6], except that we now have to take the presence of g^ε into account. This is where changing the integration from 0 to t in (4.1) into the above definition of g^ε makes the matching possible. Indeed, for $|t-\frac{\pi}{2}| = \mathcal{O}(\varepsilon)$, we compare u^ε with the function v^ε given by (5.13), where ψ is now the solution given by the long-range wave operators constructed in [14, 15]. To make this statement more precise and the link between (4.1) and the definition of g^ε more explicit, notice that we have, as $t \to \frac{\pi}{2}$,

$$\begin{aligned} g^\varepsilon\left(t, \frac{x}{\cos t}\right) &\sim \left(|x|^{-1} * |f|^2\right)\left(\frac{x}{\frac{\pi}{2}-t}\right) \log\left(\frac{\frac{\pi}{2}-t}{\varepsilon}\right) \quad \text{(phase shift for } v^\varepsilon\text{)} \\ &\sim -\left(|x|^{-1} * |f|^2\right)\left(\frac{x}{\cos t}\right) \int_{\arccos \varepsilon}^{t} \frac{d\tau}{\cos \tau} \quad \text{(compare with (4.1))}\,. \end{aligned}$$

The effects of the nonlinearity show up in g^ε. Using the scaling (5.13) we can then (formally) continue with Proposition 6.1: for $\pi/2 < t < 3\pi/2$,

$$u^\varepsilon(t,x) \sim \frac{e^{-i\frac{3\pi}{2}}}{|\cos t|^{3/2}} \left(\mathcal{F}\circ\widetilde{S}\circ\mathcal{F}^{-1}\right) f\left(\frac{x}{\cos t}\right) e^{-i\frac{x^2}{2\varepsilon}\tan t + i h^\varepsilon\left(t, \frac{x}{\cos t}\right)} \quad \text{as } \varepsilon \to 0\,,$$

where $\widetilde{S}$ is the map $\widetilde{S} : \psi_- \mapsto \psi_+$, where ψ_- is the asymptotic state of the result of [14], which yields some solution ψ to (1.9), and ψ_+ is provided by Proposition 6.1. h^ε is given by

$$h^\varepsilon(t,x) := -\left(|x|^{-1} * |\mathcal{F}\circ\widetilde{S}\circ\mathcal{F}^{-1} f|^2\right)(x) \log\left(\frac{|\cos t|}{\varepsilon}\right).$$

The action of $\mathcal{F}\circ\widetilde{S}\circ\mathcal{F}^{-1}$ on f accounts for nonlinear effects taking place at the focus, and the term h^ε accounts for nonlinear effects after the focus. So the influence of the nonlinearity will be relevant at all times.

The impossibility of defining a scattering operator for this case is one of the reasons this argument is only formal.

Remark 6.2. A rigorous result could be obtained with the same approach as in [3]. It would consist of studying the system of *linear* equations with a *nonlinear* coupling,

$$\begin{cases} i\varepsilon\partial_t \mathbf{u}^\varepsilon + \frac{1}{2}\varepsilon^2 \Delta \mathbf{u}^\varepsilon = \frac{|x|^2}{2}\mathbf{u}^\varepsilon\,, \\ i\varepsilon\partial_t u^\varepsilon + \frac{1}{2}\varepsilon^2 \Delta u^\varepsilon = \frac{|x|^2}{2} u^\varepsilon + \varepsilon\left(|x|^{-1} * |\mathbf{u}^\varepsilon|^2\right) u^\varepsilon\,. \end{cases}$$

The first equation is solved explicitly thanks to Mehler's formula, and the second is a linear Schrödinger equation with a harmonic potential and a time-dependent perturbation. With the oscillatory integral used in section 4, and adapting the results of [8], one could prove similar asymptotic relations to those stated above.

6.2. The case of an additional local strong nonlinearity. We now consider (1.6) with an additional nonlinear term that is a multiplication operator with a power of the density $|\mathfrak{u}^\varepsilon|^2$.

Such equations arise in the modeling of effective one-particle Schrödinger equations where "exchange terms" like in the Hartree–Fock equation are simplified to

functionals of the local densities, i.e., time-dependent density functional theory, with the Schrödinger–Poisson-Xα equation as the simplest of such models (see [25] and [1] for a heuristic derivation and numerical simulations). Note that the additional "local" term has the opposite sign from the Hartree term (corresponding to the physical fact that the "exchange-correlation hole" weakens the direct Coulomb interaction).

We will hence consider the following class of semiclassical Hartree equations:

$$i\varepsilon\partial_t u^\varepsilon + \frac{1}{2}\varepsilon^2\Delta u^\varepsilon = \frac{|x|^2}{2}u^\varepsilon + \varepsilon^\alpha\left(|x|^{-\gamma} * |u^\varepsilon|^2\right)u^\varepsilon - \varepsilon^\beta|u^\varepsilon|^{2\sigma}u^\varepsilon, \tag{6.1}$$

with $\alpha \geq 1$, $\beta \geq 1$, $\gamma > 0$ for $x \in \mathbb{R}^n$ and with a σ that is subcritical with respect to finite time blow-up, i.e., $\frac{2}{n} > \sigma > 0$.

We can now discern the influence of the two nonlinear terms in the classical limit in terms of

- the size of the scaling exponents α, β with respect to the critical value;
- the relation between the scaling and the "strength" of the nonlinearities determined respectively by γ and σ.

If we take $\alpha > 1$ and $\beta > 1$, by [6] and section 5 we find that the classical limit is given by the linear propagation as long as no focusing occurs. At the focus, the relevant discrimination is $\sigma = \beta/n$ or $< \beta/n$ for the power nonlinearity and $\gamma = \alpha$ or $< \alpha$ for the Hartree term. If $\sigma = \beta/n$ and $\gamma < \alpha$, the crossing of the focus will be described by the scattering operator for NLS (when it is defined); if, on the other hand, $\sigma < \beta/n$ and $\gamma = \alpha$ (and the assumptions of Proposition 5.1 are satisfied), focus crossing will be determined by the scattering operator of Proposition 5.1. If both nonlinearities are at the critical strength ($\sigma = \beta/n$ and $\gamma = \alpha$), then both will have an influence in crossing the caustic. If, on the other hand, both $\sigma < \beta/n$ and $\gamma < \alpha$, the nonlinear influence will be negligible everywhere.

If at least one of the scaling exponents α and β is equal to 1 and, at the same time, both $\sigma < \beta/n$ and $\gamma < \alpha$, the corresponding nonlinear term will be relevant in the WKB propagation before the focusing. At the focus, the nonlinear terms will not be relevant and the crossing of the focus will be as in Proposition 4.1. If $\sigma = \beta/n$ and $\gamma = \alpha$, then there will be a nonlinear influence everywhere and long-range scattering for NLS and/or Hartree has to be taken into account.

The influence of the nonlinear action for the single power NLS and the Hartree equation is summed up in two tables; for Hartree the table is given in the introduction, and for the single power nonlinear Schrödinger equation it is stated in [3]. The behavior of (6.1) can be described by independently superposing these two tables. The following table is an extract from that superposition:

	$\alpha > \gamma$ *and* $\beta > \sigma n$	$\alpha = \gamma$ *or* $\beta = \sigma n$
$\alpha > 1$ *and* $\beta > 1$	linear WKB, linear focus	linear WKB, nonlinear focus
$\alpha = 1$ *or* $\beta = 1$	nonlinear WKB, linear focus	nonlinear WKB, nonlinear focus

Here, "nonlinear WKB," respectively, "nonlinear focus," stands for an influence from at least one of the nonlinear terms away from the focus or close to the focus.

6.3. Wigner measures. We already mentioned, in the introduction, the work of Zhang, Zheng, and Mauser [31], where the (semi)classical limit of the Schrödinger–Poisson equation with no smallness assumption (on the initial data or the nonlinearity)

is studied by means of Wigner measures. Wigner measures have proven to be efficient tools for linear semiclassical problems and for homogenization limits; see [26] for an overview on Wigner measure limits of Hartree equations. Wigner measures have the merit that in phase space the caustics of physical space are somewhat unfolded and that generally, results globally in time are possible.

In [5], the Wigner measure of the nonlinear Schrödinger equation with power-like nonlinearity studied in [3] is investigated. It is shown that the Wigner measure leads to an ill-posed problem whenever nonlinear effects at the focal points come into play. In other words, the Wigner measure can be valid only as long as no caustic appears. We briefly discuss the Wigner measures of (1.6) in view of these results.

The Wigner measure of the family $(u^\varepsilon(t))_{0<\varepsilon\leq 1}$, which is bounded in L^2, is the weak limit under $\varepsilon \to 0$ (up to an extraction) of its Wigner transform,

$$W^\varepsilon(u^\varepsilon)(t,x,\xi) = \frac{1}{(2\pi)^n}\int u^\varepsilon\left(t, x-\frac{v\varepsilon}{2}\right)\overline{u^\varepsilon}\left(t,x+\frac{v\varepsilon}{2}\right)e^{i\xi\cdot v}dv.$$

This limit is a positive radon measure μ and is in general not a unique limit.

– *Linear case:* Case $\alpha > \gamma$, $\alpha > 1$.
By the result of Proposition 3.1 and the asymptotic behavior of $u_{\rm free}$ in Lemma 4.4, the Wigner measure μ^- for $t < \pi/2$ of the family $(u^\varepsilon(t))_{0<\varepsilon\leq 1}$ is

$$\mu^-(t,x,\xi) = \frac{1}{|\cos t|^n}\left|f\left(\frac{x}{\cos t}\right)\right|^2 dx\otimes\delta_{\xi = x\tan t}.$$

For $\pi/2 < t < \pi$, the Wigner measure of $(u^\varepsilon(t))_{0<\varepsilon\leq 1}$ (denoted by μ^+) is the same: $\mu^+(t,x,\xi) = \mu^-(t,x,\xi)$. At $t = \pi/2$, the limits from above and below are

$$\lim_{t\to\pi/2^-}\mu^-(t,x,\xi) = \lim_{t\to\pi/2^+}\mu^+(t,x,\xi) = |f(\xi)|^2 d\xi\otimes\delta(x).$$

– *Nonlinear WKB, linear focus:* Case $\gamma < \alpha = 1$.
The asymptotic behavior of u^ε is stated in Proposition 4.1. The additional phase term g is of order 1 and does not change the Wigner measure of $(u^\varepsilon(t))_{0<\varepsilon\leq 1}$, so in this case μ^- and μ^+ are the same as in the previous case; the Wigner measure does not "see" the nonlinear effect g.

– *Linear WKB, nonlinear focus:* Case $\gamma = \alpha > 1$.
The asymptotic relations of Proposition 5.2 involve, for $t \geq \pi/2$, the scattering operator associated with the unscaled equation (1.9). For $t < \pi/2$, the Wigner measure of $(u^\varepsilon(t))_{0<\varepsilon\leq 1}$ is still the same as above, but for $\pi/2 < t < \pi$, we have

$$\mu^+(t,x,\xi) = \frac{1}{|\cos t|^n}\left|\mathcal{F}\circ S\circ\mathcal{F}^{-1}f\left(\frac{x}{\cos t}\right)\right|^2 dx\otimes\delta_{\xi = x\tan t},$$

where S is the scattering operator for (1.9) and $\mathcal{F}$ is the Fourier transform.

– *Nonlinear WKB, nonlinear focus:* Case $\gamma = \alpha = 1$.
The asymptotic behavior for this case of (the formal computation) Proposition 6.1 includes an additional phase term which is of order $\log\varepsilon$ and a modification of the initial data of the same order of magnitude. Both do not alter the Wigner measure, since they are dominated by the scaling of the Wigner transform, and thus the Wigner measure is the same as in the previous case.

For the last two cases, the limits at $t = \pi/2$ are

$$\begin{aligned}\lim_{t\to\pi/2^-}\mu^-(t,x,\xi) &= |f(\xi)|^2 d\xi\otimes\delta(x),\\ \lim_{t\to\pi/2^+}\mu^+(t,x,\xi) &= |\mathcal{F}\circ S\circ\mathcal{F}^{-1}f(\xi)|^2 d\xi\otimes\delta(x).\end{aligned} \tag{6.2}$$

The idea of [5] is to find now two profiles f_1 and f_2 for which $|f_1|^2 \equiv |f_2|^2$, but at the same time $|\mathcal{F} \circ S \circ \mathcal{F}^{-1} f_1|^2 \not\equiv |\mathcal{F} \circ S \circ \mathcal{F}^{-1} f_2|^2$. Then the Wigner measures of the corresponding families $(u_j^\varepsilon(t))_{0<\varepsilon\leq 1}$, $j = 1, 2$, will be equal up to the focus but different after the focus, i.e., $\mu_1^- = \mu_2^-$ but $\mu_1^+ \neq \mu_2^+$. So after the caustic point the Wigner measure will not be unique anymore in the case where the nonlinearity is relevant at the focus. These profiles were constructed using an expansion of S around the origin. Since our problem is very similar to the one studied there, we expect a similar result to hold for (1.6), i.e., we expect the Wigner measure to lead to an ill-posed problem if there is a nonlinear influence at the caustic.

In view of the result of [31], note that the nonuniqueness of the weak solutions for Vlasov–Poisson with measures as initial data and the nonuniqueness of the Wigner measure of a given ε-dependent family of solutions coincide, such that there is no contradiction with the global and unique semiclassical limits of the Hartree-type equations obtained here.

Acknowledgment. This work was completed while the first author was on leave at IRMAR (University of Rennes). He would like to thank this institution for its hospitality.

REFERENCES

[1] W. Bao, N. J. Mauser, and H. P. Stimming, *Effective one particle quantum dynamics of electrons: A numerical study of the Schrödinger-Poisson-Xα model*, Commun. Math. Sci., 1 (2003), pp. 809–828.

[2] P. Bechouche, N. Mauser, and F. Poupaud, *Semiclassical limit for the Schrödinger equation in a crystal with Coulomb interaction*, Comm. Pure Appl. Math., 54 (2001), pp. 851–890.

[3] R. Carles, *Geometric optics with caustic crossing for some nonlinear Schrödinger equations*, Indiana Univ. Math. J., 49 (2000), pp. 475–551.

[4] R. Carles, *Geometric optics and long range scattering for one-dimensional nonlinear Schrödinger equations*, Comm. Math. Phys., 220 (2001), pp. 41–67.

[5] R. Carles, *Remarques sur les mesures de Wigner*, C. R. Acad. Sci. Paris Sér. I Math., 332 (2001), pp. 981–984.

[6] R. Carles, *Semi-classical Schrödinger equations with harmonic potential and nonlinear perturbation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 501–542.

[7] T. Cazenave, *Semilinear Schrödinger Equations*, Courant Lect. Notes Math. 10, New York University Courant Institute of Mathematical Sciences, New York, 2003.

[8] J. Dereziński and C. Gérard, *Scattering Theory of Quantum and Classical N-Particle Systems*, Texts Monogr. Phys., Springer-Verlag, Berlin, Heidelberg, 1997.

[9] J. J. Duistermaat, *Oscillatory integrals, Lagrange immersions and unfolding of singularities*, Comm. Pure Appl. Math., 27 (1974), pp. 207–281.

[10] R. P. Feynman and A. Hibbs, *Quantum Mechanics and Path Integrals*, International Series in Pure and Applied Physics, McGraw-Hill, Maidenhead, Berkshire, UK, 1965.

[11] P. Gérard, P. A. Markowich, N. J. Mauser, and F. Poupaud, *Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math., 50 (1997), pp. 323–379.

[12] J. Ginibre, *An Introduction to Nonlinear Schrödinger Equations*, in Nonlinear Waves (Sapporo, 1995), GAKUTO Internat. Ser. Math. Sci. Appl., R. Agemi, Y. Giga, and T. Ozawa, eds., Gakkōtosho, Tokyo, 1997, pp. 85–133.

[13] J. Ginibre and G. Velo, *On a class of nonlinear Schrödinger equations with nonlocal interaction*, Math. Z., 170 (1980), pp. 109–136.

[14] J. Ginibre and G. Velo, *Long range scattering and modified wave operators for some Hartree type equations.* II, Ann. Henri Poincaré, 1 (2000), pp. 753–800.

[15] J. Ginibre and G. Velo, *Long range scattering and modified wave operators for some Hartree type equations.* III. *Gevrey spaces and low dimensions*, J. Differential Equations, 175 (2001), pp. 415–501.

[16] N. Hayashi and P. Naumkin, *Asymptotics for large time of solutions to the nonlinear Schrödinger and Hartree equations*, Amer. J. Math., 120 (1998), pp. 369–389.

[17] N. Hayashi and Y. Tsutsumi, *Scattering theory for Hartree type equations*, Ann. Inst. H. Poincaré Phys. Théor., 46 (1987), pp. 187–213.

[18] L. Hörmander, *Symplectic classification of quadratic forms, and general Mehler formulas*, Math. Z., 219 (1995), pp. 413–449.
[19] J. Hunter and J. Keller, *Caustics of nonlinear waves*, Wave Motion, 9 (1987), pp. 429–443.
[20] S. Jin, D. Levermore, and D. McLaughlin, *The semiclassical limit of the defocusing NLS hierarchy*, Comm. Pure Appl. Math., 52 (1999), pp. 613–654.
[21] J.-L. Joly, G. Métivier, and J. Rauch, *Caustics for dissipative semilinear oscillations*, Mem. Amer. Math. Soc., 144 (2000), 72 pp.
[22] S. Kamvissis, K. McLaughlin, and P. Miller, *Semiclassical Soliton Ensembles for the Focusing Nonlinear Schrödinger Equation*, Ann. of Math. Stud. 154, Princeton University Press, Princeton, NJ, 2003.
[23] P.-L. Lions and T. Paul, *Sur les mesures de Wigner*, Rev. Mat. Iberoamericana, 9 (1993), pp. 553–618.
[24] P. A. Markowich and N. J. Mauser, *The classical limit of a self-consistent quantum-Vlasov equation in* $3D$, Math. Models Methods Appl. Sci., 3 (1993), pp. 109–124.
[25] N. J. Mauser, *The Schrödinger-Poisson-*$X\alpha$ *equation*, Appl. Math. Lett., 14 (2001), pp. 759–763.
[26] N. J. Mauser, *(Semi)classical limits of Schrödinger-Poisson systems via Wigner transforms*, in Journées "Équations aux Dérivées Partielles" (Forges-les-Eaux, 2002), Univ. Nantes, Nantes, France, 2002, pp. Exp. No. XI, 12.
[27] Y.-G. Oh, *Cauchy problem and Ehrenfest's law of nonlinear Schrödinger equations with potentials*, J. Differential Equations, 81 (1989), pp. 255–274.
[28] J. Rauch and M. Keel, *Lectures on geometric optics*, in Hyperbolic Equations and Frequency Interactions (Park City, UT, 1995), LAS/Park City Math. Ser. 5, AMS, Providence, RI, 1999, pp. 383–466.
[29] C. Sparber, P. A. Markowich, and N. J. Mauser, *Wigner functions versus WKB-methods in multivalued geometrical optics*, Asymptot. Anal., 33 (2003), pp. 153–187.
[30] W. Thirring, *A Course in Mathematical Physics. Vol.* 3. *Quantum Mechanics of Atoms and Molecules*, Lecture Notes in Phys. 141, Springer-Verlag, New York, Vienna, 1981.
[31] P. Zhang, Y. Zheng, and N. J. Mauser, *The limit from the Schrödinger-Poisson to the Vlasov-Poisson equations with general data in one dimension*, Comm. Pure Appl. Math., 55 (2002), pp. 582–632.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 57–81

WEAKLY INTERACTING PULSES IN SYNAPTICALLY COUPLED NEURAL MEDIA*

PAUL C. BRESSLOFF†

Abstract. We use singular perturbation theory to analyze the dynamics of N weakly interacting pulses in a one-dimensional synaptically coupled neuronal network. The network is modeled in terms of a nonlocal integro-differential equation, in which the integral kernel represents the spatial distribution of synaptic weights, and the output activity of a neuron is taken to be a mean firing rate. We derive a set of N coupled ordinary differential equations (ODEs) for the dynamics of individual pulses, establishing a direct relationship between the explicit form of the pulse interactions and the structure of the long-range synaptic coupling. The system of ODEs is used to explore the existence and stability of stationary N-pulses and traveling wave trains.

Key words. neural networks, localized spiral patterns, traveling pulses, integro-differential equations

AMS subject classification. 92C20

DOI. 10.1137/040616371

1. Introduction.

Synaptically coupled neuronal networks provide an important example of spatially extended excitable systems with nonlocal interactions. The network dynamics is usually modeled in terms of an integro-differential equation, in which the integral kernel represents the spatial distribution of synaptic weights and the output activity of a neuron is taken to be a mean firing rate [41, 12]. As in the case of nonlinear PDE models of diffusively coupled excitable systems [23], neuronal networks can exhibit a variety of coherent pulse-like structures including both stationary and traveling solitary pulses. Traveling pulses tend to occur when synaptic connections are predominantly excitatory and there is some form of slow local adaptation or recovery [34, 7], whereas stationary pulses occur in the presence of lateral inhibition [1, 35, 40]. Analogous solutions are found in integrate-and-fire networks, where the output of a neuron is taken to be a sequence of spikes rather than a firing rate [13, 4, 24]. The formation of localized activity states can be used to model a number of neurobiological phenomena. For example, traveling pulses have been observed in disinhibited slice preparations [6, 20, 42] using voltage-sensitive dyes and multiple electrodes. An individual pulse is generated by a brief current stimulus, whereas a train of pulses occurs in the case of repeated stimulation. A second example is given by a delayed response task, in which an animal is required to retain information about a sensory cue across a delay period between the stimulus and behavioral response. Physiological recordings in prefrontal cortex have shown that spatially localized groups of neurons fire during the recall task and then stop firing once the task has finished [18, 39]. Thus persistent localized states of activity are thought to be neural correlates of spatial working memory. An interesting question then concerns the nature of the interactions between multiple regions of localized activity induced by more complex stimuli.

Although there are an increasing number of studies regarding the behavior of solitary pulses in excitable neural media, there is relatively little known about multipulse

*Received by the editors October 5, 2004; accepted for publication (in revised form) April 13, 2005; published electronically October 3, 2005.
http://www.siam.org/journals/siap/66-1/61637.html
†Department of Mathematics, University of Utah, 155 S. 1400 E, Salt Lake City, UT 84112 (bressloff@math.utah.edu).

solutions. One approach to studying stationary N-pulse solutions in rate models is to convert the integro-differential equation into a corresponding fourth-order PDE by an appropriate choice of weight distribution, and then to search for global homoclinic connections [25, 26, 7] or bifurcations from single-pulse solutions [27]. This approach has established, for example, that stable N-pulse solutions can occur when lateral inhibition is modulated by a spatially oscillating component.

In this paper we analyze multipulse solutions of a one-dimensional neuronal network with a Heaviside firing rate function, under the assumption that the individual pulses are well separated so that their mutual interactions are weak. We use singular perturbation theory to derive equations of motion for the pulse positions, in order to investigate the existence and stability of stationary and traveling N-pulse solutions. Our analysis provides a nontrivial extension of previous studies of weakly interacting pulses in nonlinear PDE models of diffusively coupled excitable media and fluids [9, 10, 2, 3, 37, 33]. First, it applies to a nonlocal integro-differential equation that cannot be reduced to a finite-order PDE except for very specific choices of the synaptic weight distribution. Second, using a Heaviside firing rate function, it is possible to obtain exact solutions for single-pulse profiles and to carry out explicit calculations in the derivation of the dynamical equations for weakly interacting pulses. Third, and most significantly from a biological perspective, there is a direct relationship between the nature of the pulse interactions and the form of the long-range synaptic coupling.

We focus in this paper on three distinct but related models that correspond to three distinct experimental paradigms. In section 2, we analyze stationary pulses in a network with symmetric lateral inhibition, which can be interpreted as a simple model of persistent working memory. Assuming that the weights decay exponentially at large distances, the corresponding single-pulse profile also decays exponentially so that widely separated pulses interact weakly. Using singular perturbation theory, we show how the existence and stability of stationary N-pulses reduces to the problem of finding fixed points of a set of N coupled ODEs describing the motion of individual pulses. This is considerably simpler than looking for multipulse solutions of the full nonlocal equation [25, 26, 7]. Consistent with these other studies, we find that in the presence of pure lateral inhibition well separated pulses repel each other, so that a stable N-pulse solution cannot exist. On the other hand, in the case of a spatially decaying oscillatory weight distribution, stable N-pulses can occur. In section 3, we extend our analysis to the case of a network with asymmetric lateral inhibition, which has been proposed as a model of direction selective neurons in visual cortex [38, 30, 32, 43]. Localized activity pulses now tend to propagate unidirectionally rather than remain stationary. Extending the analysis of section 2 by working in the moving frame of a single traveling pulse, we derive the corresponding system of ODEs for N weakly interacting traveling pulses. These equations are then used to explore the existence and stability of traveling wave trains, with the separation between successive pulses characterized by a lattice map [10]. In the case of spatially oscillatory weights, such a map could potentially exhibit both regular and chaotic behavior. In section 4, we consider an excitatory network with an additional adaptation variable, which has been used to model wave propagation in disinhibited cortical slices [34]. In contrast to the asymmetric lateral inhibition network, waves can now travel in both directions. The resulting system of ODEs for N interacting pulses is identical in form to the previous case, but the asymptotic behavior of a single pulse profile is different. In particular, the leading edge of the pulse decays much more rapidly than the trailing edge, which also typically holds for traveling pulses in diffusively coupled excitable systems [23]. This difference in decay rates can be used to reduce the dynamics to a

kinematic form [11, 33].

2. Stationary pulses in symmetric lateral inhibition networks. Let $u(x,t)$ represent the local activity of a population of neurons at position $x \in \mathbf{R}$ in a one-dimensional neuronal network. Suppose that u evolves according to the integro-differential equation

$$\tau_m \frac{\partial u(x,t)}{\partial t} = -u(x,t) + \int_{-\infty}^{\infty} w(x-x')H[u(x',t)-\kappa]dx', \tag{2.1}$$

where τ_m is a membrane or synaptic time constant, κ is a threshold, $H(u)$ denotes the output firing rate, and $w(x-x')$ is the strength of connections from neurons at x' to neurons at x. We assume that w is a continuous function satisfying $w(-x) = w(x)$ and $\int_{-\infty}^{\infty} w(x) < \infty$. The nonlinearity H is taken to be the Heaviside function

$$H[u] = \begin{cases} 0 & \text{if } u \le 0, \\ 1 & \text{if } u > 0. \end{cases} \tag{2.2}$$

In the following we treat length and time in dimensionless units. First, we set $\tau_m = 1$ so that the unit of time is of the order 10msec. Second, the range of the synaptic coupling introduces a fundamental length scale, which we use to set the unit of length to be of the order 200μm.

Equation (2.1) was first analyzed in detail by Amari [1], who showed that there exist stationary solitary pulse solutions when the weight distribution $w(x)$ is given by a Mexican hat function:

(i) $w(x) > 0$ for $x \in [0, x_0)$ with $w(x_0) = 0$,
(ii) $w(x) < 0$ for $x \in (x_0, \infty)$,
(iii) $w(x)$ is decreasing on $[0, x_0]$,
(iv) $w(x)$ has a unique minimum on $\mathbf{R}^+$ at $x = x_1$ with $x_1 > x_0$ and $w(x)$ strictly increasing on (x_1, ∞).

More recently, it has been established that (2.1) can also support stable stationary N-pulse solutions, provided that the weight distribution has additional zeros, which would occur if there were an oscillatory modulation of the long-range connections [25, 26, 27]. We will refer to any network that has long-range inhibition (possibly alternating with long-range excitation) as a lateral inhibition network. In this section we use singular perturbation theory to analyze the dynamics of N-pulse solutions of (2.1), under the assumption that the interactions between pulses are weak. We show that in the case of a Mexican hat weight distribution, the pulses mutually repel each other so that stable N-pulse solutions cannot occur. On the other hand, when the weight distribution consists of decaying spatial oscillations, there exist configurations of the pulse locations (up to a global translation) corresponding to stable N-pulse bound states. This result is consistent with the findings of Laing et al. [25] and Laing and Tray [26, 27].

2.1. Stationary solitary pulses. Suppose that $U(x)$ is a stationary solution of (2.1): $u(x,t) = U(x)$ with

$$U(x) = \int_{-\infty}^{\infty} w(x-x')H[U(x')-\kappa]dx'. \tag{2.3}$$

Let $\mathcal{M}[U] = \{x | U(x) > \kappa\}$ be the region over which the activity U is excited (superthreshold). Equation (2.3) can then be rewritten as

$$U(x) = \int_{\mathcal{M}[U]} w(x-x')dx'. \tag{2.4}$$

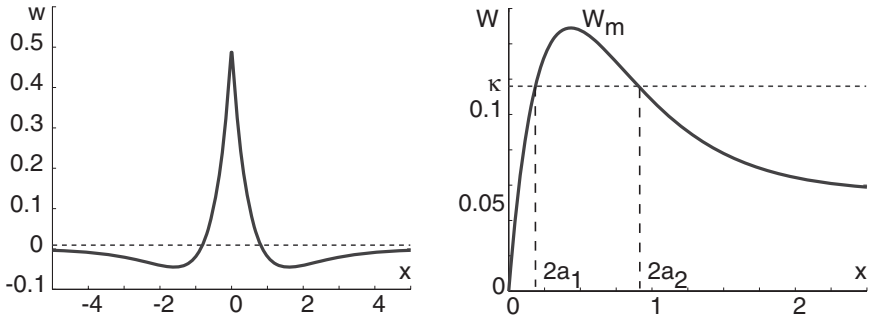


FIG. 2.1. *Construction of stationary pulses for a Mexican hat weight distribution.* (a) *Plot of* $w(x)$ *given by the difference-of-exponentials* (2.8) *with* $\sigma_E = 1.8, \sigma_I = 1.0$, *and* $\Gamma = 0.5$. (b) *Plot of corresponding function* $W(x)$. *Horizontal line shows the threshold* κ *whose intersections with* $W(2a)$ *determine the allowed pulse widths.*

We define a single pulse solution of width $2a$ to be one that is excited over the interval $(-a, a)$; any pulse solution can be arbitrarily translated so that it is centered at the origin. If we set

$$W(x) = \int_0^x w(y)dy, \tag{2.5}$$

then (2.4) reduces to the form

$$U(x) = W(a + x) - W(x - a). \tag{2.6}$$

Note that $W(0) = 0$ and $W(-x) = -W(x)$. Since $U(\pm a) = \kappa$, we obtain the following necessary condition for the existence of a stationary pulse of width $2a$:

$$W(2a) = \kappa. \tag{2.7}$$

Amari [1] showed that in the case of a Mexican hat weight distribution this condition is also sufficient. He also established that a stationary pulse is stable, provided that $W'(a) \equiv w(a) < 0$; otherwise it is unstable.

The stable and unstable pulses can be determined graphically, as illustrated in Figure 2.1 for a Mexican hat function w given by the difference-of-exponentials

$$w(x) = \mathrm{e}^{-\sigma_E|x|} - \Gamma \mathrm{e}^{-\sigma_I|x|}, \tag{2.8}$$

with $\sigma_E > \sigma_I > 0$ and $0 < \Gamma < 1$. Here σ_E^{-1} and σ_I^{-1} determine the range of excitatory and inhibitory synaptic coupling, respectively. If one neglects long-range horizontal connections (see below), then such coupling tends to extend up to around 0.8mm in cortex [29]. Integrating (2.8), we have

$$W(x) = \frac{1}{\sigma_E}\left[1 - \mathrm{e}^{-\sigma_E x}\right] - \frac{\Gamma}{\sigma_I}\left[1 - \mathrm{e}^{-\sigma_I x}\right] \tag{2.9}$$

for $x \geq 0$. The existence of single-pulse solutions depends on the relative sizes of W_m, W_∞, and κ, where

$$W_m = \max_{x>0} W(x), \quad W_\infty = \lim_{x\to\infty} W(x) = \frac{1}{\sigma_E} - \frac{\Gamma}{\sigma_I}. \tag{2.10}$$

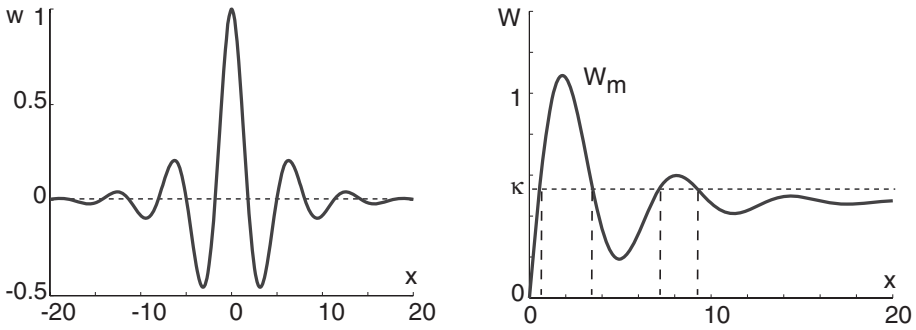


FIG. 2.2. *Construction of a stationary pulse for a spatially decaying oscillatory weight distribution.* (a) *Plot of* $w(x)$ *given by* (2.11) *with* $\sigma = 0.25$. (b) *Plot of corresponding function* $W(x)$. *The horizontal line shows the threshold* κ *whose intersections with* $W(2a)$ *determine the allowed pulse widths.*

If $0 < W_\infty < \kappa < W_m$, then there exists an unstable pulse of width a_1 and a stable pulse of width a_2 with $a_2 > a_1$, whereas there is only an unstable pulse when $0 < \kappa < W_\infty$. In the latter case the network is in a bistable regime, where the unstable pulse acts as a separatrix between a stable uniform resting state ($U \equiv 0$) and a traveling front. If $W_\infty < \kappa < 0 < W_m$, then there is a stable pulse but no unstable pulse. Outside these parameter regimes there are no pulses.

In Figure 2.2, we illustrate the corresponding graphical construction for a spatially decaying oscillatory weight distribution of the form [25]

$$w(x) = \mathrm{e}^{-\sigma|x|}\left[\cos(x) + \sigma \sin|x|\right]. \tag{2.11}$$

Integrating (2.11), we have

$$W(x) = \frac{2\sigma}{1+\sigma^2}\left[1 - \mathrm{e}^{-\sigma x}\cos(x)\right] + \frac{1-\sigma^2}{1+\sigma^2}\mathrm{e}^{-\sigma x}\sin(x), \quad x \geq 0, \tag{2.12}$$

with $W(-x) = -W(x)$. It can be seen from Figure 2.2 that, as κ is reduced below W_m, an increasing number of stable/unstable pairs of pulses are generated, assuming that condition (2.7) is sufficient to ensure $U(x) > \kappa$ for $|x| < a$ and $U(x) < \kappa$ for $|x| > a$. This has not been proven analytically for the weight distribution (2.11), although the existence of stable pulses has been confirmed numerically by Laing et al. [25]. These authors have also suggested that an anatomical substrate for the oscillatory weight distribution (2.11) might be the long-range horizontal connections found in superficial layers of cortex. Such connections extend several millimeters across cortex and are broken into discrete patches with a very regular size and spacing [36, 19, 28]. Although the horizontal connections arise almost exclusively from excitatory neurons, 20% of them terminate on interneurons that can generate significant inhibition [31]. Whether the horizontal connections have a net inhibitory or excitatory effect does not appear to be a simple function of cortical separation, however, since it also depends on the local level of activity of neurons innervated by the long-range connections [29]. Therefore, certain care has to be taken in the biological interpretation of the weight distribution (2.11).

2.2. Singular perturbation theory. Suppose that (2.1) has a stable stationary pulse solution $U(x)$ of width $2a$ centered at the origin such that in the large-$|x|$ limit,

the activity of the pulse decays exponentially, $|U(x)| \sim \mathrm{e}^{-\rho|x|}$. For example, $\rho = \sigma_I$ in the case of the Mexican hat function (2.8), and $\rho = \sigma$ in the case of the spatially decaying oscillatory function (2.11). This suggests that if two or more such pulses are placed on the real line such that the characteristic separation d between the centers of any two pulses satisfies $\mathrm{e}^{-\rho d} = \varepsilon \ll 1$, then the interactions between the pulses will be weak. In the weakly interacting regime, we can carry out a perturbation analysis of the dynamics along lines analogous to those used by Elphick, Meron, and Spiegel [10] by treating ε as a small parameter.

Following [10], we look for an N-pulse solution with individual pulses having centers at $x_n = nd + \phi_n(\tau)$, where $\phi_n(\tau)$ is a slowly varying phase and $\tau = \varepsilon t$. That is, we consider a train of pulses

$$u(x,\tau) = \sum_{n=1}^{N} U(x - nd - \phi_n(\tau)) + \varepsilon R(x,\tau). \tag{2.13}$$

The remainder term εR takes into account the fact that a superposition of widely separated pulses cannot be an exact solution, even when we allow for slowly drifting phases ϕ_n. Substituting (2.13) into (2.1) with $\partial_t \to \partial_t + \varepsilon\partial_\tau$, and using (2.3), gives

$$\varepsilon^2\partial_\tau R - \varepsilon\sum_{n=1}^{N}\dot{\phi}_n U_n' = -\varepsilon R + w * H\left(\sum_{n=1}^{N} U_n + \varepsilon R - \kappa\right) - w * \sum_{n=1}^{N} H(U_n - \kappa), \tag{2.14}$$

where $U_n(x) = U(x - nd - \phi_n(\tau))$ and $w * g$ denotes the convolution

$$w * g = \int_{-\infty}^{\infty} w(x - x')g(x')dx'. \tag{2.15}$$

We now carry out a perturbation expansion in ε by formally Taylor expanding with respect to εR inside the convolution integral:

$$w * H\left(\sum_{n=1}^{N} U_n - \kappa + \varepsilon R\right) = w * \left[H\left(\sum_{n=1}^{N} U_n - \kappa\right) + \varepsilon\delta\left(\sum_{n=1}^{N} U_n - \kappa\right) R + \mathcal{O}\left(\varepsilon^2\right)\right], \tag{2.16}$$

where δ is the Dirac delta function. This formal series expansion can be interpreted along the following lines. First, we assume that in the case of widely separated pulses, the multibump function $V \equiv \sum_n U_n(x)$ has N pairs of threshold crossing points $x_m^{\pm} \approx x_m \pm a$ such that $V(x) > \kappa$ for $x_m^- < x < x_m^+$, $V(x_m^{\pm}) = \kappa$ and $V(x) < \kappa$ otherwise. It follows that

$$\delta(V(x) - \kappa) = \sum_{m=1}^{N}\left[\frac{\delta(x - x_m^+)}{|V'(x_m^+)|} + \frac{\delta(x - x_m^-)}{|V'(x_m^-)|}\right]. \tag{2.17}$$

Similarly, the function $V + \varepsilon R$ is assumed to have threshold crossing points at $x_m^{\pm} + \varepsilon\Delta_m^{\pm}$. The convolution integral then has the explicit form

$$\begin{aligned} w * H\left(\sum_{n=1}^{N} U_n - \kappa + \varepsilon R\right)(x) &= \sum_{n=1}^{N}\int_{x_n^- + \varepsilon\Delta_n^-}^{x_n^+ + \varepsilon\Delta_n^+} w(x - x')dx' \\ &= \sum_{n=1}^{N}\int_{x_n^-}^{x_n^+} w(x - x')dx' \\ &\quad + \varepsilon\left[w(x - x_n^+)\Delta_n^+ - w(x - x_n^-)\Delta_n^-\right] + \mathcal{O}(\varepsilon^2), \end{aligned} \tag{2.18}$$

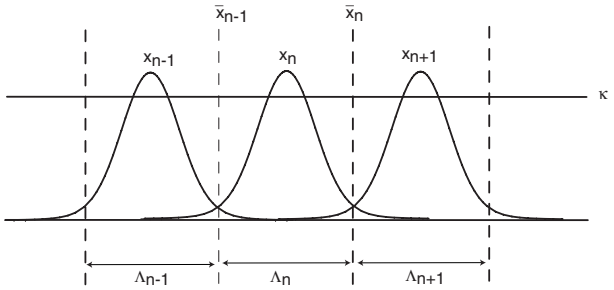


FIG. 2.3. *Illustrative sketch of a multibump solution in which the nth activity bump (region above threshold κ) is localized within the domain Λ_n. In a neighborhood of the mth activity bump we assume that $U_n(x) \sim \varepsilon^{|m-n|}$ for $m \neq n$, where $U_n(x) = U(x - x_n)$ and x_n is the center of the nth bump.*

where we have Taylor expanded with respect to the perturbations $\varepsilon\Delta_m^{\pm}$ in the locations of the activity bump boundaries. Substituting (2.17) into (2.16) shows that the latter is equivalent to (2.18) with $\Delta_m^+ = R(x_m^+)/|V'(x_m^+)|$ and $\Delta_m^- = -R(x_m^-)/|V'(x_m^-)|$.

Substituting (2.16) into (2.14) and collecting first-order terms in ε leads to the following inhomogeneous equation for R:

$$\widehat{L}R = \sum_{n=1}^{N} \dot{\phi}_n U_n' + \frac{1}{\varepsilon} w * \left[H\left(\sum_{n=1}^{N} U_n - \kappa \right) - \sum_{n=1}^{N} H(U_n - \kappa) \right], \tag{2.19}$$

where $\widehat{L}$ is the linear operator

$$\widehat{L}\psi = \psi - w * \left(\delta\left(\sum_{n=1}^{N} U_n - \kappa \right) \psi \right) \tag{2.20}$$

for any function $\psi \in L^2(\mathbf{R})$. We now show that the term in square brackets on the right-hand side of (2.19) is $\mathcal{O}(\varepsilon)$. First, partition the real line into nonoverlapping domains such that the mth activity bump lies entirely in the domain Λ_m, as illustrated in Figure 2.3. More specifically, take $\mathbf{R} = \cup_{m=1}^{N} \Lambda_m$ with $\Lambda_m = [\bar{x}_{m-1}, \bar{x}_m)$ for $m = 2, \ldots, N-1$, $\Lambda_1 = (-\infty, \bar{x}_1)$, and $\Lambda_N = (\bar{x}_{N-1}, \infty)$. Here $\bar{x}_m = (x_{m+1} - x_m)/2$ is the midpoint between neighboring bumps. The characteristic pulse separation d is assumed to be sufficiently large such that $U_n(x) \sim \varepsilon^{|m-n|}$ in a neighborhood of the mth activity bump for $m \neq n$. The given partition then allows us to carry out a formal perturbation expansion along lines similar to (2.16):

$$\begin{aligned}
&\int_{-\infty}^{\infty} w(x - x') H\left(\sum_n U_n(x') - \kappa \right) dx' = \sum_{m=-\infty}^{\infty} \int_{\Lambda_m} w(x - x') H\left(\sum_n U_n(x') - \kappa \right) dx' \\
&= \sum_{m=-\infty}^{\infty} \int_{\Lambda_m} w(x - x') H\left(U_m(x') + \sum_{n \neq m} U_n(x') - \kappa \right) dx' \\
&= \sum_{m=-\infty}^{\infty} \int_{\Lambda_m} w(x - x') \left(H(U_m(x') - \kappa) + \delta(U_m(x') - \kappa) \sum_{n=m\pm1} U_n(x') + \mathcal{O}(\varepsilon^2) \right) dx'.
\end{aligned} \tag{2.21}$$

Only nearest neighbor bumps contribute to the $\mathcal{O}(\varepsilon)$ term. The delta function $\delta(U_m(x)-\kappa)$ can be simplified using the threshold condition $U_m(\pm a+x_m)=\kappa$:

$$\delta(U_m(x)-\kappa)=\frac{\delta(x-a-x_m)}{|U'(a)|}+\frac{\delta(x+a-x_m)}{|U'(-a)|}. \tag{2.22}$$

Since $U_m(x)<\kappa$ for all $x\notin\Lambda_m$, we can replace the integral domain Λ_m in each of the integrals in (2.21) by the whole real line $(-\infty,\infty)$. Thus we find

$$w*H\left(\sum_{n=1}^{N}U_n-\kappa\right)=w*\left[\sum_{n=1}^{N}H(U_n-\kappa)+\sum_{n=1}^{N}\delta(U_n-\kappa)[U_{n+1}+U_{n-1}]+\mathcal{O}(\varepsilon^2)\right] \tag{2.23}$$

with $U_{N+1},U_0\equiv 0$.

In order for the above perturbation expansion to be valid, we require that the correction term R be finite everywhere. Following Elphick, Meran, and Spiegel [10], we will show how this implies a set of solvability conditions on (2.19), which in turn determine the leading-order dynamics of the phases ϕ_n or, equivalently, the pulse positions x_n. In order to gain insights into the analytical properties of the linear operator $\widehat{L}$ it is useful to consider the simpler operator $\widehat{L}_n$, where

$$\widehat{L}_n\psi=\psi-w*[\delta(U_n-\kappa)\psi]. \tag{2.24}$$

The latter has zero as an eigenvalue with corresponding eigenfunction $Q=U_n'$, which can be seen by differentiating (2.3). Therefore, the eigenvalue equation $\widehat{L}Q=\lambda Q$ has approximate solutions of the form $Q=\sum_{i=1}^{N}c_nU_n'$, where c_n are constants, with associated eigenvalues $\lambda=\mathcal{O}(\varepsilon)$. Assuming the standard inner product of functions P,Q on $\mathbf{R}$,

$$\langle P|Q\rangle=\int_{-\infty}^{\infty}P(x)Q(x)dx, \tag{2.25}$$

we define the adjoint operator $\widehat{L}^\dagger$ according to

$$\langle P|\widehat{L}Q\rangle=\langle\widehat{L}^\dagger P|Q\rangle, \tag{2.26}$$

so that

$$\widehat{L}^\dagger\psi=\psi-\delta\left(\sum_{n=1}^{N}U_n-\kappa\right)w*\psi. \tag{2.27}$$

The existence of N null vectors of $\widehat{L}_n$ suggests that its adjoint should also have N null vectors, which can then be used to construct eigenfunctions of $\widehat{L}^\dagger$ with $\mathcal{O}(\varepsilon)$ eigenvalues. However, since the operator $\widehat{L}_n$ is not self-adjoint, one cannot assume a priori that it has index zero. The situation is further complicated by the fact that $\widehat{L}^\dagger$ involves distributions. Therefore, we will proceed by searching for weak solutions P of the equation

$$\langle\widehat{L}^\dagger P|Q\rangle=\mathcal{O}(\varepsilon) \tag{2.28}$$

for any bounded function Q.

Comparison of (2.17) and (2.22) shows that in the weakly interacting regime,

$$\int \psi(x)\delta\left(\sum_n U_n(x)-\kappa\right)dx = \int \psi(x)\sum_{n=1}^{N}\delta\left(U_n(x)-\kappa\right)dx+\mathcal{O}(\varepsilon) \tag{2.29}$$

for arbitrary ψ. This leads to the formal decomposition of the adjoint operator given by

$$\widehat{L}^{\dagger}\psi = \widehat{L}_n^{\dagger}\psi - \sum_{j\neq n}\delta(U_j-\kappa)w*\psi+\mathcal{O}(\varepsilon) \tag{2.30}$$

for any $n=1,\ldots,N$, with $\delta(U_n-\kappa)$ satisfying (2.22). Without loss of generality, set $x_0=0$ and look for solutions of $\widehat{L}_0^{\dagger}\mathcal{P}=0$, which can be written as

$$\mathcal{P}(x) = \left[\frac{\delta(x+a)}{U'(-a)}+\frac{\delta(x-a)}{|U'(a)|}\right]\int_{-\infty}^{\infty} w(x-x')\mathcal{P}(x')dx'. \tag{2.31}$$

The formal solution is $\mathcal{P}(x)=p_1\delta(x-a)-p_2\delta(x+a)$, with coefficients p_1,p_2 satisfying the pair of algebraic equations

$$p_1 = \frac{1}{|U'(a)|}\left[p_1w(0)-p_2w(2a)\right],$$

$$p_2 = -\frac{1}{U'(-a)}\left[p_1w(2a)-p_2w(0)\right].$$

Differentiating (2.6) shows that

$$U'(-a) = -U'(a) = w(0)-w(2a), \tag{2.32}$$

and hence $p_1=p_2$. This establishes that $\widehat{L}_0^{\dagger}$ has the null vector

$$\mathcal{P}(x) = \delta(x-a)-\delta(x+a). \tag{2.33}$$

From translation symmetry, it follows that $\widehat{L}_n^{\dagger}$ has the null vector $P_n(x)$, where

$$P_n(x) \equiv \mathcal{P}(x-x_n). \tag{2.34}$$

Hence, using the decomposition (2.30) and the result $\widehat{L}_n^{\dagger}P_n=0$, we have

$$\langle \widehat{L}^{\dagger}P_n|Q\rangle = -\sum_{j\neq n}\langle\delta(U_j-\kappa)w*P_n|Q\rangle+\mathcal{O}(\varepsilon) \tag{2.35}$$

for any bounded function Q. Since $w(x)$ decays as $\mathrm{e}^{-\rho|x|}$ for large $|x|$, we see that $w(x_n-x_j)\sim\varepsilon^{|n-j|}$ for $n\neq j$, so that the inner product on the right-hand side of (2.35) is also $\mathcal{O}(\varepsilon)$. We conclude that (2.28) has the set of solutions P_n, $n=1,\ldots,N$. Interestingly, these solutions are independent of the choice of weight function w.

We now take the inner product of (2.19) with respect to P_m and use (2.23). Keeping only leading-order terms in ε then yields the following solvability condition:

$$\dot{\phi}_m\left\langle P_m\,|U_m'\right\rangle + \frac{1}{\varepsilon}\left\langle P_m\,|w*\delta(U_m-\kappa)[U_{m+1}+U_{m-1}]\right\rangle = \mathcal{O}(\varepsilon). \tag{2.36}$$

Substituting (2.33) into (2.36) and using (2.22), (2.32) shows that

$$\langle P_m\,|w*\delta(U_m-\kappa)[U_{m+1}+U_{m-1}]\rangle = U(a+x_m-x_{m+1})+U(a+x_m-x_{m-1}) - U(-a+x_m-x_{m+1})-U(-a+x_m-x_{m-1})$$

(2.37)

and

$$\langle P_m\,|U_m'\rangle = [U_m'(a+x_m)-U_m'(-a+x_m)] = -2|U'(a)|. \tag{2.38}$$

We thus obtain the following system of differential equations for the pulse positions $x_m(t) = md + \phi_m(\varepsilon t)$:

$$\begin{aligned} &\dot{x}_m = f(x_{m+1}-x_m) - f(x_m - x_{m-1}), \quad 1 < m < N, \\ &\dot{x}_1 = f(x_2-x_1), \quad \dot{x}_N = -f(x_N - x_{N-1}) \end{aligned} \tag{2.39}$$

with

$$f(x) = \frac{1}{2|U'(a)|}\left[U(x-a)-U(x+a)\right]. \tag{2.40}$$

2.3. Stationary *N*-pulses. Equations (2.6) and (2.40) show that the explicit form of the interaction function $f(x)$ for large x is determined by the asymptotic behavior of the weight distribution w. In the case of the Mexican hat function (2.8), equations (2.6), (2.9), and (2.40) imply that for large x

$$f(x) = -\frac{\Gamma\cosh(\sigma_I a)}{\sigma_I|U'(a)|}\left[\mathrm{e}^{-\sigma_I(x-a)} - \mathrm{e}^{-\sigma_I(x+a)}\right] = -Ae^{-\sigma_I x} \tag{2.41}$$

with

$$A = \frac{\Gamma\sinh(2\sigma_I a)}{\sigma_I|U'(a)|} > 0. \tag{2.42}$$

Similarly, in the case of a spatially decaying oscillatory function (2.11), we find that for large x,

$$f(x) = \frac{2\mathrm{e}^{-\sigma x}}{(1+\sigma^2)|U'(a)|}\left[A_1(a,\sigma)\cos(x) + A_2(a,\sigma)\sin(x)\right] \tag{2.43}$$

with

$$A_1(a,\sigma) = (1-\sigma^2)\sinh(\sigma a)\sin(a) + 2\sigma\cosh(\sigma a)\cos(a) - 2\sigma, \tag{2.44}$$

$$A_2(a,\sigma) = 2\sigma\sinh(\sigma a)\sin(a) - (1-\sigma^2)\cosh(\sigma a)\cos(a) + 1 - \sigma^2. \tag{2.45}$$

The function f can be written in the more compact form

$$f(x) = B\mathrm{e}^{-\sigma x}\cos(x-\Phi) \tag{2.46}$$

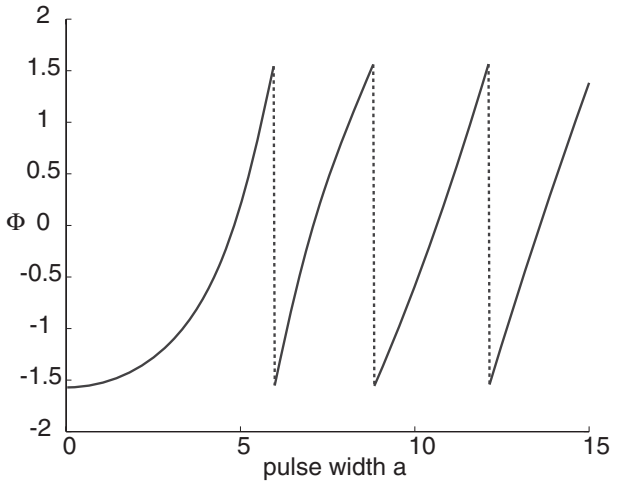


FIG. 2.4. *Plot of phase separation* Φ *for a pair of pulses as a function of pulse width* a *in the case of the spatially decaying oscillatory weight distribution* (2.11) *with* $\sigma = 0.25$.

with

$$\Phi = \tan^{-1}\frac{A_2}{A_1}, \quad B = \frac{2\sqrt{A_1^2 + A_2^2}}{(1+\sigma^2)|U'(a)|}. \tag{2.47}$$

The dependence of the phase Φ on pulse width for fixed decay rate σ is shown in Figure 2.4. Note that for $\sigma a \gg 1$ we have $\Phi \approx a - \phi$ with $\tan\phi = (1-\sigma^2)/2\sigma$.

The existence and stability of stationary N-pulse solutions can now be investigated in terms of the fixed point solutions of (2.39) for a given weight distribution. Let us first consider a pair of pulses, whose positions satisfy the pair of equations

$$\dot{x}_1 = f(x_2 - x_1), \quad \dot{x}_2 = -f(x_2 - x_1). \tag{2.48}$$

Defining the separation variable $\Delta = x_2 - x_1$, we have

$$\dot{\Delta} = -2f(\Delta). \tag{2.49}$$

It immediately follows from (2.41) that mutual interactions between pulses are repulsive for a Mexican hat weight function, since $f(\Delta) < 0$ for all Δ. Hence, the pulses repel each other and cannot form a bound 2-pulse state. A similar result holds for more than two pulses. Note that Amari [1] originally suggested that pulses were attractive at short distances, repulsive at intermediate distances, and neutral at sufficiently large distances. Our analysis suggests that repulsion actually persists to arbitrarily large distances, but that the rate of separation is slow since $f(\Delta) \sim \mathrm{e}^{-\sigma_I \Delta} \ll 1$. Following the results of Laing and colleagues [25, 27] and analogous results for PDEs [2], one expects a stable N-pulse solution to exist when the weights have an oscillatory tail. This is easily seen in the case for a pair of pulses by substituting (2.46) into (2.49):

$$\dot{\Delta} = -2B\mathrm{e}^{-\sigma\Delta}\cos(\Delta - \Phi). \tag{2.50}$$

Such an equation has a countable set of stable/unstable pairs of fixed point solutions: $\Delta = \Delta_\pm(p) = \Phi \pm \pi/2 + 2\pi p$ for integers $p \gg 1$ (so that pulses are well separated) with $\Delta_-(p)$ stable and $\Delta_+(p)$ unstable.

Higher-order N-pulse solutions can be constructed by taking the separations $\Delta_m = x_m - x_{m-1}$ to be zeros of f for all $m = 2, \ldots, N$. Stability is determined by the eigenvalues of the tridiagonal matrix

$$\mathbf{A}_N = \begin{pmatrix} -2f'(\Delta_2) & f'(\Delta_3) & 0 & \ldots & 0 \\ f'(\Delta_2) & -2f'(\Delta_3) & f'(\Delta_4) & 0 & \ldots \\ 0 & f'(\Delta_3) & -2f'(\Delta_4) & f'(\Delta_5) & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & f'(\Delta_{N-1}) & -2f'(\Delta_N) \end{pmatrix}. \tag{2.51}$$

Note that the matrix coefficients satisfy $a_{i,i+1} = a_{i+1,i} > 0$ for all i so that the eigenvalues are all real and simple. Moreover, by the Gerschgorin disk theorem (see [22]), the eigenvalues of $\mathbf{A}_N$ are contained in the union of disks defined according to $\cup_i \{|\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|\}$. Consider the simplest case in which all pulse spacings are equal, $\Delta_m = \Delta$ for all $m = 2, \ldots, N$. We then obtain the pair of conditions

$$|\lambda + 2f'(\Delta)| \leq f'(\Delta), \quad |\lambda + 2f'(\Delta)| \leq 2f'(\Delta). \tag{2.52}$$

These are circles contained within the left-half complex plane, provided that $f'(\Delta) > 0$. One can also show that there are no zero eigenvalues by noting that in the uniform case the determinant $D_N = \det[\mathbf{A}_N - \lambda \mathbf{I}]$ satisfies the iterative equation

$$D_m(\lambda) = (-2f'(\Delta) - \lambda) D_{m-1} - f'(\Delta)^2 D_{m-2}, \quad 2 \leq m \leq N, \tag{2.53}$$

with $D_1 = 1$ and $D_0 = 0$. This has the solution

$$D_N(\lambda) = \frac{\Lambda_+^N - \Lambda_-^N}{\Lambda_+ - \Lambda_-}, \tag{2.54}$$

where

$$\Lambda_\pm = \frac{1}{2}\left(-2f' - \lambda \pm \sqrt{\lambda^2 + 4f'\lambda}\right). \tag{2.55}$$

Since $D_N(0) \neq 0$, it follows that zero is not an eigenvalue. Therefore, there exists a stable uniformly spaced stationary N-pulse solution if $f(\Delta) = 0$ and $f'(\Delta) > 0$.

One can also analyze the stationary states of N pulses arranged on a ring of length L. Now the dynamics is described by the cyclic system of ODEs

$$\dot{x}_m = f(x_{m+1} - x_m) - f(x_m - x_{m-1}), \quad m = 1, \ldots, N, \tag{2.56}$$

with $x_0 = x_N$ and $x_{N+1} = x_0$. The evenly spaced solution $\Delta = L/N$ is automatically a fixed point of the dynamics, and its stability can be determined by linearizing (2.56) with $x_m = m\Delta + \theta_m$:

$$\dot{\theta}_m = -f'(\Delta)\left[2\theta_m - \theta_{m-1} - \theta_{m+1}\right]. \tag{2.57}$$

This has eigensolutions of the form $\theta_m(t) = \mathrm{e}^{\lambda(k)t}\mathrm{e}^{imk}$ with wavenumber $k = 2\pi p/N$ for $p = 1, \ldots, N$ and

$$\lambda(k) = -2f'(\Delta)[1 - \cos(k)]. \tag{2.58}$$

The zero eigenvalue at $k = 0$ reflects the translation invariance of the system. Hence, the N-pulse solution on the ring is (marginally) stable if $f'(\Delta) > 0$; otherwise it is

unstable. It then follows from (2.41) and (2.46) that a ring network with a Mexican hat weight distribution supports stable N-pulse solutions independently of the length L of the ring, whereas a spatially decaying oscillatory distribution supports such a solution only for certain ranges of L. This example illustrates how the existence and stability of multipulse solutions depends on the topology of the network as well as its weight distribution.

3. Traveling pulses in asymmetric lateral inhibition networks. In section 2 we considered a lateral inhibition network with a weight distribution that is symmetric, $w(-x) = w(x)$. Although this is usually a reasonable modeling assumption regarding the large-scale anatomy of cortical circuits, there are some examples of more specialized circuits where lateral inhibition may be asymmetric. In particular, asymmetric coupling has been suggested as providing a possible mechanism for direction selective neurons in the visual cortex [38, 30, 32, 43]. Networks with asymmetric lateral inhibition support unidirectional wave propagation rather than stationary activity pulses. If a moving external stimulus is presented to a one-dimensional network, then a superthreshold response is elicited only if the velocity of the stimulus approximately matches the direction and speed of the intrinsic waves. Here we extend the singular perturbation theory of stationary pulses in order to investigate traveling N-pulse solutions of (2.1) in the case of asymmetric lateral inhibition.

3.1. Traveling solitary pulses. Suppose that (2.1) with asymmetric w has a right-moving traveling pulse solution of the form $u(x,t) = U(x-ct)$, $c > 0$, where $U(\pm\infty) = 0$ and $U(-a) = U(0) = \kappa$. Substituting into (2.1) with $\xi = x - ct$ gives

$$\begin{aligned} -c\partial_\xi U(\xi) + U(\xi) &= \int_{-\infty}^{\infty} w(\xi - \xi')H(U(\xi') - \kappa)d\xi' \\ &= W(\xi + a) - W(\xi), \end{aligned} \tag{3.1}$$

with W defined by (2.5). Multiplying both sides of (3.1) by the integrating factor $-c^{-1}\mathrm{e}^{-\xi/c}$ and integrating from $-a$ to ξ using the threshold condition $U(-a) = \kappa$ leads to the result

$$U(\xi) = \mathrm{e}^{\xi/c}\left[\kappa\mathrm{e}^{a/c} - \frac{1}{c}\int_{-a}^{\xi}[W(\xi' + a) - W(\xi')]\mathrm{e}^{-\xi'/c}d\xi'\right]. \tag{3.2}$$

Finiteness of the solution in the limit $\xi \to \infty$ requires that the term in square brackets vanish. Hence, we can rewrite the solution for $U(\xi)$ as

$$U(\xi) = \frac{1}{c}\int_0^{\infty}(W(\xi' + \xi + a) - W(\xi' + \xi))\mathrm{e}^{-\xi'/c}d\xi'. \tag{3.3}$$

Enforcing the threshold conditions $U(0) = \kappa$ and $U(-a) = \kappa$ then generates a pair of equations that determine existence curves relating the speed c to the pulse width a for a given threshold κ.

A typical way to model asymmetric lateral inhibition is to take $w(x) = w_0(x - x_0)$ with w_0 a symmetric weight distribution such as a Mexican hat function. If $x_0 > 0$, then short-range coupling is predominantly excitatory to the right and inhibitory to the left, which leads to right-propagating waves, as illustrated in Figure 3.1. The function $W(x)$ of (2.9) may be expressed in terms of w_0 as

$$W(x) = W_0(x - x_0) + W_0(x_0), \quad W_0(x) = \int_0^x w_0(y)dy, \tag{3.4}$$

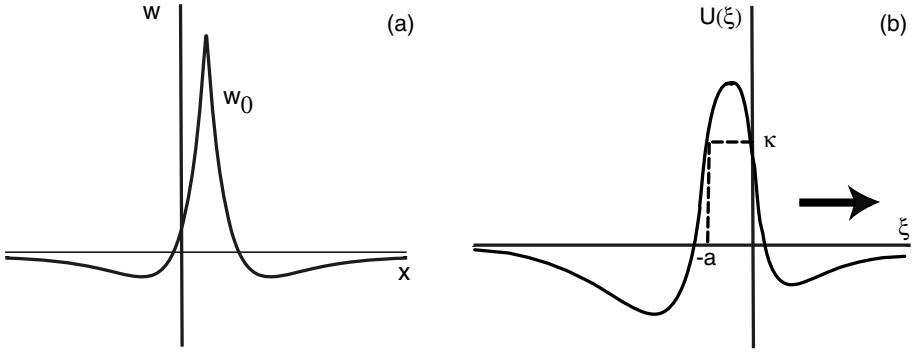


FIG. 3.1. (a) *Shifted weight distribution* $w(x) = w(x - x_0)$ *in the case of asymmetric lateral inhibition.* (b) *Illustration of a right-moving traveling pulse of width* a.

so that the wave profile becomes

$$U(\xi) = \frac{1}{c}\int_0^\infty (W_0(\xi' + \xi + a - x_0) - W_0(\xi' + \xi - x_0))\mathrm{e}^{-\xi'/c}d\xi'. \tag{3.5}$$

Equation (3.5) can be used to determine the asymptotic behavior of the solitary pulse for any exponentially decaying weight distribution. In particular, suppose that

$$w_0(x) = \mathrm{e}^{-\sigma|x|}g(x),$$

with $g(x)$ bounded for all x and $\lim_{x\to\pm\infty} g(x) = g_{\pm\infty}$. If $\xi > x_0$, then there is a common factor of $\mathrm{e}^{-\sigma\xi}$ on the right-hand side of (3.5), which can be taken outside the integral. Hence, in the limit $\xi \to \infty$,

$$U(\xi) \sim \frac{g_\infty}{c}\int_0^\infty (\mathrm{e}^{-\sigma(\xi'+\xi+a-x_0)} - \mathrm{e}^{-\sigma(\xi'+\xi-x_0)})\mathrm{e}^{-\xi'/c}d\xi' \sim -\mathrm{e}^{-\sigma\xi}. \tag{3.6}$$

On the other hand, when $\xi < x_0$, we have to partition the integral of (3.5) into the separate domains $\xi' > \xi + x_0$, $\xi + x_0 - a < \xi' < \xi + x_0$, and $\xi' < \xi + x_0 - a$ so that in the limit $\xi \to -\infty$,

$$U(\xi) \sim -\left[U_1\mathrm{e}^{\sigma\xi} + U_2\mathrm{e}^{\xi/c}\right]. \tag{3.7}$$

Therefore, the leading edge of the pulse profile decays at the rate σ determined by the weight distribution w_0, whereas the trailing edge decays at the rates σ and c^{-1}. The activity profile $U(\xi)$ of both the leading and trailing edges is negative due to the effects of inhibition; see Figure 3.1b. If $g(x)$ is taken to be an oscillatory function, then the asymptotic terms $\mathrm{e}^{-\sigma|\xi|}$ will also be oscillatory.

In Figure 3.2 we show existence curves for a traveling pulse of width a and speed c with w_0 given by the difference-of-exponentials (2.8). The pulse profile is given by $U(\xi) = U_{\sigma_E}(\xi) - \Gamma U_{\sigma_I}(\xi)$ with

$$\sigma U_\sigma(\xi) = \frac{\mathrm{e}^{-(\xi-x_0)\sigma}}{c\sigma+1} - \frac{\mathrm{e}^{-(\xi+a-x_0)\sigma}}{c\sigma+1}$$

for $\xi > x_0$,

$$\sigma U_\sigma(\xi) = 2 - \frac{\mathrm{e}^{-(\xi+a-x_0)\sigma}}{c\sigma+1} - \left[\left(\frac{2c^2\sigma^2}{c^2\sigma^2-1}\right)\mathrm{e}^{-(x_0-\xi)/c} - \frac{\mathrm{e}^{-(x_0-\xi)\sigma}}{c\sigma-1}\right]$$

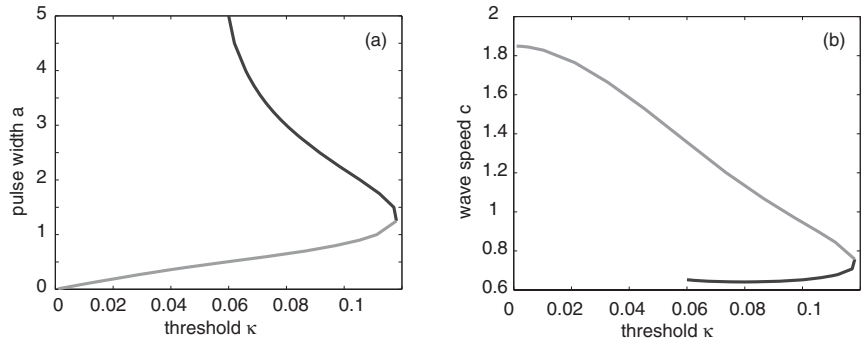


FIG. 3.2. *Existence of right-moving traveling pulses in the case of a shifted weight distribution* $w(x) = w_0(x - x_0)$ *with* $w_0(x)$ *given by the difference-of-exponentials* (2.8) *for* $\sigma_E = 1.8, \sigma_I = 1.0, \Gamma = 0.5$, *and* $x_0 = 0.5$. (a) *Plot of pulse width* a *against threshold* κ. (b) *Plot of wave speed* c *against threshold* κ. *Black (gray) curves denote stable (unstable) branches.*

for $x_0 - a < \xi < x_0$, and

$$\sigma U_\sigma(\xi) = \left[\left(\frac{2c^2\sigma^2}{c^2\sigma^2 - 1}\right) \mathrm{e}^{-(x_0-\xi-a)/c} - \frac{\mathrm{e}^{-(x_0-\xi-a)\sigma}}{c\sigma - 1}\right] - \left[\left(\frac{2c^2\sigma^2}{c^2\sigma^2 - 1}\right) \mathrm{e}^{-(x_0-\xi)/c} - \frac{\mathrm{e}^{-(x_0-\xi)\sigma}}{c\sigma - 1}\right]$$

for $\xi < x_0 - a$. Note that there are two existence branches corresponding, respectively, to narrow fast waves and wide slow waves. Given that wide pulses are stable in the stationary case (see Figure 2.1), we expect the slow branch to be stable, as can be confirmed numerically. This should be contrasted with traveling pulses in excitatory networks, where the fast branch is stable (see section 4). For the parameter values chosen in Figure 3.2, a stable pulse has a speed lying within the interval $0.6 < c < 0.8$ so that $\sigma_I < c^{-1} < \sigma_E$. Thus the dominant rate of decay for both the leading and trailing edges is σ_I. Note that if the units of length and time are taken to be 200μm and 10ms, respectively, then $c = 1$ corresponds to a wave speed of $2\mathrm{cms}^{-1}$, which is consistent with the range of speeds observed experimentally in cortical slices [20].

3.2. Singular perturbation theory. Suppose that there is a set of well separated exponentially decaying right-moving pulses. Following Elphick, Meron, and Spiegel [10], we now extend the singular perturbation theory of stationary pulses by working in the moving frame $\xi = x - ct$, where c is the speed of an isolated pulse. That is, we search for a traveling N-pulse solution with individual pulses having centers at $\xi_n = nd + \phi_n(\tau)$, where $\tau = \varepsilon t$ and $\varepsilon = \mathrm{e}^{-\rho d}$ with $\rho = \min\{\sigma, c^{-1}\}$:

$$u(\xi, \tau) = \sum_{n=1}^{N} U(\xi - nd - \phi_n(\tau)) + \varepsilon R(\xi, \tau). \tag{3.8}$$

Substituting (3.8) into (2.1) with $\partial_t \to \partial_t + \varepsilon\partial_\tau$, and performing an expansion in ε along lines identical to those of section 2, leads to the inhomogeneous equation (2.19), with the modified linear operator

$$\widehat{L}\psi = \psi - c\partial_\xi\psi - w * \left[\delta\left(\sum_{n=1}^{N} U_n - \kappa\right)\psi\right]. \tag{3.9}$$

The corresponding adjoint operator is now

$$\widehat{L}^{\dagger}\psi = \psi + c\partial_{\xi}\psi - \delta\left(\sum_{n=1}^{N} U_n - \kappa\right) w^T * \psi, \tag{3.10}$$

where $w^T(\xi) = w(-\xi) \neq w(\xi)$, since w is asymmetric. By differentiating (3.1), it can be seen that for largely separated pulses the functions U_n' are $\mathcal{O}(\varepsilon)$ null vectors of the operator $\widehat{L}$. This motivates us to seek $\mathcal{O}(\varepsilon)$ null vectors of the adjoint operator (3.10).

Proceeding along lines identical to those of section 2, we first decompose $\widehat{L}^{\dagger}$ according to (2.30) with

$$\widehat{L}_n^{\dagger}\psi = \psi + c\partial_{\xi}\psi - \delta(U_n - \kappa)w^T * \psi \tag{3.11}$$

and

$$\delta(U_n - \kappa) = \frac{\delta(\xi + a)}{U'(-a)} + \frac{\delta(\xi)}{|U'(0)|}. \tag{3.12}$$

We then look for null vectors $\mathcal{P}$ of $\widehat{L}_0^{\dagger}$ with $\xi_0 = 0$:

$$\mathcal{P}(\xi) + c\partial_{\xi}\mathcal{P}(\xi) = \left[\frac{\delta(\xi + a)}{U'(-a)} + \frac{\delta(\xi)}{|U'(0)|}\right] \int_{-\infty}^{\infty} w(\xi' - \xi)\mathcal{P}(\xi')d\xi'. \tag{3.13}$$

This has the formal solution

$$\mathcal{P}(\xi) = p_1 H(\xi + a)\mathrm{e}^{-(\xi+a)/c} - p_2 H(\xi)\mathrm{e}^{-\xi/c}, \tag{3.14}$$

with

$$p_1 c = \frac{1}{U'(-a)}\left[p_1 \int_0^{\infty} w(\xi)\mathrm{e}^{-\xi/c}d\xi - p_2 \int_0^{\infty} w(\xi + a)\mathrm{e}^{-\xi/c}d\xi\right]$$

and

$$p_2 c = -\frac{1}{|U'(0)|}\left[p_1 \int_0^{\infty} w(\xi - a)\mathrm{e}^{-\xi/c}d\xi - p_2 \int_0^{\infty} w(\xi)\mathrm{e}^{-\xi/c}d\xi\right].$$

Up to a scalar multiplication, the pair of algebraic equations for the coefficients p_1, p_2 has the solution

$$p_1 = \int_0^{\infty} w(\xi + a)\mathrm{e}^{-\xi/c}d\xi, \qquad p_2 = \int_0^{\infty} w(\xi - a)\mathrm{e}^{-\xi/c}d\xi. \tag{3.15}$$

In order to prove this, differentiate (3.3) with respect to ξ using (2.5):

$$U'(\xi) = \frac{1}{c}\int_0^{\infty} (w(\xi' + \xi + a) - w(\xi' + \xi))\mathrm{e}^{-\xi'/c}d\xi'. \tag{3.16}$$

Setting $\xi = 0$ and $\xi = -a$ then leads to the following equations for $U'(-a)$ and $|U'(0)|$:

$$\begin{aligned} U'(-a) &= \frac{1}{c}\int_0^{\infty} [w(\xi) - w(\xi - a)]\,\mathrm{e}^{-\xi/c}d\xi, \\ |U'(0)| &= \frac{1}{c}\int_0^{\infty} [w(\xi) - w(\xi + a)]\,\mathrm{e}^{-\xi/c}d\xi. \end{aligned} \tag{3.17}$$

It is now straightforward to verify (3.15).

Following the same arguments as section 2, we conclude that (2.28) has solutions of the form $P_n(\xi) = \mathcal{P}(\xi - \xi_n)$, with $\mathcal{P}$ given by (3.14) and (3.15). A dynamical equation for the pulse positions ξ_n can then be derived by taking the inner product of (2.19) with P_n, which yields an equation of the form (2.36). Substituting (3.14) and (3.12) into (2.36), we find that

$$\begin{aligned}\langle P_m \,|w * \delta(U_m - \kappa)[U_{m+1} + U_{m-1}]\rangle \\ = p_1 c\,(U(-a + \xi_m - \xi_{m+1}) + U(-a + \xi_m - \xi_{m-1})) \\ - p_2 c(U(\xi_m - \xi_{m+1}) + U(\xi_m - \xi_{m-1})),\end{aligned} \tag{3.18}$$

and $\langle P_m | U'_m \rangle = \langle \mathcal{P} | U' \rangle = K$, where

$$K = p_1 \int_0^\infty \mathrm{e}^{-\xi/c} U'(\xi - a) d\xi - p_2 \int_0^\infty \mathrm{e}^{-\xi/c} U'(\xi) d\xi. \tag{3.19}$$

Hence, (2.36) reduces to the form

$$\dot{\xi}_m = f_R(\xi_{m+1} - \xi_m) + f_L(\xi_m - \xi_{m-1}), \tag{3.20}$$

for $\xi_m(t) = nd + \phi_m(\varepsilon t)$, with

$$f_R(\xi) = \frac{c}{K}\left[p_2 U(-\xi) - p_1 U(-\xi - a)\right], \qquad f_L(\xi) = \frac{c}{K}\left[p_2 U(\xi) - p_1 U(\xi - a)\right], \tag{3.21}$$

and $U(\xi)$ determined from the underlying weight distribution according to (3.5).

3.3. Traveling wave trains. Lattice equations of the form (3.20) have been studied in considerable detail within the context of diffusive excitable systems and fluids [10, 2, 37]. Here we illustrate some of the basic results by explicitly calculating the interaction functions f_L, f_R. We define a traveling wave train as an N-pulse solution of (3.20) in which $\dot{\xi}_m = \delta c$ independently of m, where δc is a constant velocity in the frame of an isolated pulse. The spacings between pulses are then fixed, and we obtain the so-called pattern map [10]

$$\delta c = f_R(\Delta_{m+1}) + f_L(\Delta_m), \quad \Delta_m = \xi_m - \xi_{m-1}. \tag{3.22}$$

As a further simplification, we impose periodic boundary conditions by taking the pulses to be moving on a ring of length L with $\xi_0 = \xi_N$ and $\xi_{N+1} = \xi_0$. The simplest solution of (3.22) is then the fixed point $\Delta_m = \Delta = L/N$ for all m. The fixed point equation $\delta c = f_R(\Delta) + f_L(\Delta)$ determines the relationship between the total speed of the wave train $c + \delta c$ and the uniform spacing Δ between neighboring pulses. Linearizing (3.20) about the uniformly spaced wave train by setting $\xi_m = m\Delta + \delta c t + \theta_m$ gives

$$\dot{\theta}_m = f'_R(\Delta)\left[\alpha\theta_{m-1} - (1 + \alpha)\theta_m + \theta_{m+1}\right] \tag{3.23}$$

with $\alpha = -f'_L(\Delta)/f'_R(\Delta)$. This has eigensolutions of the form $\theta_m(t) = \mathrm{e}^{\lambda(k)t}\mathrm{e}^{imk}$

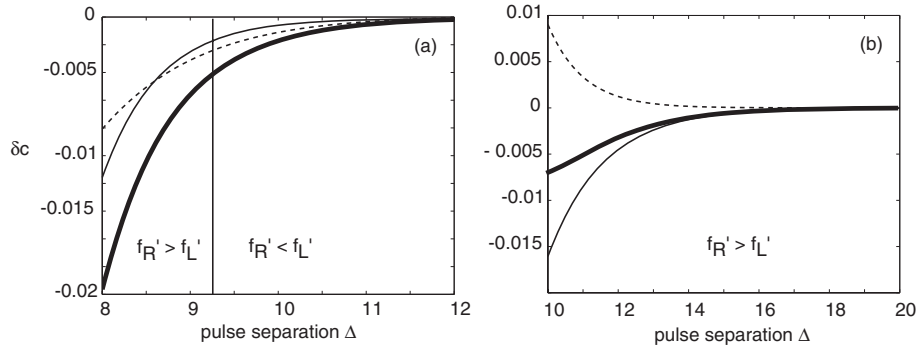


FIG. 3.3. *Plot of $\delta c = f_L(\Delta) + f_R(\Delta)$ against Δ (thick curves) in the case of a shifted weight distribution $w(x) = w_0(x - x_0)$, with $w_0(x)$ given by the difference-of-exponentials* (2.8) *for $\sigma_E = 1.8, \sigma_I = 1.0, \Gamma = 0.5$. Also shown are the plots of $f_L(\Delta)$ (dashed curves) and $f_R(\Delta)$ (thin curves).* (a) *$x_0 = 0.5$, $a = 5$, and $c = 0.65$.* (b) *$x_0 = 1$, $a = 4$, and $c = 1.58$. In both cases, the pulse width and speed of an isolated pulse are chosen to lie on the stable existence branch.*

with wavenumber $k = 2\pi p/N$ for $p = 1, \ldots, N$ and

$$\lambda(k) = -f_R'(\Delta)\left[(1+\alpha)(1-\cos(k)) \pm i(1-\alpha)\sin(k)\right]. \tag{3.24}$$

The condition for (marginal) stability of the uniformly spaced wave train is thus $f_R'(\Delta) > f_L'(\Delta)$. For the sake of illustration, consider the case of an asymmetric Mexican hat weight distribution $w(x) = w_0(x - x_0)$, with w_0 given by the difference-of-exponentials (2.8). Let the unperturbed pulse width a and speed c correspond to a solitary wave on the stable slow branch; see Figure 3.2. Two examples of dispersion curves δc versus Δ are shown in Figure 3.3. In (a), one sees that there exists a finite range of separations Δ for which $f_R'(\Delta) > f_L'(\Delta)$, corresponding to a finite band of stable wave trains. Similarly, (b) shows an example of a semi-infinite band of stable wave trains. In both examples, a given wave train moves more slowly than an isolated pulse, since $\delta c = f_L(\Delta) + f_R(\Delta) < 0$. Which wave train is actually selected will depend on initial conditions.

Suppose that we now allow for the possibility of an oscillatory weight distribution such as (2.11). If $\sigma_I < c^{-1}$, then the leading and trailing edges both consist of exponentially decaying spatial oscillations, so that for widely separated pulses the lattice dynamics takes the form

$$\begin{aligned}\dot{\xi}_m = {} & A_R \mathrm{e}^{-\sigma(\xi_{m+1}-\xi_m)} \cos(\omega(\xi_{m+1}-\xi_m) - \Phi_R) \\ & + A_L \mathrm{e}^{-\sigma(\xi_m - \xi_{m-1})} \cos(\omega(\xi_m - \xi_{m-1}) - \Phi_L).\end{aligned} \tag{3.25}$$

In the case of oscillatory interaction functions, the associated pattern map (3.22) can generate nontrivial sequences of pulse intervals $\{\ldots \Delta_{m-1}, \Delta_m, \Delta_{m+1}, \ldots\}$, including possibly chaotic sequences [17]. From a dynamical systems perspective, such wave trains can be reinterpreted in terms of nearly homoclinic orbits [2].

4. Traveling pulses in excitatory networks with adaptation. As our final example, let us return to the case of a symmetric weight distribution w, but now take w to be purely excitatory, as in the case of a disinhibited cortical slice [34]. In the absence of lateral inhibition, the scalar equation (2.1) no longer supports localized persistent states of activity but does exhibit traveling front solutions. In order to

obtain traveling localized pulses, it is necessary to introduce some form of adaptation. Therefore, following Pinto and Ermentrout [34], we extend the basic Amari model by considering the following system of equations:

$$\begin{aligned} \frac{\partial u(x,t)}{\partial t} &= -u(x,t) + \int_{-\infty}^{\infty} w(x-x')H(u(x',t)-\kappa)dx' - \beta v(x,t), \\ \frac{\partial v(x,t)}{\partial t} &= \gamma[-v(x,t)+u(x,t)], \end{aligned} \tag{4.1}$$

where $v(x,t)$ represents some form of negative feedback mechanism such as spike frequency adaptation or synaptic depression, with β, γ determining the relative strength and rate of feedback. We will extend the singular perturbation theory of section 3 in order to investigate traveling N-pulse solutions of (4.1) in the case of a positive exponentially decaying weight distribution w.

4.1. Traveling solitary pulses. In contrast to the asymmetric lateral inhibition network of section 3, the excitatory network given by (4.1) supports bidirectional wave propagation. Without loss of generality, let us consider a right-moving traveling pulse solution of the form $(u(x,t), v(x,t)) = (U(x-ct), V(x-ct))$ with $U(\pm\infty), V(\pm\infty) = 0$ and $U(-a) = U(0) = \kappa$. Substituting into (4.1) with $\xi = x - ct$ gives

$$\begin{aligned} -c\partial_\xi U(\xi) + U(\xi) + \beta V(\xi) &= \int_{-\infty}^{\infty} w(\xi-\xi')H(U(\xi')-\kappa)d\xi', \\ -c\partial_\xi V(\xi) + \gamma[V(\xi)-U(\xi)] &= 0. \end{aligned} \tag{4.2}$$

It is useful to rewrite (4.2) in the matrix form

$$\begin{pmatrix} 1 & \beta \\ -\gamma & \gamma \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix} - c\partial_\xi \begin{pmatrix} U \\ V \end{pmatrix} = [W(\xi+a) - W(\xi)] \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{4.3}$$

We proceed by diagonalizing the left-hand side of (4.3) using the right eigenvectors $\mathbf{v}$ of the matrix

$$\mathbf{M} = \begin{pmatrix} 1 & \beta \\ -\gamma & \gamma \end{pmatrix}. \tag{4.4}$$

These are given by

$$\mathbf{v}_\pm = \begin{pmatrix} \gamma - \lambda_\pm \\ \gamma \end{pmatrix}, \tag{4.5}$$

with corresponding eigenvalues

$$\lambda_\pm = \frac{1}{2}\left[1 + \gamma \pm \sqrt{(1+\gamma)^2 - 4\gamma(1+\beta)}\right]. \tag{4.6}$$

Note that $\mathbf{v}_\pm e^{\lambda_\pm \xi/c}$ are the corresponding null vectors of the linear operator on the left-hand side of (4.3); that is, they generate the complementary solution. Performing the transformation

$$\begin{pmatrix} \widetilde{U} \\ \widetilde{V} \end{pmatrix} = \mathbf{T}^{-1} \begin{pmatrix} U \\ V \end{pmatrix}, \quad \mathbf{T} = \begin{pmatrix} \mathbf{v}_+ & \mathbf{v}_- \end{pmatrix}, \tag{4.7}$$

then gives the pair of equations

$$-c\partial_\xi \widetilde{U} + \lambda_+ \widetilde{U} = \eta_+[W(\xi+a) - W(\xi)],$$
$$-c\partial_\xi \widetilde{V} + \lambda_- \widetilde{V} = \eta_-[W(\xi+a) - W(\xi)], \tag{4.8}$$

with $\eta_\pm = \mp 1/(\lambda_+ - \lambda_-)$. Integrating the equation for $\widetilde{U}$ from $-a$ to ∞, we have

$$\widetilde{U}(\xi) = \mathrm{e}^{\lambda_+\xi/c}\left[\widetilde{U}(-a)\mathrm{e}^{a\lambda_+/c} - \frac{\eta_+}{c}\int_{-a}^{\xi} \mathrm{e}^{-\lambda_+\xi'/c}[W(\xi'+a) - W(\xi')]d\xi'\right]. \tag{4.9}$$

Finiteness of $\widetilde{U}$ in the limit $\xi \to \infty$ requires that the term in square brackets cancel. Hence, we can eliminate $\widetilde{U}(-a)$ to obtain the result

$$\widetilde{U}(\xi) = \frac{\eta_+}{c}\int_0^\infty \mathrm{e}^{-\lambda_+\xi'/c}[W(\xi'+\xi+a) - W(\xi'+\xi)]d\xi'. \tag{4.10}$$

Similarly,

$$\widetilde{V}(\xi) = \frac{\eta_-}{c}\int_0^\infty \mathrm{e}^{-\lambda_-\xi'/c}[W(\xi'+\xi+a) - W(\xi'+\xi)]d\xi'. \tag{4.11}$$

Performing the inverse transformation $U = (\gamma - \lambda_+)\widetilde{U} + (\gamma - \lambda_-)\widetilde{V}$, we have

$$U(\xi) = \frac{1}{c}\int_0^\infty \left[\chi_+\mathrm{e}^{-\lambda_+\xi'/c} + \chi_-\mathrm{e}^{-\lambda_-\xi'/c}\right][W(\xi'+\xi+a) - W(\xi'+\xi)]d\xi', \tag{4.12}$$

with $\chi_\pm = (\gamma - \lambda_\pm)\eta_\pm$. Using $\lambda_+ + \lambda_- = 1 + \gamma$, we can rewrite $\chi_\pm$ as

$$\chi_+ = \frac{1-\lambda_-}{\lambda_+ - \lambda_-}, \qquad \chi_- = \frac{\lambda_+ - 1}{\lambda_+ - \lambda_-}. \tag{4.13}$$

The threshold conditions $U(-a) = \kappa$ and $U(0) = \kappa$ then yield a pair of equations whose solutions determine existence curves relating the speed c and width a of a pulse to the threshold κ [34].

For the sake of illustration, let w be given by the exponential function (2.8) with $\Gamma = 0$ and $\sigma_E = \sigma$; that is, $w(x) = \mathrm{e}^{-\sigma|x|}$. In the domain $\xi > 0$, there is a common factor of $\mathrm{e}^{-\sigma\xi}$ in the integrand of (4.12) so that $U(\xi) = \kappa\mathrm{e}^{-\sigma\xi}$ for $\xi > 0$, provided that

$$\kappa = \frac{(c\sigma + \gamma)(1 - \mathrm{e}^{-a\sigma})}{c^2\sigma^2 + c\sigma(1+\gamma) + \gamma(1+\beta)}. \tag{4.14}$$

On the other hand, when $\xi < 0$, one has to partition the integral of (4.12) into the separate domains $\xi' > |\xi|$, $|\xi| - a < \xi' < |\xi|$, and $\xi' < |\xi| - a$. This then determines the second threshold condition as well as the asymptotic behavior of $U(\xi)$ in the limit $\xi \to -\infty$:

$$U(\xi) = U_+\mathrm{e}^{\lambda_+\xi/c} + U_-\mathrm{e}^{\lambda_-\xi/c} + U_0\mathrm{e}^{\sigma\xi}, \tag{4.15}$$

where the amplitudes $U_\pm$ and U_0 can be determined from matching conditions at the threshold crossing points [34, 15]. Note that the leading edge of the pulse is positive, whereas the trailing edge is negative due to the effects of adaptation. One finds that for sufficiently slow negative feedback (small γ) and large β there exist two pulse solutions, one narrow and slow and the other wide and fast. This is illustrated in Figure 4.1. Numerically, the fast solution is found to be stable [34]. Its stability can also be established analytically using Evans function techniques [44, 8, 16].

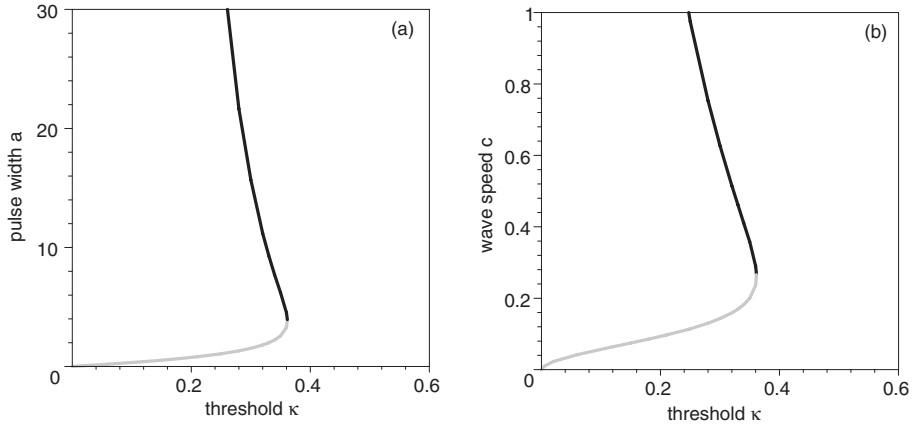


FIG. 4.1. *Existence of right-moving traveling pulses in the case of the excitatory network* (4.1) *for an exponential weight distribution with* $w(x) = e^{-\sigma|x|}$. *Here* $\sigma = 1, \gamma = 0.01$, *and* $\beta = 2.5$. (a) *Plot of pulse width* a *against threshold* κ. (b) *Plot of wave speed* c *against threshold* κ. *Stable (unstable) branches indicated by black (gray) curves.*

4.2. Singular perturbation theory. Suppose that (4.1) has a stable right-moving pulse solution $U(\xi)$ of width a and speed c. Following the model set by section 3, we search for a traveling N-pulse solution with individual pulses having centers at $\xi_n = nd + \phi_n(\tau)$, where $\tau = \varepsilon t$ and $\varepsilon = \mathrm{e}^{-\rho d}$ with $\rho = \min\{c^{-1}\lambda_{\pm}, \sigma\}$:

$$
\begin{aligned}
u(\xi,\tau) &= \sum_{n=1}^{N} U(\xi - nd - \phi_n(\tau)) + \varepsilon R(\xi,\tau),\\
v(\xi,\tau) &= \sum_{n=1}^{N} V(\xi - nd - \phi_n(\tau)) + \varepsilon \overline{R}(\xi,\tau).
\end{aligned}
\tag{4.16}
$$

Substituting (4.16) into (4.1) with $\partial_t \to \partial_t + \varepsilon\partial_\tau$, and using (4.2), gives

$$
\begin{aligned}
&-c\varepsilon\partial_\xi R + \varepsilon^2\partial_\tau R - \varepsilon\sum_{n=1}^{N}\dot{\phi}_n U_n'\\
&\qquad = -\varepsilon(R + \beta\overline{R}) + w * H\left(\sum_{n=1}^{N} U_n + \varepsilon R - \kappa\right) - w * \sum_{n=1}^{N} H(U_n - \kappa),\\
&-c\varepsilon\partial_\xi \overline{R} + \varepsilon^2\partial_\tau \overline{R} - \varepsilon\sum_{n=1}^{N}\dot{\phi}_n V_n' = -\varepsilon\gamma(\overline{R} - R).
\end{aligned}
\tag{4.17}
$$

Performing an expansion to $\mathcal{O}(\varepsilon)$ along lines identical to section 2 leads to the inhomogeneous equation

$$
\widehat{L}\begin{pmatrix} R \\ \overline{R} \end{pmatrix} = \sum_{n=1}^{N}\dot{\phi}_n\begin{pmatrix} U_n' \\ V_n' \end{pmatrix} + \frac{1}{\varepsilon} w * \left[H\left(\sum_{n=1}^{N} U_n - \kappa\right) - \sum_{n=1}^{N} H(U_n - \kappa)\right]\begin{pmatrix} 1 \\ 0 \end{pmatrix},
\tag{4.18}
$$

with the linear operator $\widehat{L}$ given by

$$\widehat{L}\psi = \begin{pmatrix} 1 & \beta \\ -\gamma & \gamma \end{pmatrix} \psi - c\partial_\xi \psi - w * \left[\delta \left(\sum_{n=1}^{N} U_n - \kappa \right) \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \psi \right], \tag{4.19}$$

where ψ now denotes a two-vector rather than a scalar. Differentiating (4.2) and using arguments similar to those of section 2, it is straightforward to show that $\widehat{L}(U_n', V_n')^{tr} = \mathcal{O}(\varepsilon)$ for all $n = 1, \ldots, N$. This again motivates us to seek $\mathcal{O}(\varepsilon)$ null vectors of the adjoint operator.

In order to determine the corresponding solvability conditions on (4.18), we seek weak solutions $(P, \overline{P})$ of the equation

$$\left\langle \widehat{L}^\dagger \begin{pmatrix} P \\ \overline{P} \end{pmatrix} \middle| \begin{pmatrix} Q \\ \overline{Q} \end{pmatrix} \right\rangle = \mathcal{O}(\varepsilon) \tag{4.20}$$

for arbitrary bounded functions $Q, \overline{Q}$, where $\widehat{L}^\dagger$ is the adjoint operator

$$\widehat{L}^\dagger \psi = \begin{pmatrix} 1 & -\gamma \\ \beta & \gamma \end{pmatrix} \psi + c\partial_\xi \psi - \delta \left(\sum_{n=1}^{N} U_n - \kappa \right) w * \left[\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \psi \right]. \tag{4.21}$$

The inner product is defined by first taking the dot product of the two vectors and then integrating over $\mathbf{R}$. Using the perturbation expansion (2.29) with $\delta(U_n - \kappa)$ given by (3.12), we obtain the following formal decomposition:

$$\widehat{L}^\dagger \psi = \widehat{L}_n^\dagger \psi - \sum_{j \neq n} \delta(U_j - \kappa) w * \left[\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \psi \right] + \mathcal{O}(\varepsilon) \tag{4.22}$$

for any $n = 1, \ldots, N$, with

$$\widehat{L}_n^\dagger \psi = \begin{pmatrix} 1 & -\gamma \\ \beta & \gamma \end{pmatrix} \psi + c\partial_\xi \psi - \delta(U_n - \kappa) w * \left[\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \psi \right]. \tag{4.23}$$

We now look for null vectors of $\widehat{L}_0^\dagger$ with $\xi_0 = 0$. We proceed by partially diagonalizing $\widehat{L}_0^\dagger$ using the left eigenvectors $\widetilde{\mathbf{v}}$ of the matrix $\mathbf{M}$ (see (4.4)):

$$\widetilde{\mathbf{v}}_\pm = \begin{pmatrix} \gamma \\ 1 - \lambda_\pm \end{pmatrix}. \tag{4.24}$$

Introducing the transformation

$$\begin{pmatrix} \mathcal{Q} \\ \overline{\mathcal{Q}} \end{pmatrix} = \widetilde{\mathbf{T}}^{-1} \begin{pmatrix} \mathcal{P} \\ \overline{\mathcal{P}} \end{pmatrix}, \quad \widetilde{\mathbf{T}} = \begin{pmatrix} \widetilde{\mathbf{v}}_+ & \widetilde{\mathbf{v}}_- \end{pmatrix}, \tag{4.25}$$

then leads to the following pair of equations:

$$c\partial_\xi \mathcal{Q} + \lambda_+ \mathcal{Q} = \chi_+ \delta(U - \kappa) w * [\mathcal{Q} + \overline{\mathcal{Q}}], \tag{4.26}$$

$$c\partial_\xi \overline{\mathcal{Q}} + \lambda_- \overline{\mathcal{Q}} = \chi_- \delta(U - \kappa) w * [\mathcal{Q} + \overline{\mathcal{Q}}], \tag{4.27}$$

with $\chi_\pm$ defined by (4.13). Using an analysis similar to that of section 3, we obtain the solution

$$\begin{aligned} \mathcal{Q}(\xi) &= \chi_+ \left[p_1 H(\xi + a) \mathrm{e}^{-\lambda_+(\xi+a)/c} - p_2 H(\xi) \mathrm{e}^{-\lambda_+ \xi/c} \right], \\ \overline{\mathcal{Q}}(\xi) &= \chi_- \left[p_1 H(\xi + a) \mathrm{e}^{-\lambda_-(\xi+a)/c} - p_2 H(\xi) \mathrm{e}^{-\lambda_- \xi/c} \right], \end{aligned} \tag{4.28}$$

with p_1, p_2 given by (3.15) and $U'(0), U'(-a)$ satisfying the self-consistency conditions

$$c = \frac{1}{U'(-a)} \int_0^\infty [w(\xi) - w(\xi - a)] \left[\chi_+ \mathrm{e}^{-\lambda_+ \xi/c} + \chi_- \mathrm{e}^{-\lambda_- \xi/c}\right] d\xi,$$

$$c = \frac{1}{U'(0)} \int_0^\infty [w(\xi + a) - w(\xi)] \left[\chi_+ \mathrm{e}^{-\lambda_+ \xi/c} + \chi_- \mathrm{e}^{-\lambda_- \xi/c}\right] d\xi. \tag{4.29}$$

The latter follow immediately from differentiating (4.12) and setting $\xi = 0, -a$. Finally, we perform the inverse transformation on $\mathcal{Q}, \overline{\mathcal{Q}}$ to obtain $\mathcal{P}, \overline{\mathcal{P}}$:

$$\begin{pmatrix} \mathcal{P}(\xi) \\ \overline{\mathcal{P}}(\xi) \end{pmatrix} = \chi_+ \begin{pmatrix} \gamma \\ 1 - \lambda_+ \end{pmatrix} \left[p_1 H(\xi + a) \mathrm{e}^{-\lambda_+(\xi+a)/c} - p_2 H(\xi) \mathrm{e}^{-\lambda_+ \xi/c}\right] + \chi_- \begin{pmatrix} \gamma \\ 1 - \lambda_- \end{pmatrix} \left[p_1 H(\xi + a) \mathrm{e}^{-\lambda_-(\xi+a)/c} - p_2 H(\xi) \mathrm{e}^{-\lambda_- \xi/c}\right]. \tag{4.30}$$

From translation symmetry, it follows that $\widehat{L}_n^\dagger$ has the null vector $(P_n, \overline{P}_n)$ with

$$P_n(\xi) = \mathcal{P}(\xi - \xi_n), \quad \overline{P}_n(\xi) = \overline{\mathcal{P}}(\xi - \xi_n). \tag{4.31}$$

Hence, applying the decomposition (4.22), we see that

$$\left\langle \widehat{L}^\dagger \begin{pmatrix} P_n \\ \overline{P}_n \end{pmatrix} \middle| \begin{pmatrix} Q \\ \overline{Q} \end{pmatrix} \right\rangle = -\sum_{j \neq n} \langle \delta(U_j - \kappa) w * P_n | Q \rangle + \mathcal{O}(\varepsilon). \tag{4.32}$$

Equations (4.30) and (4.31) imply that $P_n, \overline{P}_n$ are zero for $\xi < \xi_n - a$ and exponentially decaying for $\xi > \xi_n - a$. Evaluating the inner product on the right-hand side of (4.32) establishes that it is also $\mathcal{O}(\varepsilon)$. We conclude that (4.20) has the set of solutions $(P_n, \overline{P}_n)$, $n = 1, \ldots, N$. We now take the inner product of (4.18) with respect to the vector $(P_m, \overline{P}_m)$ for some integer $m, m = 1, \ldots, N$, and use (2.23):

$$\dot{\phi}_m \left\langle \begin{pmatrix} P_m \\ \overline{P}_m \end{pmatrix} \middle| \begin{pmatrix} U_m' \\ V_m' \end{pmatrix} \right\rangle + \frac{1}{\varepsilon} \langle P_m | w * \delta(U_m - \kappa)[U_{m+1} + U_{m-1}] \rangle = \mathcal{O}(\varepsilon). \tag{4.33}$$

Evaluating the various inner products using (4.30) and (3.12) leads to the same (3.20) and (3.21) as the scalar case, with

$$K = \gamma^{-1} \left\langle \begin{pmatrix} \mathcal{P} \\ \overline{\mathcal{P}} \end{pmatrix} \middle| \begin{pmatrix} U' \\ V' \end{pmatrix} \right\rangle. \tag{4.34}$$

However, there is a significant difference in the asymptotic behavior of the interaction functions f_L, f_R when compared to the scalar case. This is due to the fact that adaptation is slow ($\gamma \ll 1$) so that $\lambda_- \approx 0$, and thus the leading edge decays much faster than the trailing edge; see (4.6) and (4.15). Hence, we can neglect the interaction term f_L in (3.20) to obtain

$$\dot{\xi}_m = f_R(\xi_{m+1} - \xi_m) \tag{4.35}$$

with $f_R(\Delta) \sim -\mathrm{e}^{-\lambda_- \Delta/c}$. In this case the dynamics of the pulse position ξ_m depends only on the distance to the proceeding pulse $\Delta_m = \xi_{m+1} - \xi_m$ and can thus be reformulated within a kinematic framework [11, 33]. This is based on the observation that the function f_R directly determines the dispersion relation between the speed C and the pulse separation Δ of a uniformly spaced wave train, $C(\Delta) = f_R(\Delta) + c$. Thus

$$\dot{\xi_m} = C(\Delta_m) - c. \tag{4.36}$$

The condition for stability of a uniform wave train on a ring is then $f_R'(\Delta) > 0$.

5. Discussion. In this paper we have used perturbation methods to develop a theory of weakly interacting pulses in one-dimensional neuronal networks. We have shown how the pulse interactions explicitly depend on the form of the long-range synaptic coupling, and investigated how this determines the existence and stability of multipulse solutions. For simplicity, we have assumed throughout that the network is homogeneous: the coupling depends only on the distance between interacting elements in the network, and external inputs have been ignored. In a recent series of papers, we have shown that introducing a localized inhomogeneous input can generate oscillatory coherent states in the form of standing and traveling breathing pulses [5, 15, 16]. It would be interesting to develop a theory of weakly interacting breathers and to determine under what conditions long-range synaptic coupling can provide a mechanism for synchronizing the oscillations between breathers. This would provide an alternative way of thinking about stimulus-induced coherent oscillations in cortex, which are observed *in vivo* during periods of sensory processing [21, 14].

REFERENCES

[1] S. Amari, *Dynamics of pattern formation in lateral inhibition type neural fields*, Biol. Cybernet., 27 (1977), pp. 77–87.

[2] N. J. Balmforth, *Solitary waves and homoclinic orbits*, Ann. Rev. Fluid Mech., 27 (1995), pp. 335–373.

[3] N. J. Balmforth, G. R. Ierley, and R. Worthing, *Pulse dynamics in an unstable medium*, SIAM J. Appl. Math., 57 (1997), pp. 205–251.

[4] P. C. Bressloff, *Traveling waves and pulses in a one-dimensional network of excitable integrate-and-fire neurons*, J. Math. Biol., 40 (2000), pp. 169–198.

[5] P. C. Bressloff, S. Folias, A. Pratt, and Y.-X. Li, *Oscillatory waves in inhomogeneous neural media*, Phys. Rev. Lett., 91 (2003), paper 178101.

[6] R. D. Chervin, P. A. Pierce, and B. W. Connors, *Periodicity and directionality in the propagation of epileptiform discharges across neocortex*, J. Neurophysiol., 60 (1988), pp. 1695–1713.

[7] S. Coombes, G. J. Lord, and M. R. Owen, *Waves and bumps in neuronal networks with axo-dendritic synaptic interactions*, Phys. D, 178 (2003), pp. 219–241.

[8] S. Coombes and M. R. Owen, *Evans functions for integral neural field equations with Heaviside firing rate function*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 574–600.

[9] C. Elphick, E. Meron, and E. A. Spiegel, *Spatiotemporal complexity in traveling patterns*, Phys. Rev. Lett., 61 (1988), pp. 496–499.

[10] C. Elphick, E. Meron, and E. A. Spiegel, *Patterns of propagating pulses*, SIAM J. Appl. Math., 50 (1990), pp. 490–503.

[11] C. Elphick, E. Meron, J. Rinzel and E. A. Spiegel, *Impulse patterning and relaxational propagation in excitable media*, J. Theoret. Biol., 146 (1990), pp. 249–268.

[12] G. B. Ermentrout, *Neural networks as spatial pattern forming systems*, Rep. Progr. Phys., 61 (1998), pp. 353–430.

[13] G. B. Ermentrout, *The analysis of synaptically generated traveling waves*, J. Comp. Neurosci., 5 (1998), pp. 191–208.

[14] G. B. Ermentrout and D. Kleinfeld, *Traveling electrical waves in cortex: Insights from phase dynamics and speculation on a computational role*, Neuron, 29 (2001), pp. 33–44.

[15] S. E. Folias and P. C. Bressloff, *Breathing pulses in an excitatory neural network*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 378–407.

[16] S. E. Folias and P. C. Bressloff, *Stimulus-locked traveling pulses and breathers in an excitatory neural network*, SIAM J. Appl. Math., 65 (2005), pp. 2067–2092.

[17] A. C. Fowler and C. T. Sparrow, *Bifocal homoclinic orbits in four dimensions*, Nonlinearity, 4 (1991), pp. 1159–1182.

[18] J. M. Fuster and G. Alexander, *Neuron activity related to short-term memory*, Science, 173 (1971), pp. 652–654.

[19] C. D. Gilbert and T. N. Wiesel, *Clustered intrinsic connections in cat visual cortex*, J. Neurosci., 3 (1983), pp. 1116–1133.

[20] D. Golomb and Y. Amitai, *Propagating neuronal discharges in neocortical slices: Computational and experimental study*, J. Neurophysiol., 78 (1997), pp. 1199–1211.

[21] C. M. Gray, *Synchronous oscillations in neuronal systems: Mechanisms and functions*, J. Comput. Neurosci., 1 (1994), pp. 11–38.
[22] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[23] J. P. Keener, *Waves in excitable media*, SIAM J. Appl. Math., 39 (1980), pp. 528–548.
[24] C. R. Laing and C. C. Chow, *Stationary bumps in networks of spiking neurons*, Neural Comp., 13 (2001), pp. 1473–1494.
[25] C. R. Laing, W. C. Troy, B. Gutkin, and G. B. Ermentrout, *Multiple bumps in a neuronal model of working memory*, SIAM J. Appl. Math, 63 (2002), pp. 62–97.
[26] C. R. Laing and W. C. Troy, *Two-bump solutions of Amari-type models of neuronal pattern formation*, Phys. D, 178 (2003), pp. 190–218.
[27] C. R. Laing and W. C. Troy, *PDE methods for nonlocal models*, SIAM J. Appl. Dynam. Syst., 2 (2003), pp. 487–516.
[28] J. B. Levitt, D. A. Lewis, T. Yoshioka, and J. S. Lund, *Topography of pyramidal neuron intrinsic connections in macaque prefrontal cortex*, J. Comput. Neurol., 338 (1993), pp. 360–376.
[29] J. S. Lund, A. Angelucci, and P. C. Bressloff, *Anatomical substrates for functional columns in macaque monkey primary visual cortex*, Cerebral Cortex, 12 (2003), pp. 15–24.
[30] R. Maex and G. A. Orban, *Model circuit of spiking neurons generating directional selectivity in simple cells*, J. Neurophysiol., 75 (1996), pp. 1515–1545.
[31] B. A. McGuire, C. D. Gilbert, P. K. Rivlin, and T. N. Wiesel, *Targets of horizontal connections in macaque primary visual cortex*, J. Comput. Neurol., 305 (1991), pp. 370–392.
[32] P. Mineiro and D. Zipser, *Analysis of direction selectivity arising from recurrent cortical interactions*, Neural Comput., 10 (1998), pp. 353–371.
[33] M. Or-Guil, I. G. Kevrekidis, and M. Bar, *Stable bound states of pulses in an excitable medium*, Phys. D, 135 (2000), pp. 154–174.
[34] D. J. Pinto and G. B. Ermentrout, *Spatially structured activity in synaptically coupled neuronal networks*: I. *Traveling fronts and pulses*, SIAM J. Appl. Math., 62 (2001), pp. 206–225.
[35] D. J. Pinto and G. B. Ermentrout, *Spatially structured activity in synaptically coupled neuronal networks*: II. *Lateral inhibition and standing pulses*, SIAM J. Appl. Math., 62 (2001), pp. 226–243.
[36] K. S. Rockland and J. S. Lund, *Intrinsic laminar lattice connections in primate visual cortex*, J. Comput. Neurol., 216 (1983), pp. 303–318.
[37] C. P. Schenk, P. Schutz, M. Bode, and H. G. Purwins, *Interaction of self-organized quasi-particles in a two-dimensional reaction-diffusion system, The formation of molecules*, Phys. Rev. E, 57 (1998), pp. 6480–6486.
[38] H. Suarez, C. Koch, and R. Douglas, *Modeling direction selectivity of simple cells in striate visual cortex within the framework of the canonical microcircuit*, J. Neurosci., 15 (1995), pp. 6700–6719.
[39] X.-J. Wang, *Synaptic reverberation underlying mnemonic persistent activity*, Trends Neurosci., 24 (2001), pp. 455–463.
[40] H. Werner and T. Richter, *Circular stationary solutions in two-dimensional neural fields*, Biol. Cybernet., 85 (2001), pp. 211–217.
[41] H. R. Wilson and J. D. Cowan, *A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue*, Kybernetik, 13 (1973), pp. 55–80.
[42] J-Y. Wu, L. Guan, and Y. Tsau, *Propagating activation during oscillations and evoked responses in neocortical slices*, J. Neurosci., 19 (1999), pp. 5005–5015.
[43] X. Xie and M. A. Giese, *Nonlinear dynamics of direction-selective recurrent neural media*, Phys. Rev. E, 65 (2002), paper 051904.
[44] L. Zhang, *On the stability of traveling wave solutions in synaptically coupled neuronal networks*, Differential Integral Equations, 16 (2003), pp. 513–536.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 82–97

BLOCKING PROBABILITIES FOR AN UNDERLOADED OR OVERLOADED LINK WITH TRUNK RESERVATION*

CHARLES KNESSL† AND JOHN A. MORRISON‡

Abstract. A single link in a circuit-switched network is considered. The link has C circuits, R of which are reserved for the primary traffic. Offered calls arrive in independent Poisson streams with mean rates λ and ν for the primary and secondary traffic, respectively, and corresponding independent and exponentially distributed holding times with means 1 and $1/\kappa$. Both primary and secondary calls require 1 circuit. A primary call is blocked on arrival if all C circuits are busy, whereas a secondary call is blocked if more than $C - R - 1$ circuits are busy. Blocked calls are lost to the link. It is assumed that $R = O(1)$ and $\lambda \gg 1$, $\nu = O(\lambda)$ and $C = O(\lambda)$, and that the link is either underloaded, corresponding to $\lambda + \nu/\kappa < C$, or overloaded, corresponding to $\lambda + \nu/\kappa > C$. Asymptotic approximations to the blocking probabilities B_1 and B_2 of the primary and secondary calls, respectively, are derived. Numerical results are presented to illustrate the accuracy of the approximations.

Key words. queueing, blocking, trunk reservation, asymptotics

AMS subject classifications. 34E10, 60K25, 90B18

DOI. 10.1137/S0036139903426599

1. Introduction.

The concept of trunk reservation is of fundamental importance in circuit-switched communication networks. On any link of the network, which has a fixed number of circuits, some of them may be reserved for the primary traffic which is offered directly to the link. Secondary traffic, which is rerouted because of a busy link on its direct route, is accepted on an alternate link only if there are enough unreserved links available. State-dependent routing on symmetric loss networks with trunk reservation has been investigated by Mitra, Gibbens, and Huang [4], [5] and Mitra and Gibbens [3]. Basic to the investigations is an analysis of a single link. The network is then analyzed by means of fixed point approximations [1], [2] which are based on the assumption that each link acts independently; the traffic streams which are Poisson when offered to the network remain Poisson when offered to the links by virtue of this assumption.

The investigation of a single link with trunk reservation is also of interest when integrated traffic is considered. Thus, some classes of traffic may have less stringent blocking requirements than others, e.g., voice as compared to data. It is crucial for practical applications to investigate the effect of trunk reservation on the blocking probabilities, since this is a prime quality of service (QoS) requirement. In this paper we obtain asymptotic approximations to these probabilities for two classes under appropriate traffic conditions. The approximations provide insight into the effect of trunk reservation, and they are easily evaluated numerically.

*Received by the editors April 18, 2003; accepted for publication (in revised form) March 31, 2005; published electronically October 3, 2005.

http://www.siam.org/journals/siap/66-1/42659.html

†Department of Mathematics, Statistics, and Computer Science (M/C 249), University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607-7045 (knessl@uic.edu). This author's research was partially supported by NSF grants DMS 99-71656 and DMS 02-02815, and NSA grant MDA 904-03-1-0036.

‡Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, NJ 07974-0636 (johnmorrison@lucent.com).

A different aspect has been considered by Hunt and Laws [7] who investigated control policies for a single link in heavy traffic. They allowed for several classes of traffic with different capacity requirements, which earn different rewards for being accepted, rather than blocked and lost. The goal is to maximize the expected reward per unit time. They show that the optimal policy for accepting or rejecting calls offered to the link is of (generalized) trunk reservation from. However, their results do not provide a concrete guideline for setting trunk reservation parameters, but they do suggest an insensitivity to the value of such parameters.

In this paper we consider a single link with C circuits, R of which are reserved for the primary traffic. The offered calls arrive in independent Poisson streams with mean rates λ and ν for the primary and secondary traffic, respectively, and corresponding independent and exponentially distributed holding times with means 1 and $1/\kappa$. Both primary and secondary calls require 1 circuit. A primary call is blocked on arrival if all C circuits are busy, whereas a secondary call is blocked if more than $C-R-1$ circuits are busy. Blocked calls are lost to the link.

When $\kappa = 1$, the single link corresponds to the one investigated by Mitra and Gibbens [3]. Exact expressions for the blocking probabilities B_1 and B_2 of the primary and secondary calls, respectively, may be obtained. Mitra and Gibbens [3] derived asymptotic approximations to B_1 and B_2 under the assumptions $\lambda \gg 1$, $C-\lambda = O(\sqrt{\lambda}\,)$, $R = O(\sqrt{\lambda}\,)$, and $\nu = O(\sqrt{\lambda}\,)$. Since $C-\lambda-\nu/\kappa = O(\sqrt{\lambda}\,)$, the link is critically loaded. Morrison [6] investigated the single link when $\kappa \neq 1$, for which no exact solution is known, by means of singular perturbation techniques. For $\nu = \gamma\sqrt{\lambda}$, he derived the first two terms in the power series expansion in γ of the leading term in the asymptotic approximations to B_1 and B_2. For $R = O(1)$ and $\nu = O(\sqrt{\lambda}\,)$, Morrison [6] obtained explicit expressions for B_1 and B_2.

In this paper we consider $R = O(1)$ and the asymptotic regime $\lambda \gg 1$, $\nu = O(\lambda)$, and $C = O(\lambda)$ for both an underloaded link, corresponding to $\lambda+\nu/\kappa < C$, and for an overloaded link, corresponding to $\lambda+\nu/\kappa > C$. It is assumed that $(\lambda+\nu/\kappa-C)/\lambda$ is not small. The leading term in the asymptotic approximations to B_1 and B_2 is derived. Numerical results are presented to illustrate the accuracy of the approximations.

The paper is organized as follows. In section 2 the problem is formulated, and the asymptotic regime is introduced. An underloaded link is investigated in section 3, and an overloaded one in section 4. The leading term in the lowest order asymptotic approximation to the joint steady-state distribution of the number of primary and secondary calls is investigated for an overloaded link in section 5. Numerical results are presented in section 6. In section 7 we discuss in more detail the ranges of validity of the asymptotic approximations.

2. Formulation. For the model described in the previous section, let $p(n_1, n_2)$ denote the steady-state probability that n_1 primary and n_2 secondary calls are in progress, and let $I(\cdot)$ be the indicator function. Then, by standard arguments,

$$\begin{aligned}
&[\lambda I(n_1+n_2+1 \le C) + \nu I(n_1+n_2+1 \le C-R) + n_1 + \kappa n_2]p(n_1,n_2)\\
&= \lambda I(n_1 \ge 1)p(n_1-1,n_2) + \nu I(n_1+n_2 \le C-R)I(n_2 \ge 1)p(n_1,n_2-1)\\
&\quad + I(n_1+n_2+1 \le C)(n_1+1)p(n_1+1,n_2)\\
&\quad + \kappa I(n_1+n_2+1 \le C)I(n_2+1 \le C-R)(n_2+1)p(n_1,n_2+1),\\
&\qquad n_1 \ge 0, \quad 0 \le n_2 \le C-R, \quad n_1+n_2 \le C. \qquad (2.1)
\end{aligned}$$

The normalization condition is

$$\sum_{n_2=0}^{C-R}\sum_{n_1=0}^{C-n_2} p(n_1,n_2)=1. \tag{2.2}$$

The terms on the right-hand side of (2.1) correspond to different transitions that leave the system with n_1 primary and n_2 secondary customers. The first term corresponds to a primary arrival, the second to a secondary arrival, the third to a primary departure, and the fourth to a secondary departure. The indicator functions show when such a transition is possible; their use allows us to avoid writing down a large number of "boundary equations."

The blocking probabilities B_1 and B_2 of the primary and secondary calls, respectively, are

$$B_1=\sum_{n_1=R}^{C} p(n_1,C-n_1),\quad B_2=\sum_{\ell=0}^{R}\sum_{n_1=\ell}^{C-R+\ell} p(n_1,C-R+\ell-n_1). \tag{2.3}$$

We let

$$n_1=m,\quad n_1+n_2=C-R+\ell,\quad p(n_1,n_2)=P_\ell(m). \tag{2.4}$$

Then, from (2.1)–(2.3), we obtain

$$\begin{aligned}&[\lambda I(\ell\le R-1)+\nu I(\ell\le -1)+m+\kappa(C-R+\ell-m)]P_\ell(m)\\&=\lambda I(m\ge 1)P_{\ell-1}(m-1)+\nu I(\ell\ge 0)I(C-R+\ell-m\ge 1)P_{\ell-1}(m)\\&+I(\ell\le R-1)[(m+1)P_{\ell+1}(m+1)+\kappa I(m\ge \ell+1)(C-R+\ell-m+1)P_{\ell+1}(m)],\end{aligned}$$

$$0\le m\le C,\quad R-C+m\le \ell\le R,\quad R\ge 0, \tag{2.5}$$

$$\sum_{\ell=R-C}^{R}\sum_{m=\max(\ell,0)}^{C-R+\ell} P_\ell(m)=1, \tag{2.6}$$

and

$$B_1=\sum_{m=R}^{C} P_R(m),\quad B_2=\sum_{\ell=0}^{R}\sum_{m=\ell}^{C-R+\ell} P_\ell(m). \tag{2.7}$$

We consider the asymptotic regime in which

$$C-R=\frac{\sigma\lambda}{\kappa},\quad \nu=(\rho-\kappa)\lambda>0,\quad \lambda\gg 1,\quad R=O(1), \tag{2.8}$$

and let

$$m=\zeta\lambda+x\sqrt{\lambda},\quad P_\ell(m)=p_\ell(x),\qquad x=O(1),\quad \ell=O(1), \tag{2.9}$$

where ζ is to be determined, with $0<\zeta<\sigma/\kappa$. Then (2.5) becomes

$$\begin{aligned}&[\lambda I(\ell\le R-1)+(\rho-\kappa)\lambda I(\ell\le -1)+(1-\kappa)(\zeta\lambda+x\sqrt{\lambda}\,)+\sigma\lambda+\kappa\ell]p_\ell(x)\\&=\lambda p_{\ell-1}\left(x-\frac{1}{\sqrt{\lambda}}\right)+(\rho-\kappa)\lambda I(\ell\le 0)p_{\ell-1}(x)\\&\quad+I(\ell\le R-1)\left\{\left(\zeta\lambda+x\sqrt{\lambda}+1\right)p_{\ell+1}\left(x+\frac{1}{\sqrt{\lambda}}\right)\right.\\&\qquad\left.+[(\sigma-\kappa\zeta)\lambda-\kappa x\sqrt{\lambda}+\kappa(\ell+1)]p_{\ell+1}(x)\right\},\quad \ell\le R.\end{aligned} \tag{2.10}$$

We assume that

$$p_\ell(x) = A(\lambda)\left[p_\ell^{(0)}(x) + \frac{1}{\sqrt{\lambda}}p_\ell^{(1)}(x) + O\left(\frac{1}{\lambda}\right)\right]. \tag{2.11}$$

Then, from (2.10), we obtain at the first two orders

$$\begin{aligned}&[I(\ell \le R-1) + (\rho-\kappa)I(\ell \le -1) + \sigma + \zeta - \kappa\zeta]p_\ell^{(0)}(x)\\ &= [1 + (\rho-\kappa)I(\ell \le 0)]p_{\ell-1}^{(0)}(x) + I(\ell \le R-1)(\sigma+\zeta-\kappa\zeta)p_{\ell+1}^{(0)}(x), \quad \ell \le R,\end{aligned} \tag{2.12}$$

and

$$\begin{aligned}&[I(\ell \le R-1) + (\rho-\kappa)I(\ell \le -1) + \sigma + \zeta - \kappa\zeta]p_\ell^{(1)}(x) + (1-\kappa)xp_\ell^{(0)}(x)\\ &= [1 + (\rho-\kappa)I(\ell \le 0)]p_{\ell-1}^{(1)}(x) - \frac{dp_{\ell-1}^{(0)}}{dx}\\ &\quad + I(\ell \le R-1)\left[(\sigma+\zeta-\kappa\zeta)p_{\ell+1}^{(1)}(x) + (1-\kappa)xp_{\ell+1}^{(0)}(x) + \zeta\frac{dp_{\ell+1}^{(0)}}{dx}\right], \quad \ell \le R.\end{aligned} \tag{2.13}$$

It will be shown that $p_\ell^{(0)}(x)$, $0 \le \ell \le R$, is proportional to $\exp(-\beta x^2)$, where $\beta > 0$. Hence, from (2.7), (2.9), and (2.11), the blocking probabilities are asymptotically given by

$$B_1 \sim \sqrt{\lambda}A(\lambda)\int_{-\infty}^{\infty} p_R^{(0)}(x)dx, \qquad B_2 \sim \sqrt{\lambda}A(\lambda)\sum_{\ell=0}^{R}\int_{-\infty}^{\infty} p_\ell^{(0)}(x)dx. \tag{2.14}$$

The constant $A(\lambda)$ will be ultimately determined by the normalization condition (2.6).

3. Underloaded link. We here consider the case $\rho < \sigma$, so that $\lambda + \nu/\kappa < C - R$, and the link is underloaded. In the interior of the admissible region, $p(n_1, n_2)$ is given asymptotically by the solution corresponding to $C = \infty$, since the blocking probabilities will be exponentially small. Hence,

$$p(n_1, n_2) \sim e^{-\lambda}e^{-\nu/\kappa}\frac{\lambda^{n_1}(\nu/\kappa)^{n_2}}{n_1!n_2!}, \qquad 0 \le n_1 + n_2 \ll C - R. \tag{3.1}$$

The precise range of validity of (3.1) is discussed in section 7. If we use Stirling's formula, we find that $p(n_1, n_2)$ attains its maximum for $n_1 \sim \lambda$, $n_2 \sim \nu/\kappa$. Also, for $n_1 \sim \zeta\lambda$ and $n_1 + n_2 = \sigma\lambda/\kappa$, we find that $p(n_1, n_2)$ attains its maximum for $\zeta = \sigma/\rho$, which is the appropriate choice for ζ in (2.9). For

$$n_1 = \frac{\sigma\lambda}{\rho} + x\sqrt{\lambda}, \qquad n_1 + n_2 = \frac{\sigma\lambda}{\kappa} + \ell, \tag{3.2}$$

we obtain from (3.1)

$$p_\ell(x) \sim A(\lambda)\left(\frac{\rho}{\sigma}\right)^\ell \exp\left[-\frac{\rho^2x^2}{2\sigma(\rho-\kappa)}\right], \qquad -\ell \gg 1, \ \frac{-\ell}{\sqrt{\lambda}} \ll 1, \tag{3.3}$$

where

$$A(\lambda) = \frac{e^{(\sigma-\rho)\lambda/\kappa}\left(\frac{\rho}{\sigma}\right)^{\frac{\sigma\lambda}{\kappa}+1}}{2\pi\lambda\sqrt{\frac{\rho}{\kappa}-1}}. \tag{3.4}$$

From (2.12), with $\zeta = \sigma/\rho$, we have

$$\left(1+\frac{\sigma}{\rho}\right) p_\ell^{(0)}(x) = p_{\ell-1}^{(0)}(x) + \frac{\sigma}{\rho} p_{\ell+1}^{(0)}(x), \qquad \ell \le -1. \tag{3.5}$$

Hence, matching with the interior solution (3.3) for $-\ell \gg 1$ and noting that $\rho < \sigma$, we obtain

$$p_\ell^{(0)}(x) = \left(\frac{\rho}{\sigma}\right)^\ell \exp\left[-\frac{\rho^2 x^2}{2\sigma(\rho-\kappa)}\right] + H_0(x), \qquad \ell \le 0. \tag{3.6}$$

Here $H_0(x)$ is to be determined shortly. If $R \ge 1$, then, from (2.12) with $\zeta = \sigma/\rho$, we have

$$\begin{aligned} &\left[I(\ell \le R-1) + \frac{\sigma}{\rho}(1+\rho-\kappa)\right] p_\ell^{(0)}(x) \\ &= p_{\ell-1}^{(0)}(x) + I(\ell \le R-1)\frac{\sigma}{\rho}(1+\rho-\kappa) p_{\ell+1}^{(0)}(x), \qquad 1 \le \ell \le R. \end{aligned} \tag{3.7}$$

Hence,

$$p_\ell^{(0)}(x) = \widetilde{a}^\ell F_0(x), \qquad 0 \le \ell \le R, \tag{3.8}$$

where

$$\widetilde{a} = \frac{\rho}{\sigma(1+\rho-\kappa)} < 1, \tag{3.9}$$

since $\kappa < \rho < \sigma$.

It follows from (3.6) and (3.8), by continuity at $\ell = 0$, that

$$F_0(x) = \exp\left[-\frac{\rho^2 x^2}{2\sigma(\rho-\kappa)}\right] + H_0(x). \tag{3.10}$$

But, from (2.12) with $\zeta = \sigma/\rho$,

$$\begin{aligned} &\left[I(R \ge 1) + \frac{\sigma}{\rho}(1+\rho-\kappa)\right] p_0^{(0)}(x) \\ &= (1+\rho-\kappa) p_{-1}^{(0)}(x) + I(R \ge 1)\frac{\sigma}{\rho}(1+\rho-\kappa) p_1^{(0)}(x). \end{aligned} \tag{3.11}$$

Hence, from (3.6) and (3.8)–(3.11), since $\rho < \sigma$, we obtain $H_0(x) = 0$, and

$$p_\ell^{(0)}(x) = \widetilde{a}^\ell \exp\left[-\frac{\rho^2 x^2}{2\sigma(\rho-\kappa)}\right], \qquad 0 \le \ell \le R. \tag{3.12}$$

From (2.14), (3.4), and (3.12), we obtain the asymptotic approximations to the blocking probabilities,

$$\begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \sim \sqrt{\frac{\kappa}{2\pi\sigma\lambda}} e^{(\sigma-\rho)\lambda/\kappa} \left(\frac{\rho}{\sigma}\right)^{\sigma\lambda/\kappa} \begin{bmatrix} \widetilde{a}^R \\ \frac{(1-\widetilde{a}^{R+1})}{(1-\widetilde{a})} \end{bmatrix}, \tag{3.13}$$

and $\widetilde{a}$ is given by (3.9). We note that B_1 and B_2 become exponentially small as $\lambda \to \infty$.

We now illustrate the matching of this result for an underloaded link with the one for a critically loaded link [6] when $R = O(1)$. In the notation of [6], $C - R = \lambda - \beta\sqrt{\lambda}$ and $\nu = \gamma\sqrt{\lambda}$. We let

$$\omega = \beta + \frac{\gamma}{\kappa}, \quad -\omega \gg 1, \ -\frac{\omega}{\sqrt{\lambda}} \ll 1. \tag{3.14}$$

Then, from (2.8), we have

$$\sigma = \kappa + \frac{\gamma - \kappa\omega}{\sqrt{\lambda}}, \qquad \rho = \kappa + \frac{\gamma}{\sqrt{\lambda}}. \tag{3.15}$$

Hence,

$$\frac{\sigma}{\kappa} \sim 1, \quad 1 - \frac{\rho}{\sigma} \sim -\frac{\omega}{\sqrt{\lambda}}, \quad \frac{\lambda\sigma}{\kappa}\left[1 - \frac{\rho}{\sigma} + \log\left(\frac{\rho}{\sigma}\right)\right] \sim -\frac{\omega^2}{2}. \tag{3.16}$$

Also, from (3.9), $\widetilde{a} \sim 1$ so that $\widetilde{a}^R \sim 1$ and $(1 - \widetilde{a}^{R+1})/(1 - \widetilde{a}) \sim R + 1$, and from (3.13) we obtain

$$\begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \sim \frac{e^{-\omega^2/2}}{\sqrt{2\pi\lambda}} \begin{bmatrix} 1 \\ R+1 \end{bmatrix}. \tag{3.17}$$

But, from (7.16)–(7.18) in [6], with $t = 1$,

$$\begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \sim \frac{1}{W_0(\omega)\sqrt{\lambda}} \begin{bmatrix} 1 \\ R+1 \end{bmatrix}, \tag{3.18}$$

where

$$W_0(\omega) = e^{\omega^2/2} \int_\omega^\infty e^{-u^2/2} du \sim \sqrt{2\pi} e^{\omega^2/2}, \qquad -\omega \gg 1. \tag{3.19}$$

Hence the results match asymptotically.

To summarize, we have obtained the blocking probabilities for the underloaded case; cf. (3.13) with (3.9) and (2.8). The joint probability distribution is given, asymptotically, by (3.3) with (3.2) and (3.4) for $\ell < 0$, and by (3.12) with (2.11) for $0 \le \ell \le R$.

4. Overloaded link. We now consider the case

$$\rho > \max(\kappa, \sigma) > 0, \tag{4.1}$$

so that $\lambda + \nu/\kappa > \sigma\lambda/\kappa = C - R$, and the link is overloaded. We note that the link is underloaded or overloaded, by the primary traffic alone, according as $\sigma > \kappa$ or $\sigma < \kappa$, respectively. If $R \ge 1$, then, from (2.12), we have

$$\begin{aligned} &[I(\ell \le R-1) + \sigma + \zeta - \kappa\zeta] p_\ell^{(0)}(x) \\ &= p_{\ell-1}^{(0)}(x) + I(\ell \le R-1)(\sigma + \zeta - \kappa\zeta) p_{\ell+1}^{(0)}(x), \qquad 1 \le \ell \le R. \end{aligned} \tag{4.2}$$

We let

$$a = \frac{1}{(\sigma + \zeta - \kappa\zeta)} > 0, \tag{4.3}$$

since $0 < \zeta < \sigma/\kappa$. It follows from (4.2) that

$$p_\ell^{(0)}(x) = a^\ell F_0(x), \qquad 0 \le \ell \le R, \tag{4.4}$$

where $F_0(x)$ is to be determined.

From (2.12), we have

$$\begin{aligned} &[1+\rho+\sigma-\kappa+(1-\kappa)\zeta]p_\ell^{(0)}(x) \\ &= (1+\rho-\kappa)p_{\ell-1}^{(0)}(x) + (\sigma+\zeta-\kappa\zeta)p_{\ell+1}^{(0)}(x), \qquad \ell \le -1. \end{aligned} \tag{4.5}$$

We let

$$b = \frac{(1+\rho-\kappa)}{(\sigma+\zeta-\kappa\zeta)} > a, \tag{4.6}$$

from (4.3), since $\rho > \kappa$. We will show that $b > 1$. Then, the solution of (4.5) which decreases geometrically as $-\ell$ increases is

$$p_\ell^{(0)}(x) = b^\ell F_0(x), \qquad \ell \le 0. \tag{4.7}$$

If we sum on ℓ in (2.10), neglect exponentially small terms, and expand $p_{\ell-1}(x - 1/\sqrt{\lambda}\,)$ and $p_{\ell+1}(x + 1/\sqrt{\lambda}\,)$ in powers of $1/\sqrt{\lambda}$, we obtain

$$\begin{aligned} &\sum_{\ell=-\infty}^{R}\left[\sqrt{\lambda}\zeta\frac{dp_\ell}{dx} + \frac{\zeta}{2}\frac{d^2p_\ell}{dx^2} + x\frac{dp_\ell}{dx} + p_\ell(x)\right] \\ &= \sum_{\ell=-\infty}^{R-1}\left(\sqrt{\lambda}\frac{dp_\ell}{dx} - \frac{1}{2}\frac{d^2p_\ell}{dx^2}\right) + O\left(\frac{1}{\sqrt{\lambda}}\right). \end{aligned} \tag{4.8}$$

Hence, from (2.11), we have

$$\zeta\sum_{\ell=-\infty}^{R}\frac{dp_\ell^{(0)}}{dx} = \sum_{\ell=-\infty}^{R-1}\frac{dp_\ell^{(0)}}{dx}. \tag{4.9}$$

It follows from (4.4) and (4.7) that

$$h(\zeta) \equiv \zeta\left[\frac{1}{(b-1)} + \frac{(1-a^{R+1})}{(1-a)}\right] - \left[\frac{1}{(b-1)} + \frac{(1-a^R)}{(1-a)}\right] = 0, \tag{4.10}$$

where a and b are functions of ζ, as given by (4.3) and (4.6). The limiting values as $a \to 1$ are to be taken in (4.10) if $a = 1$.

In Appendix A we show that there is a solution of (4.10) with $b > 1$ and $0 < \zeta\sigma/\kappa$. In fact, $0 < \zeta < \min(1, \sigma/\kappa)$ if $\sigma < 1 + \rho - \kappa$, and $0 \le \zeta^* < \zeta < \min(1, \sigma/\kappa)$ if $\sigma \ge 1 + \rho - \kappa$, which implies $\kappa > 1$, where

$$\zeta^* = \frac{(\kappa+\sigma-1-\rho)}{(\kappa-1)}. \tag{4.11}$$

Moreover, in the next section, we show that $F_0(x)$ is proportional to $\exp(-\gamma x^2/2)$, where $\gamma > 0$, and without loss of generality we may take $\int_{-\infty}^{\infty} F_0(x)dx = 1$. Then, from (2.6), (2.9), (2.11), (4.4), and (4.7), we obtain

$$1 \sim \sqrt{\lambda}A(\lambda)\sum_{\ell=-\infty}^{R} p_\ell^{(0)}(x)dx = \sqrt{\lambda}A(\lambda)\left[\frac{1}{(b-1)} + \frac{(1-a^{R+1})}{(1-a)}\right]. \tag{4.12}$$

Hence, from (2.14), the blocking probabilities are asymptotically given by

$$\begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \sim \left[\frac{1}{(b-1)} + \frac{(1-a^{R+1})}{(1-a)}\right]^{-1} \begin{bmatrix} a^R \\ \frac{(1-a^{R+1})}{(1-a)} \end{bmatrix}, \tag{4.13}$$

where a and b are given by (4.3) and (4.6).

We now illustrate the matching of this result for an overloaded link with the one for a critically loaded link [6] when $R = O(1)$. Now ω, as defined in (3.14), satisfies $\omega \gg 1$ and $\omega/\sqrt{\lambda} \ll 1$. With $\zeta \sim 1 - \delta/\sqrt{\lambda}$, it follows from (4.3) and (4.6) that

$$a - 1 \sim \frac{1}{\sqrt{\lambda}} \left[\kappa\omega - \gamma + (1-\kappa)\delta\right], \qquad b - 1 \sim \frac{1}{\sqrt{\lambda}} \left[\kappa\omega + (1-\kappa)\delta\right]. \tag{4.14}$$

After some straightforward algebra, it is found from (4.10) that $\delta = \omega$. Hence $a \sim 1$, $b - 1 \sim \omega/\sqrt{\lambda}$, and (4.13) implies that

$$\begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \sim \frac{\omega}{\sqrt{\lambda}} \begin{bmatrix} 1 \\ R+1 \end{bmatrix}. \tag{4.15}$$

However,

$$W_0(\omega) = \int_0^\infty e^{-v^2/2} e^{-\omega v} dv = \frac{1}{\omega} \int_0^\infty \exp\left[-\frac{\eta^2}{2\omega^2}\right] e^{-\eta} d\eta \sim \frac{1}{\omega}, \qquad \omega \ll 1, \tag{4.16}$$

and (4.15) is consistent with (3.18).

To summarize, we have obtained the blocking probabilities for the overloaded case; cf. (4.13) with (4.3), (4.6), and (4.10). The joint probability distribution is given, asymptotically, by (4.4) and (4.7), with (2.4), (2.9), (2.11), and (4.12), with $F_0(x)$ being the Gaussian defined below (4.11).

5. Lowest order function. We here derive an equation for $F_0(x)$, which appears in (4.4) and (4.7). We emphasize that it is crucial to show that $F_0(x)$ is exponentially small for $x \to \pm\infty$, so that (2.14) and (4.12) hold. From (2.13), (4.6), and (4.7), we obtain

$$\begin{aligned} &(1+\rho-\kappa)\left[\frac{1}{b} p_{\ell+1}^{(1)}(x) - \left(1 + \frac{1}{b}\right) p_\ell^{(1)}(x) + p_{\ell-1}^{(1)}(x)\right] \\ &= b^{\ell-1}[b(1-b)(1-\kappa)xF_0(x) + (1-\zeta b^2)F_0'(x)], \qquad \ell \le -1, \end{aligned} \tag{5.1}$$

where the prime denotes derivative. The solution, which becomes small for large values of $-\ell$, is

$$p_\ell^{(1)}(x) = b^\ell \left\{F_1(x) - \frac{\ell}{(1+\rho-\kappa)} \left[b(1-\kappa)xF_0(x) + \frac{(\zeta b^2 - 1)}{(b-1)} F_0'(x)\right]\right\}, \quad \ell \le 0, \tag{5.2}$$

where $F_1(x)$ is not determined.

If $R \ge 1$, then, from (2.13), (4.3), and (4.4), we have

$$\begin{aligned} &I(\ell \le R-1)\frac{1}{a} p_{\ell+1}^{(1)}(x) - \left[I(\ell \le R-1) + \frac{1}{a}\right] p_\ell^{(1)}(x) + p_{\ell-1}^{(1)}(x) \\ &= a^{\ell-1}[a(1-\kappa)xF_0(x) + F_0'(x)] \\ &\quad - I(\ell \le R-1)a^{\ell+1}[(1-\kappa)xF_0(x) + \zeta F_0'(x)], \qquad 1 \le \ell \le R. \end{aligned} \tag{5.3}$$

Since $p_0^{(1)}(x) = F_1(x)$, it is found that, for $a \neq 1$,

$$p_\ell^{(1)}(x) = a^\ell \left\{ F_1(x) - \ell \left[a(1-\kappa)xF_0(x) + \frac{(1-\zeta a^2)}{(1-a)} F_0'(x) \right] \right\}$$

$$+ a^{R+1} \frac{(1-a^\ell)}{(1-a)^2}(1-a\zeta)F_0'(x), \qquad 0 \leq \ell \leq R. \tag{5.4}$$

The expression in (5.4) remains finite as $a \to 1$, so that the result for $a = 1$ may be obtained in this limit. Consequently, we subsequently assume that $a \neq 1$.

If we let $\ell = 0$ in (2.13), then

$$[I(R \geq 1) + \sigma + \zeta - \kappa\zeta]p_0^{(1)}(x) + (1-\kappa)xp_0^{(0)}(x)$$
$$= (1 + \rho - \kappa)p_{-1}^{(1)}(x) - \frac{dp_{-1}^{(0)}}{dx}$$
$$+ I(R \geq 1) \left[(\sigma + \zeta - \kappa\zeta)p_1^{(1)}(x) + (1-\kappa)xp_1^{(0)}(x) + \zeta \frac{dp_1^{(0)}}{dx} \right]. \tag{5.5}$$

It may be verified from (4.3), (4.4), (4.6), (4.7), (5.2), and (5.4) that this equation is satisfied in view of (4.10), which follows from (4.9).

Next, from (2.11), (4.8), and (4.9), we obtain

$$\sum_{\ell=-\infty}^{R} \left[\zeta \frac{dp_\ell^{(1)}}{dx} + \frac{\zeta}{2} \frac{d^2 p_\ell^{(0)}}{dx^2} + x \frac{dp_\ell^{(0)}}{dx} + p_\ell^{(0)}(x) \right]$$
$$= \sum_{\ell=-\infty}^{R-1} \left[\frac{dp_\ell^{(1)}}{dx} - \frac{1}{2} \frac{d^2 p_\ell^{(0)}}{dx^2} \right]. \tag{5.6}$$

We define

$$\Omega \equiv \frac{(\zeta b^2 - 1)(1-\zeta)}{(1+\rho-\kappa)(b-1)} \sum_{\ell=-\infty}^{-1} \ell b^\ell + \frac{(1-\zeta a^2)}{(1-a)} \left(\sum_{\ell=0}^{R-1} \ell a^\ell - \zeta \sum_{\ell=0}^{R} \ell a^\ell \right)$$
$$+ \frac{(\zeta a - 1)}{(1-a)^2} a^{R+1} \left[\zeta \sum_{\ell=0}^{R} (a^\ell - 1) - \sum_{\ell=0}^{R-1} (a^\ell - 1) \right] + \zeta \left(\sum_{\ell=-\infty}^{-1} b^\ell + \sum_{\ell=0}^{R} a^\ell \right) \tag{5.7}$$

and

$$\Gamma \equiv \frac{b(1-\kappa)(1-\zeta)}{(1+\rho-\kappa)} \sum_{\ell=-\infty}^{-1} \ell b^\ell + a(1-\kappa) \left(\sum_{\ell=0}^{R-1} \ell a^\ell - \zeta \sum_{\ell=0}^{R} \ell a^\ell \right)$$
$$+ \sum_{\ell=-\infty}^{-1} b^\ell + \sum_{\ell=0}^{R} a^\ell. \tag{5.8}$$

Then from (4.4), (4.7), (5.2), (5.4), and (5.6) we obtain

$$\Omega F_0''(x) + \Gamma[xF_0'(x) + F_0(x)] = 0. \tag{5.9}$$

In Appendices B and C we use (4.10) to express ζ, and hence Ω and Γ, in terms of a and b. Then, using the facts that $b > a > 0$ and $b > 1$, it is shown that $\Omega > 0$ and $\Gamma > 0$. Hence $F_0(x)$ is proportional to $\exp(-\gamma x^2/2)$, where $\gamma = \Gamma/\Omega > 0$. This establishes the validity of the asymptotic approximation (4.13) to the blocking probabilities.

TABLE 1

C	B_1(exact)	B_2(exact)	B_1(asymptotic)	B_2(asymptotic)
50	6.57×10^{-4}	9.09×10^{-2}	7.64×10^{-4}	7.87×10^{-2}
60	4.54×10^{-4}	6.44×10^{-2}	5.23×10^{-4}	5.82×10^{-2}
70	3.22×10^{-4}	4.67×10^{-2}	3.69×10^{-4}	4.34×10^{-2}
80	2.34×10^{-4}	3.43×10^{-2}	2.66×10^{-4}	3.26×10^{-2}
90	1.72×10^{-4}	2.55×10^{-2}	1.95×10^{-4}	2.46×10^{-2}
100	1.28×10^{-4}	1.92×10^{-2}	1.44×10^{-4}	1.87×10^{-2}
250	2.42×10^{-6}	3.81×10^{-4}	2.57×10^{-6}	3.82×10^{-4}
500	5.26×10^{-9}	8.39×10^{-7}	5.42×10^{-9}	8.41×10^{-7}
750	1.32×10^{-11}	2.11×10^{-9}	1.34×10^{-11}	2.12×10^{-9}
1000	3.50×10^{-14}	5.64×10^{-12}	3.56×10^{-14}	5.64×10^{-12}

6. Numerical results. We present numerical results for the blocking probabilities B_1 and B_2 of the primary and secondary calls, respectively. Exact results may be obtained from (2.3) by numerically solving the recursion (2.1) for the joint steady-state probability $p(n_1, n_2)$ of the numbers of primary and secondary calls, subject to the normalization condition (2.2). However, this is feasible for values of C, the number of circuits in the link, up to about 50. If the mean holding times of the primary and secondary calls are equal, so that $\kappa = 1$, then the distribution of the total number of calls, and hence B_1 and B_2, may be calculated from a well-known single recursion for values of C up to at least 1000.

Asymptotic approximations to B_1 and B_2 were derived under the assumptions in (2.8), where λ and ν are the mean arrival rates of the primary and secondary calls, respectively, with corresponding mean holding times 1 and $1/\kappa$, and R circuits are reserved for the prime calls. In the case of an underloaded link, corresponding to $\lambda + \nu/\kappa = \rho\lambda/\kappa < \sigma\lambda/\kappa = C - R$, the approximations to B_1 and B_2 are given by (3.13), where $\tilde{a}$ is given by (3.9). In the case of an underloaded link, corresponding to $\lambda + \nu/\kappa > C - R$, so that $\rho > \sigma$, the approximations to B_1 and B_2 are given by (4.13), where a and b are given by (4.3) and (4.6), and ζ is a solution of (4.10) with $b > 1$ and $0 < \zeta < \min(1, \sigma/\kappa)$. We have assumed that $\lambda \gg 1$ and $R = O(1)$. It follows from (4.4) and (5.4) that for the asymptotic expansion (2.11) to be valid we need $R/\sqrt{\lambda} \ll 1$. Consideration of the first order correction term in (2.11) for an underloaded link leads to the same restriction.

In Table 1 we compare the asymptotic approximations to B_1 and B_2 and the exact values for an underloaded link, with $R = 5,\ \kappa = 1$, and $\lambda = 0.4C = \nu$, for different values of C. The analogous comparisons are made in Table 2 for an overloaded link, with $R = 5, \kappa = 1$, and $\lambda = 0.6C = \nu$. It is seen that the approximations improve as C increases. In Tables 3 and 4 the comparisons are made for $R = 2$ and $\kappa = 2$ and $1/2$, for an underloaded link with $\lambda = 0.2C = \nu/\kappa$, and for an overloaded link with $\lambda = 0.8C = \nu/\kappa$. In Tables 5 and 6 the comparisons are made for $R = 2,\ \kappa = 1$, and $\lambda/\nu = 1, 1/3$, and 3, for an underloaded link with $\lambda + \nu = 0.8C$, and for an overloaded link with $\lambda + \nu = 1.6C$. Again, the approximations improve as C increases.

7. Discussion. Finally, we discuss the range of validity of our asymptotic approximations to $p(n_1, n_2)$ and also indicate how to obtain results of a more "global" nature. We view the domain of the difference equation (2.1) as consisting of the union of the lattice triangle $T = \{(n_1, n_2) : n_1 \geq 0, n_2 \geq 0, n_1 + n_2 \leq C - R\}$ and the oblique strip $S = \{(n_1, n_2) : 0 \leq n_2 \leq C - R, C - R - n_2 \leq n_1 \leq C - n_2\}$. We refer to $\mathcal{I} = T \cap S = \{(n_1, n_2) : n_1 + n_2 = C - R, n_1 \geq 0, n_2 \geq 0\}$ as an "interface."

TABLE 2

C	B_1(exact)	B_2(exact)	B_1(asymptotic)	B_2(asymptotic)
50	1.88×10^{-2}	0.509	2.30×10^{-2}	0.477
60	1.75×10^{-2}	0.484	2.04×10^{-2}	0.452
70	1.66×10^{-2}	0.465	1.86×10^{-2}	0.434
80	1.59×10^{-2}	0.451	1.74×10^{-2}	0.420
90	1.54×10^{-2}	0.439	1.65×10^{-2}	0.409
100	1.49×10^{-2}	0.429	1.58×10^{-2}	0.401
250	1.24×10^{-2}	0.371	1.25×10^{-2}	0.354
500	1.15×10^{-2}	0.348	1.15×10^{-2}	0.339
750	1.12×10^{-2}	0.340	1.11×10^{-2}	0.333
1000	1.10×10^{-2}	0.336	1.10×10^{-2}	0.331

TABLE 3

C	κ	B_1(exact)	B_2(exact)	B_1(asymptotic)	B_2(asymptotic)
10	2	7.54×10^{-4}	3.72×10^{-2}	8.36×10^{-4}	3.59×10^{-2}
20	2	1.97×10^{-5}	1.12×10^{-3}	2.08×10^{-5}	1.11×10^{-3}
30	2	6.54×10^{-7}	3.89×10^{-5}	6.80×10^{-7}	3.88×10^{-5}
40	2	2.35×10^{-8}	1.43×10^{-6}	2.42×10^{-8}	1.43×10^{-6}
50	2	8.82×10^{-10}	5.43×10^{-8}	9.03×10^{-10}	5.42×10^{-8}
10	1/2	2.04×10^{-3}	3.92×10^{-2}	3.34×10^{-3}	4.35×10^{-2}
20	1/2	6.58×10^{-5}	1.25×10^{-3}	8.33×10^{-5}	1.31×10^{-3}
30	1/2	2.34×10^{-6}	4.42×10^{-5}	2.72×10^{-6}	4.56×10^{-5}
40	1/2	8.67×10^{-8}	1.64×10^{-6}	9.70×10^{-8}	1.67×10^{-6}
50	1/2	3.30×10^{-9}	6.24×10^{-8}	3.61×10^{-9}	6.34×10^{-8}

TABLE 4

C	κ	B_1(exact)	B_2(exact)	B_1(asymptotic)	B_2(asymptotic)
10	2	0.181	0.782	0.213	0.787
20	2	0.153	0.719	0.162	0.713
30	2	0.143	0.693	0.147	0.686
40	2	0.138	0.679	0.140	0.673
50	2	0.134	0.670	0.136	0.664
10	1/2	0.230	0.737	0.278	0.722
20	1/2	0.210	0.668	0.227	0.648
30	1/2	0.203	0.640	0.212	0.621
40	1/2	0.198	0.624	0.204	0.608
50	1/2	0.196	0.614	0.200	0.600

The boundary segment $\mathcal{B} = \{(n_1, n_2) : n_1 + n_2 = C, 0 \le n_2 \le C - R\}$ also plays an important role in the analysis.

If we view the problem (2.1) on a coarse spatial scale with $(\xi, \eta) = C^{-1}(n_1, n_2)$, then the differences become small in the new variables, and the domain is approximately

$$\{(\xi, \eta) : \xi \ge 0, \eta \ge 0, \xi + \eta \le 1\}.$$

For $C \to \infty$ with $R = O(1)$ the interface $\mathcal{I}$ and boundary $\mathcal{B}$ are close to (within $O(C^{-1})$ of) one another. The scaling (2.9) corresponds to a small neighborhood of the point $(\xi, \eta) = (\zeta_1, 1 - \zeta_1)$, where $\zeta_1 = \zeta\lambda/C = O(1)$. This is the region where most of the probability mass concentrates, and analysis of this range is sufficient for obtaining asymptotically the blocking probabilities in (2.14).

TABLE 5

C	λ	ν	B_1(exact)	B_2(exact)	B_1(asymptotic)	B_2(asymptotic)
10	4	4	3.65×10^{-2}	3.33×10^{-1}	3.53×10^{-2}	2.47×10^{-1}
20	8	8	1.77×10^{-2}	1.67×10^{-1}	1.65×10^{-2}	1.37×10^{-1}
30	12	12	1.06×10^{-2}	1.01×10^{-1}	1.01×10^{-2}	8.86×10^{-2}
40	16	16	6.97×10^{-3}	6.68×10^{-2}	6.75×10^{-3}	6.09×10^{-2}
50	20	20	4.80×10^{-3}	4.62×10^{-2}	4.72×10^{-3}	4.32×10^{-2}
10	2	6	9.85×10^{-3}	2.81×10^{-1}	8.82×10^{-3}	1.85×10^{-1}
20	4	12	4.58×10^{-3}	1.36×10^{-1}	4.12×10^{-3}	1.06×10^{-1}
30	6	18	2.71×10^{-3}	8.18×10^{-2}	2.52×10^{-3}	6.93×10^{-2}
40	8	24	1.77×10^{-3}	5.37×10^{-2}	1.69×10^{-3}	4.78×10^{-2}
50	10	30	1.21×10^{-3}	3.69×10^{-2}	1.18×10^{-3}	3.40×10^{-2}
10	6	2	7.53×10^{-2}	3.89×10^{-1}	7.93×10^{-2}	3.26×10^{-1}
20	12	4	3.81×10^{-2}	2.02×10^{-1}	3.71×10^{-2}	1.76×10^{-1}
30	18	6	2.33×10^{-2}	1.25×10^{-1}	2.27×10^{-2}	1.13×10^{-1}
40	24	8	1.54×10^{-2}	8.28×10^{-2}	1.52×10^{-2}	7.73×10^{-2}
50	30	10	1.07×10^{-2}	5.75×10^{-2}	1.06×10^{-2}	5.47×10^{-2}

TABLE 6

C	λ	ν	B_1(exact)	B_2(exact)	B_1(asymptotic)	B_2(asymptotic)
10	8	8	0.207	0.757	0.250	0.750
20	16	16	0.185	0.690	0.199	0.676
30	24	24	0.176	0.663	0.184	0.649
40	32	32	0.172	0.648	0.177	0.636
50	40	40	0.169	0.638	0.172	0.628
10	4	12	0.072	0.660	0.091	0.636
20	8	24	0.062	0.584	0.067	0.561
30	12	36	0.058	0.555	0.061	0.535
40	16	48	0.056	0.539	0.058	0.522
50	20	60	0.055	0.529	0.056	0.515
10	12	4	0.336	0.825	0.391	0.826
20	24	8	0.309	0.770	0.329	0.762
30	36	12	0.298	0.746	0.310	0.738
40	48	16	0.292	0.733	0.300	0.725
50	60	20	0.288	0.725	0.294	0.718

To obtain results for $p(n_1, n_2)$ that have a wider range of validity, we can consider the scale $\ell = C - R - n_1 - n_2 = O(1)$ with $\xi \in (0, 1)$. For the overloaded case we can use a WKB-type expansion of the form

$$p(n_1, n_2) = \frac{1}{\sqrt{C}} e^{C\phi(\xi;R)} \left[A_\ell(\xi; R) + \frac{1}{C} A_\ell^{(1)}(\xi; R) + O(C^{-2}) \right]. \tag{7.1}$$

Using (7.1) in (a scaled form of) (2.1) and (2.5), the derivative of the function ϕ can be characterized as a root of a transcendental equation. More precisely, $e^{\phi'}$ can be obtained as an algebraic function of ξ. By expanding (7.1) about $\xi = \zeta_1$, we find that $\phi(\zeta_1; R) = \phi'(\zeta_1; R) = 0$, and near $\xi = \zeta_1$, (7.1) reduces to the form we obtained in section 4. The exponential factor in (7.1) becomes proportional to the Gaussian factor $F_0(x)$, and $A_\ell(\zeta_1; R)$ yields the geometric factors a^ℓ in (4.4) and b^ℓ in (4.7). However, for some parameter ranges within the overloaded case, the function $A_\ell(\xi)$ may, for certain $\xi \in (0, 1)$, become negative and/or not decay for $\ell \to -\infty$. Thus (7.1) will be valid for ξ only in some subset of the unit interval $[0, 1]$. However, it can be shown that ζ_1 is always contained in this subinterval.

We can obtain results even more global than (7.1) by considering the problem on the (ξ, η) scale. Then employing a geometrical optics expansion of the form

$$p(n_1, n_2) = \frac{1}{\sqrt{C}} e^{C\Phi(\xi,\eta)} \left[K(\xi, \eta) + \frac{1}{C} K^{(1)}(\xi, \eta) + O(C^{-2}) \right] \tag{7.2}$$

will yield PDEs satisfied by Φ and K. Solving the PDE for Φ subject to the condition $\Phi(\xi, 1-\xi) = \phi(\xi)$ will determine the solution in either the entire triangular domain or some subset thereof. In the latter case the expansion (7.2) would need to be modified, and different boundary conditions used to determine $\Phi(\xi, \eta)$ in the complementary subset(s). In general, obtaining a complete set of results on the (ξ, η) scale is likely to lead to consideration of several subregions of the basic triangle.

For the underloaded case ($\lambda + \nu/\kappa < C$) we can first consider the problem on the (ξ, η) scale. Employing an expansion of the type (7.2), with the initial factor $1/\sqrt{C}$ replaced by $1/C$, we again find that Φ satisfies a nonlinear PDE. Solving this by the method of characteristics (ray method) and using the characteristic curves that start from the equilibrium point $(\xi, \eta) = C^{-1}(\lambda, \nu/\kappa)$, we can regain the asymptotic expansion of $\log[p(n_1, n_2)]$ in (3.1). Note that $p(n_1, n_2)$ is maximal near this equilibrium point. The range of validity of this expansion is determined by the portion of the triangle $\xi + \eta < 1$ that is filled by the rays. When $\kappa = 1$, this can be shown to be the entire triangle, but in general for $\kappa \neq 1$ the rays will not fill the entire triangle, and thus have a "shadow region." In the shadow the product form expression in (3.1) is not the appropriate approximation to $p(n_1, n_2)$, and an entirely separate analysis is needed. However, we can show that the point $(\xi, \eta) = (\lambda\kappa, \nu)/(\nu + \lambda\kappa)$ always lies in the region illuminated by the rays from the equilibrium, and our analysis used (3.1) only near this point (cf. (3.2)).

This brief discussion shows that obtaining more global results for $p(n_1, n_2)$, with the scaling (2.8), is nontrivial and would require considerably more analysis. We are also presently investigating the case of critical loading, where $\lambda + \nu/\kappa = C + O(\sqrt{C})$. This analysis will likely lead to a two-dimensional diffusion approximation. Another interesting asymptotic limit is to have $R \to \infty$ with $R/C = r = O(1)$. Then even on the coarse (ξ, η) scale, both the triangle T (with $\xi + \eta < 1 - r$) and the oblique strip S (with $1 - r < \xi + \eta < 1, 0 < \eta < 1 - r$) have nonvanishing volume. The clearer separation between the interface $\mathcal{I}$ and boundary $\mathcal{B}$ may actually simplify the asymptotic analysis, over that with the present assumption $R = O(1)$.

Appendix A. We here establish that there is a solution of (4.10) with $b > 1$ and $0 < \zeta < \sigma/\kappa$, where a and b are functions of ζ, as given by (4.3) and (4.6), and (4.1) holds. We let

$$f(\zeta) = \sigma + \zeta - \kappa\zeta > 0, \qquad 0 \leq \zeta \leq \frac{\sigma}{\kappa}. \tag{A.1}$$

From (4.3) and (4.6), we have

$$a(\zeta) = \frac{1}{f(\zeta)}, \qquad b(\zeta) = \frac{(1 + \rho - \kappa)}{f(\zeta)}. \tag{A.2}$$

If $\sigma/\kappa > 1$, then $b(1) > 1$ since $\rho > \sigma$, and $a(1) > 0$, so that $h(1) > 0$. If $\sigma/\kappa \leq 1$, then $b(\sigma/\kappa) = \kappa(1 + \rho - \kappa)/\sigma > 1$ since $\rho > \kappa$. Also, $h(\sigma/\kappa) = (\rho - \kappa)/(b - 1) > 0$. Hence,

$$h\left(\min\left(1, \frac{\sigma}{\kappa}\right)\right) > 0. \tag{A.3}$$

There are two cases to consider.

Case 1. $\sigma < 1 + \rho - \kappa$. Then $b(0) = (1 + \rho - \kappa)/\sigma > 1$, and $b(\zeta) > 1$ for $0 \le \zeta \le \min(1, \sigma/\kappa)$. However, $h(0) < 0$. Hence, from (A.3), $0 < \zeta < \min(1, \sigma/\kappa)$.

Case 2. $\sigma \ge 1 + \rho - \kappa$. Then $\kappa > 1$, since $\rho > \sigma$, and $b(\zeta^*) = 1$, where $\zeta^* < 1$ is given by (4.11). However,

$$\frac{\sigma}{\kappa} - \zeta^* = \frac{[\kappa(\rho - \kappa) + (\kappa - \sigma)]}{\kappa(\kappa - 1)}. \tag{A.4}$$

Hence, if $\sigma \le \kappa$, then $\zeta^* < \sigma/\kappa$, and so $0 \le \zeta^* < \min(1, \sigma/\kappa)$. If follows that $b(\zeta) > 1$ for $\zeta^* < \zeta \le \min(1, \sigma/\kappa)$. Moreover, $b(\zeta) \to 1+$ as $\zeta \to \zeta^*+$, so that $h(\zeta) \to -\infty$ as $\zeta \to \zeta^*+$. Hence, from (A.3), $\zeta^* < \zeta < \min(1, \sigma/\kappa)$.

Appendix B. We here show that $\Omega > 0$, where Ω is given by (5.7). Now,

$$\sum_{\ell=-\infty}^{-1} b^\ell = \frac{1}{(b-1)}, \qquad \sum_{\ell=-\infty}^{-1} \ell b^\ell = -\frac{b}{(b-1)^2}, \tag{B.1}$$

and, for $a \ne 1$,

$$\sum_{\ell=0}^{R-1} a^\ell = \frac{(1-a^R)}{(1-a)}, \qquad \sum_{\ell=0}^{R-1} \ell a^\ell = a\left[\frac{(1-a^R)}{(1-a)^2} - \frac{Ra^{R-1}}{(1-a)}\right]. \tag{B.2}$$

Also, from (4.3) and (4.6),

$$b = (1 + \rho - \kappa)a. \tag{B.3}$$

If we use (4.10) to express ζ in terms of a and b, then after some straightforward algebra, it is found that

$$\left[\frac{1}{(b-1)} + \frac{(1-a^{R+1})}{(1-a)}\right]^2 \Omega = c_0(a) + \frac{c_1(a)}{(b-1)} + \frac{c_2(a)}{(b-1)^2} + \frac{c_3(a)}{(b-1)^3}, \tag{B.4}$$

where

$$\begin{aligned} c_0(a) = {} & \frac{a^R}{(1-a)^2}(1 + a - a^{R+1})\left[\frac{a(1-a^R)}{(1-a)} - R\right] \\ & + \frac{a^{R+1}}{(1-a)^2}\left[\frac{(1-a^R)}{(1-a)} - Ra^R\right] + \frac{(1-a^R)(1-a^{R+1})^2}{(1-a)^3}, \end{aligned} \tag{B.5}$$

$$\begin{aligned} c_1(a) = {} & -a^{R+1}\frac{(1-a^R)}{(1-a)} + a^{R+1}(1+a)\frac{(1-a^R)}{(1-a)^2} - Ra^R\frac{(2+2a-a^{R+1})}{(1-a)} \\ & + \frac{a^{R+1}}{(1-a)}\left[\frac{2(1-a^R)}{(1-a)} - Ra^R\right] + \frac{(1-a^{R+1})}{(1-a)^2}(3 - 2a^R - a^{R+1}), \end{aligned} \tag{B.6}$$

$$\begin{aligned} c_2(a) = {} & -a^{R+1}\frac{(3-a-2a^R)}{(1-a)} - R(1+a)a^R \\ & + a^{R+1}\frac{(1-a^R)}{(1-a)} + \frac{(3-a^R-2a^{R+1})}{(1-a)}, \end{aligned} \tag{B.7}$$

and

$$c_3(a) = 1 + a^{R+1}(a^R - 2) = (1 - a^{R+1})^2 + (1 - a)a^{2R+1}. \tag{B.8}$$

After simplification in (B.5), we obtain

$$c_0(a) = (1 + a^{R+2})\frac{(1 - a^R)}{(1 - a)^3} - \frac{R(1 + a)a^R}{(1 - a)^2}. \tag{B.9}$$

Then, from (B.6) and (B.7), it is found that

$$c_1(a) = 2(1 - a)c_0(a) + \frac{(1 - a^{R+1})^2}{(1 - a)^2} \tag{B.10}$$

and

$$c_2(a) = (1 - a)^2 c_0(a) + a^{2R+1} + \frac{2(1 - a^{R+1})^2}{(1 - a)}. \tag{B.11}$$

Hence, from (B.4), (B.8), (B.10), and (B.11), we have

$$\begin{aligned}&\left[\frac{1}{(b - 1)} + \frac{(1 - a^{R+1})}{(1 - a)}\right]^2 \Omega \\ &= \frac{(b - a)}{(b - 1)^3}\left[(b - a)(b - 1)c_0(a) + (b - a)\frac{(1 - a^{R+1})^2}{(1 - a)^2} + a^{2R+1}\right].\end{aligned} \tag{B.12}$$

However,

$$\sum_{\ell=0}^{R} \ell a^{R-\ell} = \sum_{\ell=0}^{R} (R - \ell)a^\ell = \frac{R}{(1 - a)} - \frac{a(1 - a^R)}{(1 - a)^2}, \tag{B.13}$$

from (B.2). Hence, from (B.9),

$$c_0(a) = \frac{1}{(1 - a)}\sum_{\ell=0}^{R} \ell(a^{\ell-1} - a^{2R-\ell+1}) = \sum_{\ell=0}^{R} \ell a^{\ell-1} \sum_{m=0}^{2R+1-2\ell} a^m \geq 0, \tag{B.14}$$

since $a > 0$ from (4.3). But $b > 1$ and $b > a$ from (4.6). It follows from (B.12) and (B.14) that $\Omega > 0$.

Appendix C. We here show that $\Gamma > 0$, where Γ is given by (5.8). If we use (4.10) to express ζ in terms of a and b, then it is found that

$$\begin{aligned}&\left[\frac{1}{(b - 1)} + \frac{(1 - a^{R+1})}{(1 - a)}\right]\Gamma = \left[\frac{1}{(b - 1)} + \frac{(1 - a^{R+1})}{(1 - a)}\right]^2 \\ &+ (1 - \kappa)a^{R+1}\left[\frac{a(1 - a^R)}{(1 - a)^2} - \frac{R}{(1 - a)} - \frac{b}{(b - 1)^2} - \frac{R}{(b - 1)}\right].\end{aligned} \tag{C.1}$$

Now,

$$\frac{1}{(b - 1)} + \frac{(1 - a^{R+1})}{(1 - a)} = \frac{1}{(b - 1)}\left[(b - a)\frac{(1 - a^R)}{(1 - a)} + a^R b\right]. \tag{C.2}$$

We let

$$f(a,b) = b + R(b-1) + \left[\frac{R}{(1-a)} - \frac{a(1-a^R)}{(1-a)^2}\right](b-1)^2 \tag{C.3}$$

and

$$g(a,b) = \left[(b-a)\frac{(1-a^R)}{(1-a)} + a^R b\right]^2 - a^{R+1} f(a,b). \tag{C.4}$$

Then,

$$(b-1)^2\left[\frac{1}{(b-1)} + \frac{(1-a^{R+1})}{(1-a)}\right]\Gamma = \kappa a^{R+1} f(a,b) + g(a,b). \tag{C.5}$$

Since $a > 0$ and $b > 1$, it follows from (B.13) and (C.3) that $f(a,b) > 0$. We may express $f(a,b)$ in the form

$$f(a,b) = \frac{R(b-a)(b-1)}{(1-a)} + \frac{(1-a^R)}{(1-a)^2}(b-a)(1-ab) + a^R b. \tag{C.6}$$

From (C.4), after some algebra, it is found that

$$g(a,b) = (b-a)[\ell(a)(b-a) + (R+1)a^{R+1}], \tag{C.7}$$

where

$$\ell(a) = \frac{(1-a^{R+1})^2}{(1-a)^2} - a^{R+1}\left[\frac{R}{(1-a)} - \frac{a(1-a^R)}{(1-a)^2}\right]. \tag{C.8}$$

Hence, from (B.13),

$$\ell(a) = \sum_{r=0}^{R} a^r \sum_{s=0}^{R} a^s - \sum_{\ell=0}^{R} \ell a^{2R-\ell+1} = \sum_{n=0}^{R} (n+1)a^n > 0, \tag{C.9}$$

since $a > 0$. However, $b > a$, and it follows from (C.7) that $g(a,b) > 0$. Finally, since $\kappa > 0$ and $f(a,b) > 0$, we obtain $\Gamma > 0$.

Acknowledgment. The authors would like to thank the referees for their helpful suggestions for improving the presentation.

REFERENCES

[1] F. P. Kelly, *Blocking probabilities in large circuit-switched networks*, Adv. Appl. Probab., 18 (1986), pp. 473–505.

[2] F. P. Kelly, *Loss networks*, Ann. Appl. Probab., 1 (1991), pp. 319–378.

[3] D. Mitra and R. J. Gibbens, *State-dependent routing on symmetric loss networks with trunk reservations* II: *Asymptotics, optimal design*, Ann. Oper. Res., 35 (1992), pp. 3–30.

[4] D. Mitra, R. J. Gibbens, and B. D. Huang, *Analysis and optimal design of aggregated-least-busy-alternative routing on symmetric loss networks with trunk reservations*, in Teletraffic and Datatraffic in a Period of Change, Proceedings of the 13th International Teletraffic Congress (ITC-13), Copenhagen, 1991, A. Jensen and V. B. Iversen, eds., Elsevier (North–Holland), Amsterdam, 1991, pp. 477–482.

[5] D. Mitra, R. J. Gibbens, and B. D. Huang, *State-dependent routing on symmetric loss networks with trunk reservations* I, IEEE Trans. Comm., 41 (1993), pp. 400–411.

[6] J. A. Morrison, *Blocking probabilities for a single link with trunk reservation*, J. Math. Anal. Appl., 203 (1996), pp. 401–434.

[7] P. J. Hunt and C. N. Laws, *Optimization via trunk reservation in single resource loss systems under heavy traffic*, Ann. Appl. Probab., 7 (1997), pp. 1058–1079.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 98–121

NONADIABATIC CORRECTIONS TO THE HANNAY–BERRY PHASE*

SEAN B. ANDERSSON†

Abstract. The effect of the Coriolis force on a moving system can be described as a holonomy with respect to a particular connection known as the Cartan–Hannay–Berry connection. The resulting geometric phase is called the Hannay–Berry phase, and it provides direct information about the imposed motion on the system. This approach assumes that the imposed motion is adiabatic. In this paper we describe the use of Hamiltonian perturbation theory to develop nonadiabatic corrections to the Hannay–Berry phase for a moving system. The technique is illustrated by applying it to a rotating free-floating spring-jointed equal-sided four-bar mechanism.

Key words. geometric phases, perturbations, systems with slow and fast motions, averaging of perturbations

AMS subject classifications. 81Q70, 37J40, 70K70, 70K65

DOI. 10.1137/040606600

1. Introduction.

When a system undergoes an imposed motion, the external forces can alter the natural dynamics. A simple example is the Foucault pendulum, where the rotation of the Earth causes a precession of the swing plane of the pendulum. Classically the effect of the external forces is captured by introducing fictitious forces (the centrifugal, Coriolis, and Euler forces) into the moving system (see, e.g., [26]). When the rate of the imposed motion is slow with respect to the time scale of the nominal dynamics, that is, with respect to the dynamics in the absence of any external forces, one does not expect the imposed motion to fundamentally alter the behavior of the system. (This is essentially a statement of the averaging principle. It should be kept in mind that this principle is simply one of physical intuition and not a theorem. See [6] for further comments.) For the Foucault pendulum, the rotation of the Earth slowly shifts the swing plane; on short time scales the motion of the bob is well approximated by ignoring the effects of the imposed rotation. (See [2] for a historical narrative of Foucault and the pendulum experiment.)

An effect analogous to the precession of the swing plane of the Foucault pendulum due to the Coriolis force has been observed in many other rotating dynamical systems including rotating vibrating beams [18], tuning forks [27], and shells [10]. Vibratory gyroscopes take advantage of the effect to sense the imposed rotation, and a wide variety of designs have been proposed and built [24]. Most modern analyses of slowly rotating systems are linear in nature and view the Coriolis force as providing a coupling between two vibratory modes of the system (see, e.g., [30]). However, it is desirable to have a method which, at least in principle, can be extended to a nonlinear theory and which provides a unified setting for understanding a variety of systems in which

*Received by the editors April 12, 2004; accepted for publication (in revised form) March 30, 2005; published electronically October 3, 2005. This research was supported in part by Army Research Office ODDR&E MURI97 Program grant DAAG55-97-1-0114 to the Center for Dynamics and Control of Smart Structures (through Harvard University), Army Research Office ODDR&E MURI01 Program grant DAAD19-01-1-0465 to the Center for Communicating Networked Control Systems (through Boston University), and Air Force Office of Scientific Research grant F49620-01-0415, and by a fellowship from the ARCS foundation.

http://www.siam.org/journals/siap/66-1/60660.html

†Division of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 (sanderss@deas.harvard.edu).

the Coriolis force plays a role. A technique developed by Marsden, Montgomery, and Ratiu, known as the moving systems approach, provides such a method [25]. The moving systems approach descends from the classical work of Cartan [11] and describes the effect of the Coriolis force as a geometric phase (holonomy) with respect to a particular Ehresmann connection known as the Cartan–Hannay–Berry connection. Research on geometric phases in physical systems was spurred by the work of Berry in the early 1980s on the geometric shift of the phase in a quantum system as a result of an adiabatic variation of a parameter-dependent Hamiltonian [7]. In previous work we have applied the moving systems approach to show that the Hannay–Berry phase in a rotating equal-sided four-bar mechanism is zero [4] and to show that the precession of the nodal points in a rotating vibrating ring can be understood as a Hannay–Berry phase [3].

Inherent in the moving systems approach is an assumption of adiabaticity; that is, it is assumed that the imposed motion is infinitely slow. In practice, of course, while the imposed motion may be extremely slow with respect to the nominal dynamics of the system, it is not infinitely slow, and neglecting this fact introduces some error. By accounting for the nonadiabatic nature of the imposed motion, more accurate models of the resulting response of the system can be developed.

Since Berry's original work on geometric phases in quantum systems, various techniques have been proposed to account for the finite rate of change of the parameters in the Hamiltonian. Berry developed an iterative scheme in which the geometric phase at each step is incorporated into the nominal dynamics [8], while other authors have showed that the Berry phase can be viewed as the first-order term in a perturbation expansion of the system [15, 33]. This work has been applied to systems such as nuclear quadrupole resonance [34], hysteresis loops in manganese acetate crystals [16], and magnetic resonance [19]. A few authors have also considered the effect of the finite rate of change of the parameters on the Hannay angles, the geometric phase in classical integrable systems. Bhattacharjee and Sen used a perturbative method [9], which was later compared by Gjaja and Bhattacharjee to a classical analogue of the iterative scheme of Berry [20].

In this work we consider a perturbative approach to account for the nonadiabatic nature of the imposed motion. However, classical perturbation techniques ignore the underlying geometric structure of the Hamiltonian system to which the moving systems method applies. We will show that the averaged Hamiltonian which gives rise to the Hannay–Berry phase Poisson commutes with the nominal Hamiltonian, using the canonical Poisson bracket on the phase space. This leads us to Hamiltonian perturbation theory and Hamiltonian normal forms. We note that when not considering the imposed motion as a small perturbation from the nominal system, the effect of the external forces is often incorporated into an amended potential (see, e.g., [26]). Here we wish to develop a technique which takes advantage of the fact that the imposed motion is slow when describing the resulting system dynamics. If the original system is integrable, then the techniques developed here are similar to a perturbation expansion in the action-angle variables as in [23].

In the next section we briefly recall the moving system approach, and in section 3 give a short introduction to Hamiltonian normal form theory. In section 4 the theory of normal forms is applied to the moving systems approach to develop higher-order corrections. The technique is illustrated in section 5 by applying it to a spring-jointed equal-sided four-bar mechanism, where we show that the effect of an imposed rotation is zero to second order.

2. The moving systems approach. Inspired by classical examples such as the Foucault pendulum and the ball in a hoop, one is naturally led to consider the effect in the phase space of a mechanical system as a parameter is slowly varied along a closed loop in parameter space. Well-known examples in quantum physics, optics, and other settings [7, 31, 35] reveal that the essential calculation is a geometric one and is in fact captured by the holonomy of a connection on a fiber bundle. (For background material on fiber bundles, connections, and holonomy see [17, 28].)

Marsden, Montgomery, and Ratiu have developed an approach to moving systems using the tools of Ehresmann connections on fiber bundles [25]. Here we provide a brief review of their approach. Let S be a Riemannian manifold, and let M be the space of embeddings of a manifold Q into S. We think of S as the ambient space in which Q is being moved and of Q as the configuration space for a system of interest. A tangent vector to M at m is a map $u_m : Q \to TS$ such that $u_m(q) \in T_{m(q)}S$. Given a tangent vector $u_m(q)$, one can construct a tangent vector to T_qQ as follows. Relative to the metric on S, orthogonally project $u_m(q)$ to $T_{m(q)}m(Q) \subset (T_qm)(T_qQ)$, denote this vector $u_m^T(q)$, and then pull-back $u_m^T(q)$ by Tm^{-1} to T_qQ. This natural construction defines an Ehresmann connection on the product bundle $\pi : Q \times M \to M$ as follows.

Definition 2.1 (see [25]). *The* Cartan connection *on $\pi : Q \times M \to M$ is given by the vertical-valued one-form γ_c defined by*

$$\gamma_c(q, m)(v_q, u_m) = (v_q + (T^{-1}m \circ u_m^T)q, 0). \tag{2.1}$$

The Cartan connection induces a connection on $\rho : T^*Q \times M \to M$ as follows.

Definition 2.2 (see [25]). *The* induced Cartan connection *on $\rho : T^*Q \times M \to M$ is given by the vertical-valued one-form γ_o defined by*

$$\gamma_o(\alpha_q, m)(U_{\alpha_q}, u_m) = \left(U_{\alpha_q} + X_{\mathcal{P}(u_m)}(\alpha_q), 0\right), \tag{2.2}$$

where $\mathcal{P}(u_m)$ is the function defined by

$$(\mathcal{P}(u_m))\,\alpha_q = \alpha_q \cdot (T^{-1}m \circ u_m^T)(q) \tag{2.3}$$

and $X_{\mathcal{P}(u_m)}$ is the Hamiltonian vector field of $\mathcal{P}(u_m)$.

To separate the effects of the imposed motion on the system (as defined by the embeddings m_t) from the nominal dynamics (when the imposed motion is zero) we assume we are given a left action of a compact Lie group G on T^*Q with respect to which we can define an average and then make the following definition.

Definition 2.3 (see [25]). *The* Cartan–Hannay–Berry connection *on $\rho : T^*Q \times M \to M$ is given by the vertical-valued one-form γ defined by*

$$\gamma(\alpha_q, m)(U_{\alpha_q}, u_m) = \left(U_{\alpha_q} + X_{\langle \mathcal{P}(u_m)\rangle}(\alpha_q), 0\right), \tag{2.4}$$

where $\langle\cdot\rangle$ denotes the average with respect to the action of the Lie group G.

In [25] Marsden, Montgomery, and Ratiu show that this is an Ehresmann connection. The horizontal lift of a vector field Z on M relative to γ is

$$(\mathrm{hor}\, Z)(\alpha_q, m) = \left(-X_{\langle \mathcal{P}(Z(m))\rangle}(\alpha_q), Z(m)\right). \tag{2.5}$$

Definition 2.4 (see [25]). *The holonomy of the Cartan–Hannay–Berry connection is called the* Hannay–Berry phase *for a moving system.*

2.1. The adiabatic assumption. The Hannay–Berry phase captures the effects of the imposed motion on a system under the assumption that this imposed motion is slow with respect to the nominal dynamics. To better understand this adiabatic assumption, we consider the following system (as in [26]). If a particle in Q is following a curve $q(t)$ and if Q is in turn being moved in the ambient space S by superposing the motion m_t, then the path of the particle in S is given by $m_t(q(t))$. The velocity in S is then

$$T_{q(t)}m_t\dot{q}(t) + \mathcal{Z}_t(m_t(q(t))), \tag{2.6}$$

where $\mathcal{Z}_t(m_t(q)) = \frac{d}{dt}m_t(q)$ (with q viewed as fixed). The standard Lagrangian is given by the kinetic energy minus the potential energy,

$$L(q,v) = \frac{1}{2}\|T_{q(t)}m_t v + \mathcal{Z}_t(m_t(q))\|^2 - V(q) - U(m_t(q)). \tag{2.7}$$

Here V is a given potential on Q, and U is a given potential on S. To compute the associated Hamiltonian we take the Legendre transform. Taking the derivative of L with respect to v in the direction w yields

$$\frac{\partial L}{\partial v}\cdot w = p\cdot w = \langle T_{q(t)}m_t v + \mathcal{Z}_t(m_t(q(t)))^T, T_{q(t)}m_t w\rangle_{m_t(q(t))},$$

where $p\cdot w$ is the natural pairing between the covector $p\in T^*_{q(t)}Q$ and the vector $w\in T_{q(t)}Q$, $\langle\cdot,\cdot\rangle_{m_t(q(t))}$ denotes the metric inner product on S at the point $m_t(q(t))$, and ^{T} denotes the orthogonal projection to $Tm_t(Q)$ using the metric of S at $m_t(q(t))$. Q inherits a metric from S such that m_t is an isometry for each t. Thus

$$\begin{aligned} p\cdot w &= \left\langle v + \left(T_{q(t)}m_t\right)^{-1}\mathcal{Z}_t^T(m_t(q(t)))^T, w\right\rangle_{q(t)} \\ \Rightarrow p &= \left(v + \left(T_{q(t)}m_t\right)^{-1}\mathcal{Z}_t^T(m_t(q(t)))^T\right)^\flat \triangleq (v+Z_t)^\flat, \end{aligned} \tag{2.8}$$

where we have defined the tangent vector Z_t. Here $\flat : T_qQ \to T_q^*Q$ is the map defined by

$$z^\flat\cdot w = \langle z, w\rangle_q \quad \forall w\in T_qQ. \tag{2.9}$$

The Hamiltonian for the moving system is given by

$$H(q,p) = \frac{1}{2}\|p\|^2 - \mathcal{P}(Z_t) - \frac{1}{2}\|\mathcal{Z}_t^\perp\|^2 + V(q) + U(m_t(q)), \tag{2.10}$$

where $\mathcal{Z}_t^\perp = \mathcal{Z}_t - \mathcal{Z}_t^T$ is the orthogonal complement of $\mathcal{Z}_t$ and $\mathcal{P}(Z_t)$ is the function on T^*Q (defined in (2.3)) given by

$$\mathcal{P}(Z_t)(q,p) = p\cdot Z_t(q).$$

The nominal Hamiltonian H_0 is defined by setting $\mathcal{Z}_t = 0$ and $U = 0$. The term $\mathcal{P}(Z_t)$ captures what are classically referred to as the Coriolis terms, and $\|\mathcal{Z}_t^\perp\|^2$ captures the centrifugal terms.

Recall now that we have a compact Lie group G acting on T^*Q on the left. We assume that the group action leaves the nominal Hamiltonian invariant. Applying the corresponding average, we obtain

$$\langle H\rangle(q,p) = \frac{1}{2}\|p\|^2 - \langle\mathcal{P}(Z_t)\rangle - \frac{1}{2}\langle\|\mathcal{Z}_t^\perp\|^2\rangle + V(q) + \langle U(m_t(q))\rangle. \tag{2.11}$$

Invoking the adiabatic assumption, we discard $\langle \|\mathcal{Z}_t^{\perp}\|^2 \rangle$ since it is small with respect to the other terms in the averaged Hamiltonian. After discarding the centrifugal terms, the dynamics of the Hamiltonian system are governed by the Hamiltonian vector field

$$X_{\langle H \rangle} = X_{H_0} - X_{\langle \mathcal{P}(Z_t) \rangle} + X_{\langle U \circ m_t \rangle}. \tag{2.12}$$

The second term captures the effect of the imposed motion in the adiabatic limit and is precisely the term given by the horizontal lift of the vector field $\mathcal{Z}_t$ with respect to the Cartan–Hannay–Berry connection as defined in (2.5).

2.2. Geometric character of the Hannay–Berry phase. The effect of the vector field $X_{\langle \mathcal{P}(Z_t) \rangle}$ is geometric in nature. By this we mean that the resulting change in the system is independent of the parametrization of the curve followed in the base space M; i.e., the effect depends only on the loop itself and not on how it is traversed. To see this explicitly, recall that the vector field $-X_{\langle \mathcal{P}(Z_t) \rangle}$ is the horizontal lift of a vector field $\mathcal{Z}_t$ on the base space M to the fiber T^*Q with respect to the Cartan–Hannay–Berry connection and is thus a linear map of $\mathcal{Z}_t$. Denoting points in T^*Q by z, the ordinary differential equation defining the Hannay–Berry phase may be expressed as

$$\frac{dz}{dt} = -X_{\langle \mathcal{P}(Z_t) \rangle} = D(z)\mathcal{Z}_t. \tag{2.13}$$

In coordinates, $D(z)$ is a matrix taking tangent vectors on M to tangent vectors on T^*Q. We now change the time parametrization by taking $t \mapsto \tau(t)$ with $\frac{d\tau}{dt}$ strictly positive. Under this new parametrization, the vector field $\mathcal{Z}_t$ is scaled by $\frac{d\tau}{dt}$, and thus

$$\frac{dz}{dt} = \frac{dz}{d\tau}\frac{d\tau}{dt} = D(z)\left(\frac{d\tau}{dt}\right)\mathcal{Z}_\tau. \tag{2.14}$$

We then have

$$\frac{dz}{d\tau} = D(z)\mathcal{Z}_\tau, \tag{2.15}$$

which has the same form as the differential equation in the original parametrization.

3. Hamiltonian normal forms. The theory of Hamiltonian normal forms (or Hamiltonian perturbation theory) is a generalization of Lie perturbation techniques (see, e.g., [12]), which in turn is built upon the perturbation methods developed by Poincaré and von Ziepel (see [5] for historical comments). In this section we provide a brief description of the theory and refer the reader to [13, 14] for more details and further references.

Recall the definition of a Poisson manifold.

DEFINITION 3.1. *A* Poisson manifold *is a smooth manifold M together with a $\mathbb{R}$-bilinear map on $C^\infty(M)$,*

$$\{\cdot,\cdot\} : C^\infty(M) \times C^\infty(M) \to C^\infty(M),$$

which for all $f, g, h \in C^\infty(M)$ satisfies

(i) *skew symmetry:* $\{f, g\} = -\{g, f\}$,
(ii) *Leibniz identity:* $\{f, gh\} = \{f, g\}h + g\{f, h\}$,
(iii) *Jacobi identity:* $\{f, \{g, h\}\} + \{g, \{h, f\}\} + \{h, \{f, g\}\} = 0$.

Consider a Poisson manifold $(M, \{\cdot,\cdot\})$. Let $\mathcal{F}(M)$ be the vector space of formal power series in ϵ with coefficients in $C^\infty(M)$. That is,

$$\mathcal{F}(M) = \left\{ f_\epsilon \in C^\infty(M) \,\middle|\, f_\epsilon = \sum_{i=0}^{\infty} \epsilon^i f_i,\ f_i \in C^\infty(M) \right\}. \tag{3.1}$$

Let $\mathrm{ad}_f g = \{g, f\}$ and define $\mathrm{ad}_f^0 g = g$. Recursively define ad_f^i by

$$\mathrm{ad}_f^i g = \{\mathrm{ad}_f^{i-1} g, f\}. \tag{3.2}$$

DEFINITION 3.2. *The* Lie series *of f is the formal power series*

$$\phi_\epsilon^f = \exp(\epsilon \mathrm{ad}_f) = \sum_{i=0}^{\infty} \frac{\epsilon^i}{i!} \mathrm{ad}_f^i. \tag{3.3}$$

Here ϕ_ϵ^f is the formal flow of the Hamiltonian vector field X_f with ϵ as the time parameter.

DEFINITION 3.3. *For $f \in \mathcal{F}(M)$, X_f is said to have* periodic flow *if there exists a positive, smooth function T on M such that for all $m \in M$ and for all $g \in \mathcal{F}(M)$, $((\phi_T^f)^*(g))(m) = g(m)$.*

DEFINITION 3.4. *Consider $H \in \mathcal{F}(M)$ and suppose X_{H_0} has periodic flow. H is said to be in* normal form with respect to H_0 *if $\{H_0, H_i\} = 0$ for $i = 1, 2, \ldots$ and in* normal form up to order n with respect to H_0 *if $\{H_0, H_i\} = 0$, $i = 1, 2, \ldots, n$.*

To bring a Hamiltonian into normal form we will use a formal change of coordinates of the form ϕ_ϵ^f for some appropriate $f \in \mathcal{F}(m)$. The following lemma from [14] shows how the Hamiltonian is modified under such a change of coordinates.

LEMMA 3.5 (see [14]). *Let $H, f \in \mathcal{F}(M)$. If ϕ_ϵ^f is the flow of X_f, then*

$$\left(\phi_\epsilon^f\right)^* H = \exp(\epsilon \mathrm{ad}_f) H. \tag{3.4}$$

The use of (3.3) in (3.4) yields

$$\begin{aligned} \left(\phi_\epsilon^f\right)^* H &= \sum_{i=0}^{\infty} \frac{\epsilon^i}{i!} \mathrm{ad}_f^i \left(\sum_{j=0}^{\infty} \epsilon^j H_j \right) \\ &= H_0 + \epsilon(H_1 + \mathrm{ad}_f H_0) + \epsilon^2 \left(H_2 + \mathrm{ad}_f H_1 + \frac{1}{2} \mathrm{ad}_f^2 H_0 \right) + O(\epsilon^3). \end{aligned} \tag{3.5}$$

To bring H into first-order normal form, we seek a function $f \in \mathcal{F}(M)$ such that

$$\{H_0, H_1 + \mathrm{ad}_f H_0\} = 0. \tag{3.6}$$

To find this function we use the following lemma from [13].

LEMMA 3.6 (see [13]). *If X_{H_0} has periodic flow on M, then*

$$C^\infty(M) = \ker\left(\mathrm{ad}_{H_0}\right) \oplus \mathrm{im}\left(\mathrm{ad}_{H_0}\right). \tag{3.7}$$

Let $\langle\cdot\rangle$ denote the average over the orbits of H_0; i.e., for $g \in C^\infty(M)$

$$\langle g \rangle = \frac{1}{T} \int_0^T \left(\phi_t^{H_0}\right)^* g \, dt. \tag{3.8}$$

To use Lemma 3.6 we first show the following.

LEMMA 3.7. *Let $g \in C^\infty(M)$. Then $\langle g \rangle \in \ker(\mathrm{ad}_{H_0})$.*

Proof. The equation $\mathrm{ad}_{H_0}\langle g\rangle = \rho$ is equivalent to the dynamical system

$$\frac{d}{dt}\left(\phi_t^{H_0}\right)^* \langle g\rangle = \left(\phi_t^{H_0}\right)^* \rho$$

(see, e.g., Proposition 10.2.3 of [26]). We show that $\rho = 0$. The use of the definition of the average of g yields

$$\begin{aligned}
\frac{d}{dt}\left(\phi_t^{H_0}\right)^* \langle g\rangle &= \frac{d}{dt}\left(\phi_t^{H_0}\right)^* \frac{1}{T}\int_0^T \left(\phi_\tau^{H_0}\right)^* g\, d\tau \\
&= \frac{1}{T}\frac{d}{dt}\int_0^T \left(\phi_{t+\tau}^{H_0}\right)^* g\, d\tau \\
&= \frac{1}{T}\frac{d}{dt}\int_t^{T+t} \left(\phi_\sigma^{H_0}\right)^* g\, d\sigma \\
&= \frac{1}{T}\left(\left(\phi_{T+t}^{H_0}\right)^* g - \left(\phi_t^{H_0}\right)^* g\right) \\
&= \frac{1}{T}\left(\left(\phi_t^{H_0}\right)^* g - \left(\phi_t^{H_0}\right)^* g\right) \\
&= 0.
\end{aligned}$$

Therefore $\mathrm{ad}_{H_0}\langle g\rangle = 0$. □

To put the Hamiltonian into normal form to first order we write

$$H_1 = \langle H_1\rangle + (H_1 - \langle H_1\rangle) \tag{3.9}$$

and then substitute (3.9) into (3.6). Thus

$$\begin{aligned}
0 &= \{H_0, \langle H_1\rangle + (H_1 - \langle H_1\rangle) + \mathrm{ad}_f H_0\} \\
&= \{H_0, \langle H_1\rangle\} + \{H_0, (H_1 - \langle H_1\rangle) + \mathrm{ad}_f H_0\} \\
&= \{H_0, (H_1 - \langle H_1\rangle) + \mathrm{ad}_f H_0\}, \qquad (3.10)
\end{aligned}$$

where the last step follows from the fact that, from 3.7, $\langle H_1\rangle \in \ker(\mathrm{ad}_{H_0})$. We then seek a solution to the *homological equation*

$$\mathrm{ad}_f H_0 = -(H_1 - \langle H_1\rangle), \tag{3.11}$$

where f is the unknown function.

PROPOSITION 3.8 (see [13]). *The solution to* (3.11) *is given by*

$$f = \frac{1}{T}\int_0^T t\left(\phi_t^{H_0}\right)^* (H_1 - \langle H_1\rangle)\, dt. \tag{3.12}$$

Proof. Let $g = -\mathrm{ad}_f H_0 = \mathrm{ad}_{H_0} f$. This is equivalent to the dynamical system

$$\frac{d}{dt}\left(\phi_t^{H_0}\right)^* f = \left(\phi_t^{H_0}\right)^* g.$$

We show that $g = H_1 - \langle H_1 \rangle$ by direct substitution. Thus

$$\begin{aligned}
\frac{d}{dt}\left(\phi_t^{H_0}\right)^* f &= \frac{d}{dt}\left(\frac{1}{T}\int_0^T \tau \left(\phi_\tau^{H_0}\right)^* (H_1 - \langle H_1\rangle) d\tau\right) \\
&= \frac{1}{T}\frac{d}{dt}\int_0^T \tau \left(\phi_{t+\tau}^{H_0}\right)^* (H_1 - \langle H_1\rangle) d\tau \\
&= \frac{1}{T}\frac{d}{dt}\int_t^{t+T} (\sigma - t)\left(\phi_\sigma^{H_0}\right)^* (H_1 - \langle H_1\rangle) d\sigma \\
&= \frac{1}{T}\left(T\left(\phi_{t+T}^{H_0}\right)^* (H_1 - \langle H_1\rangle) - \int_t^{t+T} \left(\phi_\sigma^{H_0}\right)^* (H_1 - \langle H_1\rangle) d\sigma\right) \\
&= \left(\phi_{t+T}^{H_0}\right)^* (H_1 - \langle H_1\rangle) - \frac{1}{T}\int_t^{t+T} \left(\phi_\sigma^{H_0}\right)^* H_1 d\sigma + \langle H_1\rangle \\
&= \left(\phi_{t+T}^{H_0}\right)^* (H_1 - \langle H_1\rangle) - \langle H_1\rangle + \langle H_1\rangle \\
&= \left(\phi_t^{H_0}\right)^* (H_1 - \langle H_1\rangle).
\end{aligned}$$

Therefore $g = (H_1 - \langle H_1\rangle)$. From this the proposition follows. □

With this choice of f, the Hamiltonian in (3.4) becomes

$$\exp(\epsilon \mathrm{ad}_f) H = H_0 + \epsilon\langle H_1\rangle + \epsilon^2\left(H_2 + \mathrm{ad}_f H_1 + \frac{1}{2}\mathrm{ad}_f^2 H_0\right) + O(\epsilon^3), \tag{3.13}$$

which is in first-order normal form. Notice that if we wish to bring the Hamiltonian into normal form only up to first order, then there is no need to explicitly calculate the generating function f.

To bring the function into normal form up to second order we repeat the process, now on the once transformed Hamiltonian. This time we seek a generating function of the form ϵg. Application of the corresponding change of coordinates results in

$$\begin{aligned}
&\exp(\epsilon \mathrm{ad}_{\epsilon g})\left(\exp(\epsilon \mathrm{ad}_f) H\right) = \sum_{i=0}^{\infty} \frac{\epsilon^i}{i!}\mathrm{ad}_{\epsilon g}^i \left(\exp(\epsilon \mathrm{ad}_f) H\right) \\
&\quad = H_0 + \epsilon\langle H_1\rangle + \epsilon^2\left(H_2 + \mathrm{ad}_f H_1 + \frac{1}{2}\mathrm{ad}_f^2 H_0 + \mathrm{ad}_g H_0\right) + O(\epsilon^3).
\end{aligned} \tag{3.14}$$

The homological equation which needs to be solved is

$$\mathrm{ad}_g H_0 = -\left(H_2 + \mathrm{ad}_f H_1 + \frac{1}{2}\mathrm{ad}_f^2 H_0 - \left(\langle H_2\rangle + \langle \mathrm{ad}_f H_1\rangle + \left\langle \frac{1}{2}\mathrm{ad}_f^2 H_0\right\rangle\right)\right). \tag{3.15}$$

From Proposition 3.8, the solution to this equation is

$$\begin{aligned}
g = \frac{1}{T}\int_0^T t\left(\phi_t^{H_0}\right)^* \Big[&H_2 - \langle H_2\rangle + \mathrm{ad}_f H_1 - \langle \mathrm{ad}_f H_1\rangle \\
&+ \frac{1}{2}\left(\mathrm{ad}_f^2 H_0 - \langle \mathrm{ad}_f^2 H_0\rangle\right)\Big] dt.
\end{aligned} \tag{3.16}$$

With this choice our transformed Hamiltonian becomes

$$\begin{aligned}
&\exp(\epsilon \mathrm{ad}_{\epsilon g})\left(\exp(\epsilon \mathrm{ad}_f) H\right) \\
&\quad = H_0 + \epsilon\langle H_1\rangle + \epsilon^2\left(\langle H_2\rangle + \langle \mathrm{ad}_f H_1\rangle + \frac{1}{2}\langle \mathrm{ad}_f^2 H_0\rangle\right) + O(\epsilon^3).
\end{aligned} \tag{3.17}$$

The Hamiltonian can be placed into normal form up to arbitrary order n by repeating this process.

In practice one places the system into normal form up to some desired order and then truncates the higher-order terms. The truncated Hamiltonian gives an approximation to the original system. Since the coefficients of ϵ^i in the transformed Hamiltonian all commute with H_0 for $i = 1, 2, \ldots, n$, the flow of the corresponding Hamiltonian vector field of the higher-order terms also commutes with the flow of the nominal system. Thus for a Hamiltonian in first-order normal form we have

$$\phi_t^{H_0+\epsilon\langle H_1\rangle}(m) = \phi_t^{\epsilon\langle H_1\rangle} \circ \phi_t^{H_0}(m), \qquad m \in M, \tag{3.18}$$

and the first-order terms give rise naturally to a first-order correcting symplectic map given by the flow of the Hamiltonian system $\epsilon\langle H_1\rangle$. For systems in higher-order normal form, however, while the functions at each order do Poisson commute with H_0, they do not in general commute with each other, and thus a system in nth-order normal form defines a single nth-order correcting symplectic map.

4. Normal forms and the Hannay–Berry phase. In the setting of the moving systems approach the Poisson manifold is T^*Q together with the canonical Poisson bracket defined by

$$\{f, g\} = \sum_{i=1}^{n} \frac{\partial f}{\partial q^i}\frac{\partial g}{\partial p_i} - \frac{\partial f}{\partial p_i}\frac{\partial g}{\partial q^i}, \quad f, g \in C^\infty(M). \tag{4.1}$$

To apply Hamiltonian normal form theory we make a few additional assumptions on the Hamiltonian of a moving system, (2.10). We first assume that the potential U on S is constant and drop it from the Hamiltonian. Next we assume that $\mathcal{Z}_t(m_t(q))$ can be written in the form

$$\mathcal{Z}_t(m_t(q)) = \epsilon\hat{\mathcal{Z}}_t(m_t(q)) \tag{4.2}$$

for some parameter ϵ. For example, if M is a Riemannian manifold and $\mathcal{Z}_t$ is a constant magnitude vector field, then we may take $\epsilon = \|\mathcal{Z}_t\|$ and $\hat{\mathcal{Z}}_t = \frac{\mathcal{Z}_t}{\|\mathcal{Z}_t\|}$. If $\|\mathcal{Z}_t\|$ is not constant, then one could take ϵ to be the average magnitude of $\|\mathcal{Z}_t\|$ over the loop in M starting at the given initial condition. However, the form of $\mathcal{Z}_t$ in (4.2) is often natural to the problem, and in general $\hat{\mathcal{Z}}_t$ is not a unit vector. Under these assumptions the Hamiltonian can be written as

$$H(q, p) = H_0(q, p) + \epsilon H_1(q, p) + \epsilon^2 H_2(q, p), \tag{4.3}$$

where

$$H_0(q, p) = \frac{1}{2}\|p\|^2 + V(q), \tag{4.4}$$

$$H_1(q, p) = -\mathcal{P}(\hat{Z}_t), \tag{4.5}$$

$$H_2(q, p) = -\frac{1}{2}\|\hat{\mathcal{Z}}_t^{\perp}\|^2. \tag{4.6}$$

Finally we assume that H_0 has periodic flow with period T. We then have a natural action of S^1 on T^*Q, given by $\phi_t^{H_0}$, the flow of X_{H_0}. Let $\langle\cdot\rangle$ denote the average with respect to this group action; i.e., for a smooth function f on T^*Q we have

$$\langle f\rangle = \frac{1}{T}\int_0^T \left(\phi_t^{H_0}\right)^* f\, dt. \tag{4.7}$$

In general, the parameter ϵ captures the rate of the imposed motion on the system. In the adiabatic limit, then, ϵ goes to zero and the terms in ϵ^2 are negligible. (Note that the terms in ϵ are not negligible in the adiabatic limit since the time parametrization of the (slow) imposed motion scales as $\frac{1}{\epsilon}$. See section 2.2.) In what follows we are interested in relaxing the adiabatic condition; i.e., we assume that while ϵ is small, the terms in ϵ^2 are not negligible. The truncated Hamiltonian defined by

$$\langle H\rangle^{(1)}(q,p) = H_0(q,p) + \epsilon\langle H_1\rangle(q,p) = H_0(q,p) - \epsilon\langle \mathcal{P}(\hat{Z}_t)\rangle \tag{4.8}$$

is in first-order normal form. The flow of the system to first order is then given by

$$\phi_t^{H_0-\epsilon\langle \mathcal{P}(\hat{Z}_t)\rangle}(q,p) = \phi_t^{-\epsilon\langle \mathcal{P}(\hat{Z}_t)\rangle} \circ \phi_t^{H_0}(q,p), \tag{4.9}$$

and the flow of $-\langle \mathcal{P}(\hat{Z}_t)\rangle$ defines the correcting symplectic map to first order. Thus, in the setting where the group action on T^*Q is given by the flow of the nominal dynamics, the Hannay–Berry phase is the first-order correction at the completion of a closed loop in parameter space.

To find a more accurate expression we express the Hamiltonian in normal form to a higher order before truncating. Let G be the generator of a change of coordinates bringing the original Hamiltonian into first-order normal form. From Proposition 3.8 and the form of H_1 in (4.5), G is given by

$$G = \frac{1}{T}\int_0^T t\left(\phi_t^{H_0}\right)^*\left[\langle \mathcal{P}(\hat{Z}_t)\rangle - \mathcal{P}(\hat{Z}_t)\right]dt. \tag{4.10}$$

From (3.17) and the form of H_2 in (4.6), the second-order truncated normal form is

$$\begin{aligned}\langle H\rangle^{(2)}(q,p) &= H_0(q,p) - \epsilon\langle \mathcal{P}(\hat{Z}_t)\rangle \\ &\quad -\epsilon^2\left(\frac{1}{2}\langle\|\hat{\mathcal{Z}}_t^{\perp}\|^2\rangle + \langle \mathrm{ad}_G\mathcal{P}(\hat{Z}_t)\rangle - \frac{1}{2}\langle \mathrm{ad}_G^2 H_0\rangle\right).\end{aligned} \tag{4.11}$$

Notice that the terms at second order in the Hamiltonian not only account for the average effect of the centrifugal force but also include additional terms involving the first-order change of coordinates. The flow of the system to second order is

$$\begin{aligned}&\phi_t^{H_0-\epsilon\langle \mathcal{P}(\hat{Z}_t)\rangle-\epsilon^2\left(\frac{1}{2}\langle\|\hat{\mathcal{Z}}_t^{\perp}\|^2\rangle+\langle \mathrm{ad}_G\mathcal{P}(\hat{Z}_t)\rangle-\frac{1}{2}\langle \mathrm{ad}_G^2 H_0\rangle\right)}(q,p) \\ &= \phi_t^{-\epsilon\langle \mathcal{P}(\hat{Z}_t)\rangle-\epsilon^2\left(\frac{1}{2}\langle\|\hat{\mathcal{Z}}_t^{\perp}\|^2\rangle+\langle \mathrm{ad}_G\mathcal{P}(\hat{Z}_t)\rangle-\frac{1}{2}\langle \mathrm{ad}_G^2 H_0\rangle\right)} \circ \phi_t^{H_0}(q,p),\end{aligned} \tag{4.12}$$

and this in general defines a correcting symplectic map to second order. If in addition the terms in ϵ Poisson commute with the terms in ϵ^2, then the second-order terms define a second-order correcting symplectic map. In this case the three Hamiltonian systems can be solved independently and their flows composed to obtain the second-order solution. This is captured in the following lemma.

LEMMA 4.1. *If*

$$\left\{\langle \mathcal{P}(\hat{Z}_t)\rangle, \frac{1}{2}\langle\|\hat{\mathcal{Z}}_t^{\perp}\|^2\rangle + \langle \mathrm{ad}_G\mathcal{P}(\hat{Z}_t)\rangle - \frac{1}{2}\langle \mathrm{ad}_G^2 H_0\rangle\right\} = 0, \tag{4.13}$$

then

$$\begin{aligned}&\phi_t^{H_0-\epsilon\langle \mathcal{P}(\hat{Z}_t)\rangle-\epsilon^2\left(\frac{1}{2}\langle\|\hat{\mathcal{Z}}_t^{\perp}\|^2\rangle+\langle \mathrm{ad}_G\mathcal{P}(\hat{Z}_t)\rangle-\frac{1}{2}\langle \mathrm{ad}_G^2 H_0\rangle\right)}(q,p) \\ &= \phi_t^{-\epsilon^2\left(\frac{1}{2}\langle\|\hat{\mathcal{Z}}_t^{\perp}\|^2\rangle+\langle \mathrm{ad}_G\mathcal{P}(\hat{Z}_t)\rangle-\frac{1}{2}\langle \mathrm{ad}_G^2 H_0\rangle\right)} \circ \phi_t^{-\epsilon\langle \mathcal{P}(\hat{Z}_t)\rangle} \circ \phi_t^{H_0}(q,p).\end{aligned} \tag{4.14}$$

Proof. The proof is immediate by the assumption of the Poisson commutativity of the functions. ∎

4.1. Time-dependence of nonadiabatic corrections. In section 2.2 we showed that the Hannay–Berry phase is a geometric phenomenon by showing that the corresponding ordinary differential equation is independent of the time parametrization. We now show that the terms in ϵ^2 in the moving systems Hamiltonian do not result in a geometric effect. Consider (4.11). For simplicity assume that the generating function for the change of coordinates is $G = 0$ and that $\{\langle H_1\rangle, \langle H_2\rangle\} = 0$ so that we can calculate the effect on the system from these two terms separately. Denote points in T^*Q by z. Noticing that $X_{\langle \|\mathcal{Z}^\perp\|^2\rangle}$ is a quadratic form in the vector field $\mathcal{Z}$ on the base space, we define

$$Y(\mathcal{Z}_t, z) = -X_{\langle \|\mathcal{Z}_t^\perp|^2\rangle}, \tag{4.15}$$

where $Y(a\mathcal{Z}_t, z) = a^2 Y(\mathcal{Z}_t, z)$. The corresponding ordinary differential equation is

$$\dot{z} = Y(\mathcal{Z}_t, z). \tag{4.16}$$

We now change the time parametrization (as in section 2.2) by taking $t \mapsto \tau(t)$ with $\frac{d\tau}{dt}$ strictly positive. Under this parametrization, the vector field $\mathcal{Z}_t$ is scaled by $\frac{d\tau}{dt}$ and thus

$$\frac{dz}{dt} = \frac{dz}{d\tau}\frac{d\tau}{dt} = Y\left(\frac{d\tau}{dt}\mathcal{Z}_\tau, z\right) = \left(\frac{d\tau}{dt}\right)^2 Y(\mathcal{Z}_\tau, z). \tag{4.17}$$

From this we have

$$\frac{dz}{d\tau} = \frac{d\tau}{dt} Y(\mathcal{Z}_\tau, z), \tag{4.18}$$

which shows the dependence on the time parametrization.

5. The equal-sided spring-jointed free-floating four-bar mechanism. In this section we apply the method developed in section 4 to an equal-sided spring-jointed free-floating four-bar mechanism which is being rotated about its center of mass. The study of the four-bar mechanism has a long history, dating at least back to the work of Grashof in the mid-nineteenth century [22]. (See also [29] and references therein.) Building upon an analysis of four-bar linkages due to Yang and Krishnaprasad [37], this system is analyzed in [4, 3] using the moving system approach, and it is shown that the Hannay–Berry phase is zero. After briefly recalling this result, we will show that the second-order effects of the imposed rotation are also zero.

Consider an equal-sided four-bar mechanism as shown in Figure 5.1. By a "bar" we mean a planar rigid body on which the center of mass and pin joints are arbitrarily located. The identical bars are labeled sequentially from 0 to 3, and on each a body-fixed frame is defined such that its origin is at the body center of mass and the x-axis is parallel to the line connecting the pin joints. The positive direction of the x-axis of the ith bar is defined to be towards the $(i+1)$th bar for $i = 0, 1, 2, 3$, where we adopt

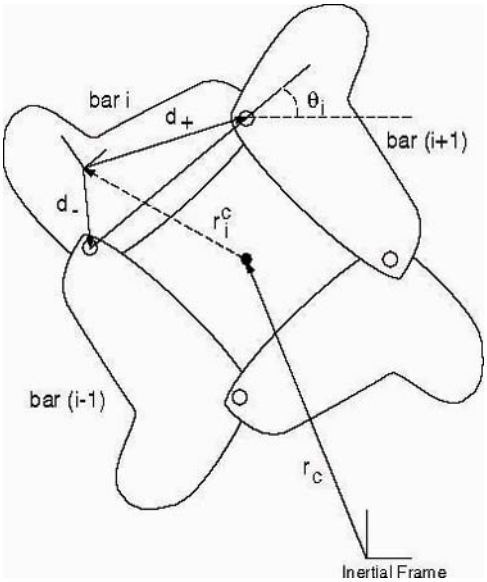


FIG. 5.1. *Equal-sided four-bar mechanism.*

the convention of modulo four addition for subscripts. We define the following:

- $\mathbf{d}_+$ the vector from the body center of mass of the ith bar to the pin joint with the $(i+1)$th bar,
- $\mathbf{d}_-$ the vector from the body center of mass of the ith bar to the pin joint with the $(i-1)$th bar,
- l the length of each bar $\|\mathbf{d}_+ - \mathbf{d}_-\|$, where $\|\cdot\|$ is the standard Euclidean norm,
- $\mathbf{r}_i^c$ the vector from the system center of mass to the ith body center of mass,
- $\mathbf{r}_c$ the vector from the origin of the inertial system to the system center of mass,
- θ_i the angle between the ith bar frame and the inertial frame,
- $\theta_{i,j}$ the angle $\theta_i - \theta_j$ between the ith and jth bars,
- I, m the moment of inertia and mass of each bar.

From Figure 5.1 we have that

$$r_{i+1}^c = r_i^c + R(\theta_i)\mathbf{d}_+ - R(\theta_{i+1})\mathbf{d}_-, \quad i = 0, 1, 2, 3, \tag{5.1}$$

where $R(\theta_i)$ is the rotation matrix given by

$$R(\theta_i) = \begin{pmatrix} \cos(\theta_i) & -\sin(\theta_i) \\ \sin(\theta_i) & \cos(\theta_i) \end{pmatrix}.$$

We use (5.1) recursively to define the loop closure constraint

$$F(r) = \sum_{i=0}^{3} R(\theta_i)(\mathbf{d}_+ - \mathbf{d}_-) = 0. \tag{5.2}$$

In [32] it is shown that the configuration space for a free-floating four-link open chain is $\mathcal{R} = \mathbb{R}^2 \times S^1 \times S^1 \times S^1 \times S^1$. The configuration space for a general four-bar mechanism is thus $S_{\text{gen}} = \{r \in \mathcal{R} | F(r) = 0\}$. For a four-bar mechanism with identical bars it can be shown that if the system is not allowed to pass through any singularities (joint angles of 0 or 2π), then the configuration space $S = \{r \in \mathcal{R} | F(r) = 0, \theta_{i+1,i} \neq 0, \pi\}$ is a smooth submanifold of $\mathcal{R}$ [3].

While in the general four-bar mechanism the relations between the angles θ_i can be quite complicated (see, for example, [29]), they have a particularly simple form for the equal-sided case, namely

$$\theta_2 = \theta_0 + \pi, \qquad \theta_3 = \theta_0 - \pi, \tag{5.3}$$

which leads to the following equalities:

$$\theta_{32} = \theta_{10}, \quad \theta_{21} = \theta_{03} = \pi - \theta_{10}, \quad \theta_{13} = \theta_{20} = \pi. \tag{5.4}$$

For the free-floating equal-sided four-bar linkage the configuration is completely specified by the choice of one global angle and one joint angle. We arbitrarily choose θ_0 and θ_{10}. After removing the singular points $\theta_{10} = 0, \pi$, the configuration space is given by $S^1 \times \{(0, \pi) \cup (0, -\pi)\}$. Since the joint angle is not allowed to pass through the singular points, we may arbitrarily choose either one of the connected components of this space to describe the configuration of our system, with the additional requirement that the initial condition lie in the component we have chosen. Without loss of generality, then, we take $S = S^1 \times (0, \pi)$ as the configuration space of the free-floating equal-sided four-bar mechanism.

The total kinetic energy of the system in the center of mass frame is given by

$$T = \frac{1}{2} I \sum_{i=0}^{3} \omega_i^2 + \frac{1}{2} m \sum_{i=0}^{3} \|\dot{r}_i^c\|^2. \tag{5.5}$$

Following [37], this can be written

$$T = \frac{1}{2} \langle \tilde{\omega}, \tilde{M} \tilde{\omega} \rangle, \tag{5.6}$$

where $\tilde{\omega} = (\omega_0, \omega_1, \omega_2, \omega_3)'$ and $\tilde{M}$ is a 4×4 symmetric matrix whose elements for the equal-sided four-bar system are

$$\tilde{M}_{ii} = I + \frac{3m}{8} \left(\|\mathbf{d}_+\|^2 + \|\mathbf{d}_-\|^2 \right), \tag{5.7}$$

$$\tilde{M}_{i,i+1} = \frac{m}{8} \left(\mathbf{d}'_- R_{i+1,i} \mathbf{d}_+ - 3 \mathbf{d}'_+ R_{i+1,i} \mathbf{d}_- \right), \tag{5.8}$$

$$\tilde{M}_{i,i+2} = -\frac{m}{8} \left(\mathbf{d}'_+ R_{i+2,i} \mathbf{d}_+ + \mathbf{d}'_- R_{i+2,i} \mathbf{d}_- \right), \tag{5.9}$$

where $'$ indicates transpose and $R_{i,j} = R(\theta_i - \theta_j)$. From (5.3) we have

$$\begin{pmatrix} \omega_2 \\ \omega_3 \end{pmatrix} = \begin{pmatrix} \omega_0 \\ \omega_1 \end{pmatrix}, \tag{5.10}$$

where $\omega_i = \dot{\theta}_i$. Let $\mathbb{I}$ denote the identity matrix. Define

$$M = \begin{pmatrix} \mathbb{I} & \mathbb{I} \end{pmatrix} \tilde{M} \begin{pmatrix} \mathbb{I} \\ \mathbb{I} \end{pmatrix}. \tag{5.11}$$

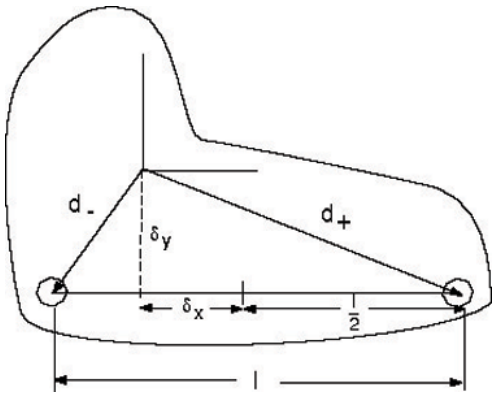


FIG. 5.2. *Single bar diagram.*

Then

$$T = \frac{1}{2} \begin{pmatrix} \omega_0 & \omega_1 \end{pmatrix} M \begin{pmatrix} \omega_0 \\ \omega_1 \end{pmatrix}. \tag{5.12}$$

M is symmetric. It depends only on the joint angles and thus, given the relations (5.4), it depends only on θ_{10}. We would like to express the entries in this matrix in terms of the parameters of the four-bar linkage. From (5.4), (5.7), (5.9), and (5.11), we have

$$M_{00} = 2I + m(\|\mathbf{d}_+\|^2 + \|\mathbf{d}_-\|^2), \tag{5.13}$$

$$M_{11} = M_{00}. \tag{5.14}$$

From (5.4), (5.8), and (5.11) we have

$$M_{10} = \frac{m}{4} \left(\mathbf{d}'_- R_{1,0} \mathbf{d}_+ - 3\mathbf{d}'_+ R_{1,0} \mathbf{d}_- + \mathbf{d}'_- R_{\pi-\theta_{10}} \mathbf{d}_+ - 3\mathbf{d}'_+ R_{\pi-\theta_{10}} \mathbf{d}_- \right).$$

Since

$$R_{\pi-\theta_{10}} = \begin{pmatrix} -\cos(\theta_{10}) & -\sin(\theta_{10}) \\ \sin(\theta_{10}) & -\cos(\theta_{10}) \end{pmatrix} = -R_{0,1}, \tag{5.15}$$

this may be rewritten as

$$M_{10} = m \left(\mathbf{d}'_- R_{1,0} \mathbf{d}_+ - \mathbf{d}'_+ R_{1,0} \mathbf{d}_- \right) = 2m(d^1_+ d^2_- - d^2_+ d^1_-) \sin(\theta_{10}), \tag{5.16}$$

where $d^j_\pm$ is the jth component of $d_\pm$.

Consider the diagram of a single bar shown in Figure 5.2. From the figure we have

$$\mathbf{d}_+ = \begin{pmatrix} \frac{l}{2} + \delta_x \\ \delta_y \end{pmatrix}, \qquad \mathbf{d}_- = \begin{pmatrix} -(\frac{l}{2} - \delta_x) \\ \delta_y \end{pmatrix}.$$

Thus

$$d^1_+ d^2_- - d^2_+ d^1_- = \left(\frac{l}{2} + \delta_x \right) (\delta_y) + (\delta_y) \left(\frac{l}{2} - \delta_x \right) = l\delta_y,$$

and so

$$M_{10} = 2ml\delta_y \sin(\theta_{10}). \tag{5.17}$$

The kinetic energy defines a Riemannian metric K on S given by

$$K(\theta_{10})(X, W) = X'\hat{M}(\theta_{10})W, \quad X, W \in T_{(\theta_0,\theta_{10})}S. \tag{5.18}$$

Each joint is equipped with an identical spring. Let the spring potential for each be given by $V_s(\theta_{i+1,i})$, $i = 0, 1, 2, 3$, with V_s twice continuously differentiable. The total potential energy is then

$$V(\theta_0, \theta_{10}) = 2(V_s(\theta_{10}) + V_s(\pi - \theta_{10})) \stackrel{\triangle}{=} V(\theta_{10}), \tag{5.19}$$

where the relations in (5.4) have been used to simplify the expression. We assume that the potential energy is such that there exists $\alpha \in \{0, \pi\}$ such that

$$\left.\frac{\partial V}{\partial \theta_{10}}\right|_\alpha = 0, \qquad \left.\frac{\partial^2 V}{\partial \theta_{10}^2}\right|_\alpha > 0, \tag{5.20}$$

and without loss of generality we take $V(\alpha) = 0$. The standard Lagrangian is then given by

$$L(\theta_{10}, \omega_{10}) = \frac{1}{2} \begin{pmatrix} \omega_0 & \omega_{10} \end{pmatrix} \hat{M}(\theta_{10}) \begin{pmatrix} \omega_0 \\ \omega_{10} \end{pmatrix} - V(\theta_{10}). \tag{5.21}$$

Consider now the following action Φ_g of the Lie group S^1 on S:

$$\Phi_g(\theta_0, \theta_{10}) = (\theta_0 + g, \theta_{10}). \tag{5.22}$$

The quadruple (S, K, V, S^1) is a simple mechanical system with symmetry, where the action of S^1 on S is given by (5.22). (For a definition and discussion of simple mechanical systems with symmetry, see [1].) Since the action is both free and proper, the reduced space is a manifold. Recall that we defined $S = S^1 \times (0, \pi)$; the reduced (or shape) space is then $Q = (0, \pi)$ with the coordinate θ_{10}.

In the language of the moving systems approach, Q is the configuration space and S is the ambient space. To slowly rotate the mechanism set $\theta_0 = \Omega t + \hat{\theta}_0$ for some fixed initial offset $\hat{\theta}_0$. (Note that θ_0 and $\theta_0 + 2\pi$ are identified.) The imposed motion on the four-bar system is captured by the parametrized family of embeddings from Q into S given by

$$m_t(\theta_{10}) = \begin{pmatrix} \Omega t + \hat{\theta}_0 \\ \theta_{10} \end{pmatrix}. \tag{5.23}$$

5.1. The nominal dynamics. To apply the method developed in section 4 the dynamics of the system in the absence of any imposed motion must be periodic. Consider the nominal Lagrangian for the four-bar mechanism, defined by setting $\Omega = 0$ in (5.21):

$$L_0(\theta_{10}, \omega_{10}) = \frac{M_{11}}{2}\omega_{10}^2 - V(\theta_{10}). \tag{5.24}$$

Applying the Legendre transform, the conjugate momentum is found to be

$$p_{10} = M_{11}\omega_{10},$$

and thus the Hamiltonian for the nominal system is

$$H_0(\theta_{10}, p_{10}) = \frac{p_{10}^2}{2M_{11}} + V(\theta_{10}). \tag{5.25}$$

The nominal dynamics is

$$\dot{\theta}_{10} = \frac{p_{10}}{M_{11}}, \quad \dot{p}_{10} = -\frac{\partial V}{\partial \theta_{10}}. \tag{5.26}$$

From (5.20) and (5.26), we see that there is an equilibrium point at $(\theta_{10} = \alpha, p_{10} = 0)$. The existence of periodic solutions near this equilibrium point can be assured by appealing to the following theorem by Weinstein.

THEOREM 5.1 (see [36]). *Consider $H : \mathbb{R}^{2n} \to \mathbb{R}$. If H is twice continuously differentiable near an equilibrium point z and the Hessian matrix at the equilibrium point is positive definite, then for sufficiently small ϵ any energy surface $H(z) = H(0) + \epsilon^2$ contains at least n periodic orbits of the associated Hamiltonian system.*

The Hessian matrix of the nominal system, evaluated at the equilibrium $(\alpha, 0)$, is

$$\left(\begin{array}{cc} \frac{\partial^2 H}{\partial \theta_{10}^2} & 0 \\ 0 & \frac{\partial^2 H}{\partial p_{10}^2} \end{array} \right) \Bigg|_{(\alpha,0)} = \left(\begin{array}{cc} \frac{\partial^2 V}{\partial \theta_{10}^2} & 0 \\ 0 & M_{11}^{-1} \end{array} \right) \Bigg|_{(\alpha,0)}. \tag{5.27}$$

M is positive definite by its construction, and so $M_{11} > 0$. Combining this with the assumption in (5.20), we see that the Hessian matrix of the Hamiltonian at the equilibrium point is positive definite, and therefore by Theorem (5.1) there is a periodic solution around the equilibrium if the energy is sufficiently small.

Since this is a one-degree-of-freedom system, it is integrable, and thus there exist action-angle coordinates (J, ψ) [6]. These coordinates will prove particularly convenient for evaluating both the Hannay–Berry phase and the higher-order corrections. Let $\Gamma(h)$ be the trajectory in phase space corresponding to the energy h. Then

$$J = \frac{1}{2\pi} \oint_{\Gamma(h)} p_{10} d\theta_{10}. \tag{5.28}$$

The trajectory $\Gamma(h)$, and thus the action, depends on the form of $V(\theta_{10})$. We can write in general

$$\begin{array}{ll} J = g_1(\theta_{10}, p_{10}), & \theta_{10} = f_1(J, \psi), \\ \psi = g_2(\theta_{10}, p_{10}), & p_{10} = f_2(J, \psi). \end{array} \tag{5.29}$$

For the remainder of this paper we will assume that the initial conditions are such that the periodic solutions of the nominal system are of small amplitude. The expansion of $V(\theta_{10})$ about the equilibrium point $(\alpha, 0)$ yields

$$V(\theta_{10}) = \frac{1}{2} \left. \frac{\partial^2 V}{\partial \theta_{10}^2} \right|_{\alpha} (\theta_{10} - \alpha)^2 + O((\theta_{10} - \alpha)^3). \tag{5.30}$$

In the small angle limit the potential is taken only to second order. Since the springs on each bar are identical, this is equivalent to taking

$$V_s(\theta_{10}) = \frac{k_s}{2} (\theta_{10} - \alpha_s)^2 \tag{5.31}$$

for the potential of each spring. Here $\alpha_s \in S^1$. From (5.19) the total potential is

$$V(\theta_{10}) = k_s \left[(\theta_{10} - \alpha_s)^2 + (\pi - \theta_{10} - \alpha_s)^2 \right]. \tag{5.32}$$

The equilibrium point α is given by setting the derivative of V with respect to θ_{10} to zero. This yields

$$0 = \left. \frac{\partial V}{\partial \theta_{10}} \right|_{\theta_{10}=\alpha} = 2k_s \left(2\alpha - \pi \right) \tag{5.33}$$

and thus $\alpha = \frac{\pi}{2}$. The second derivative of V with respect to θ_{10} is

$$\frac{\partial^2 V}{\partial \theta_{10}^2} = 4k_s \triangleq k. \tag{5.34}$$

With this choice of spring potential the nominal Hamiltonian is given by

$$H_0 = \frac{p_{10}^2}{2M_{11}} + \frac{k}{2}(\theta_{10} - \alpha)^2. \tag{5.35}$$

This is the Hamiltonian for a harmonic oscillator. From [6] the angle variable is the phase of the oscillation, and the action is

$$J = \frac{h}{\omega}, \tag{5.36}$$

where $\omega = \sqrt{\frac{k}{M_{11}}}$ is the frequency of oscillation and h is the energy corresponding to a given initial condition. From (5.35) and (5.36)

$$J = \frac{p_{10}^2 + kM_{11}(\theta_{10} - \alpha)^2}{2\sqrt{kM_{11}}}. \tag{5.37}$$

Therefore

$$\theta_{10} = \alpha + \left[\frac{2J}{\sqrt{kM_{11}}} \right]^{\frac{1}{2}} \cos\psi = f_1(J, \psi), \tag{5.38}$$

$$p_{10} = -\left[2J\sqrt{kM_{11}} \right]^{\frac{1}{2}} \sin\psi = f_2(J, \psi). \tag{5.39}$$

The use of (5.38), (5.39) in (5.35) yields

$$H_0(\theta_{10}, p_{10}) = H_0(J) = \sqrt{\frac{k}{M_{11}}} J, \tag{5.40}$$

leading to the following dynamics in the action-angle variables.

$$\dot{\psi} = \sqrt{\frac{k}{M_{11}}}, \quad \dot{J} = 0. \tag{5.41}$$

The average $\langle \cdot \rangle$ is then the average over one cycle of the angle variable ψ.

5.2. The Hannay–Berry phase of the four-bar linkage. Recall that the Hannay–Berry phase is the holonomy of the Cartan–Hannay–Berry connection and is determined by solving the Hamiltonian system associated to the averaged momentum function defined in (2.3). From (5.23), the velocity vector of the motion in S is

$$\frac{d}{dt}\left(m_t(\theta_{10})\right) = \begin{pmatrix} 0 \\ \omega_{10} \end{pmatrix} + \begin{pmatrix} \Omega \\ 0 \end{pmatrix},$$

and thus the tangent vector which must be projected to $T_{m_t(q)}m_t(Q)$ is

$$\mathcal{Z} \triangleq \mathcal{Z}_t(m_t(q(t))) = \begin{pmatrix} \Omega \\ 0 \end{pmatrix}. \tag{5.42}$$

The projection of $\mathcal{Z}$ to $T_{m_t(q)}m_t(Q)$ with respect to the kinetic energy metric on S is given by $\mathcal{Z}^T = \mathcal{Z} - \mathcal{Z}^\perp$, where $\mathcal{Z}^\perp$ satisfies the orthogonality condition

$$K(\theta_{10})(\mathcal{Z}^\perp, X) = 0 \quad \forall X \in T_{m_t(q)}m_t(Q).$$

Application of the orthogonality condition yields

$$\mathcal{Z}^\perp = \begin{pmatrix} \Omega \\ -\Omega\left[\frac{M_{10}(\theta_{10})+M_{11}}{M_{11}}\right] \end{pmatrix} \tag{5.43}$$

and thus

$$\mathcal{Z}^T = \begin{pmatrix} 0 \\ \Omega\left[\frac{M_{10}(\theta_{10})+M_{11}}{M_{11}}\right] \end{pmatrix}. \tag{5.44}$$

The pull-back of $\mathcal{Z}^T$ to T_qQ by $[Tm]^{-1}$ is given by projection onto the second factor,

$$Z \triangleq [Tm]^{-1}\,\mathcal{Z}^T = \Omega\left[\frac{M_{10}(\theta_{10}) + M_{11}}{M_{11}}\right]. \tag{5.45}$$

The function $\mathcal{P}(Z)$ is then (following (2.3))

$$\begin{aligned}\mathcal{P}(Z)(J,\psi) &= \Omega\left[\frac{M_{10}(\theta_{10}(J,\psi)) + M_{11}}{M_{11}}\right] p_{10}(J,\psi) \\ &= -\Omega\left[1 + \frac{2ml\delta_y}{M_{11}}\sin\left(\alpha + \left[\frac{2J}{\sqrt{kM_{11}}}\right]^{\frac{1}{2}}\cos\psi\right)\right]\left[2J\sqrt{kM_{11}}\right]^{\frac{1}{2}}\sin\psi, \end{aligned} \tag{5.46}$$

where we have expressed the function in terms of the action-angle coordinates given by (5.38), (5.39) and substituted for M_{10} using (5.17).

The flow of the nominal system induces an S^1 action on T^*Q, and the average with respect to this action is simply the average over one cycle of the angle coordinate ψ. Thus the Hamiltonian function defining the lift with respect to the Cartan–Hannay–Berry connection is

$$\begin{aligned}\langle P(Z)\rangle &= \frac{1}{2\pi}\int_0^{2\pi} P(Z)(J,\psi)d\psi \\ &= -\frac{\Omega\left[2J\sqrt{kM_{11}}\right]^{\frac{1}{2}}}{2\pi}\left[\int_0^{2\pi}\sin\psi d\psi + \int_0^{2\pi}\frac{2ml\delta_y}{M_{11}}\sin\left(\alpha + \left[\frac{2J}{\sqrt{kM_{11}}}\right]^{\frac{1}{2}}\cos\psi\right)\sin\psi d\psi\right] \\ &= -\frac{\Omega\sqrt{kM_{11}}}{2\pi}\left[\frac{2ml\delta_y}{M_{11}}\right]\cos\left(\alpha + \left[\frac{2J}{\sqrt{kM_{11}}}\right]^{\frac{1}{2}}\cos\psi\right)\Bigg|_0^{2\pi} = 0. \end{aligned} \tag{5.47}$$

Therefore, under the assumption of linear springs, the Hannay–Berry phase for the equal-sided spring-jointed four-bar mechanism is zero.

5.3. Nonadiabatic corrections. From the kinetic energy metric in (5.18) and the form of $\mathcal{Z}^{\perp}$ in (5.43) we have

$$\|\mathcal{Z}^{\perp}\|^2 = \Omega^2 \left[\frac{M_{11}^2 - M_{10}^2(J,\psi)}{M_{11}}\right]. \tag{5.48}$$

Using this and the form of $\mathcal{P}(Z)$ for the four-bar in (5.46), the Hamiltonian for the rotating four-bar may be written as

$$H(J,\psi) = H_0(J) + \Omega H_1(J,\psi) + \Omega^2 H_2(J,\psi), \tag{5.49}$$

where

$$H_0(J) = \sqrt{\frac{k}{M_{11}}}J, \tag{5.50}$$

$$H_1(J,\psi) = \left[\frac{b\sin\left(\alpha + \left[\frac{2J}{\sqrt{kM_{11}}}\right]^{\frac{1}{2}}\cos\psi\right) + M_{11}}{M_{11}}\right]\left[2J\sqrt{kM_{11}}\right]^{\frac{1}{2}}\sin\psi, \tag{5.51}$$

$$H_2(J,\psi) = -\frac{1}{2}\left[\frac{M_{11}^2 - b^2\sin^2\left(\alpha + \left[\frac{2J}{\sqrt{kM_{11}}}\right]^{\frac{1}{2}}\cos\psi\right)}{M_{11}}\right], \tag{5.52}$$

where we have defined the constant $b = 2ml\delta_y$ to ease the notation. To find the effect of the imposed rotation to second order we must first find the generating function for the change of coordinates bringing the system into first-order normal form. From Proposition 3.8 we have

$$\begin{aligned} G(J) &= \frac{1}{2\pi}\int_0^{2\pi} \psi\left(H_1(J,\psi) - \langle H_1(J,\psi)\rangle\right)d\psi \\ &= \frac{1}{2\pi}\int_0^{2\pi} \psi\left[\frac{b\sin\left(\alpha + \left[\frac{2J}{\sqrt{kM_{11}}}\right]^{\frac{1}{2}}\cos\psi\right) + M_{11}}{M_{11}}\right]\left[2J\sqrt{kM_{11}}\right]^{\frac{1}{2}}\sin\psi d\psi, \end{aligned} \tag{5.53}$$

where we have used (5.47). Define

$$c_1(J) = \left[\frac{2J}{\sqrt{kM_{11}}}\right]^{\frac{1}{2}}, \quad c_2(J) = \left[2J\sqrt{kM_{11}}\right]^{\frac{1}{2}}. \tag{5.54}$$

Then

$$G(J) = \frac{c_2(J)b}{2\pi M_{11}}\int_0^{2\pi}\psi\sin\psi\sin(\alpha + c_1(J)\cos\psi)d\psi + \frac{c_2(J)}{2\pi}\int_0^{2\pi}\psi\sin\psi d\psi. \tag{5.55}$$

A simple integration by parts shows that

$$\int_0^{2\pi}\psi\sin\psi d\psi = -2\pi. \tag{5.56}$$

Similarly

$$\begin{aligned}\int_0^{2\pi} \psi \sin\psi \sin(\alpha + c_1(J)\cos\psi) d\psi \\ &= \frac{2\pi}{c_1(J)} \cos(\alpha + c_1(J)) - \frac{1}{c_1(J)} \int_0^{2\pi} \cos(\alpha + c_1(J)\cos\psi) d\psi \\ &= \frac{2\pi}{c_1(J)} \left[\cos(\alpha + c_1(J)) - \cos(\alpha) J_0(c_1(J))\right], \end{aligned} \tag{5.57}$$

where J_0 is a Bessel function of the first kind. Therefore

$$G(J) = \frac{c_2(J) b}{c_1(J)} \left[\cos(\alpha + c_1(J)) - \cos(\alpha) J_0(c_1(J))\right] - c_2(J). \tag{5.58}$$

To determine the second-order terms in the truncated second-order normal form Hamiltonian for a moving system, as in (4.11), we must find the terms $\mathrm{ad}_G^2 H_0$ and $\mathrm{ad}_G H_1$. Using the canonical bracket (4.1), we have

$$\mathrm{ad}_G H_0 = \frac{\partial H_0}{\partial \psi}\frac{\partial G}{\partial J} - \frac{\partial H_0}{\partial J}\frac{\partial G}{\partial \psi} = 0, \tag{5.59}$$

since H_0 and G are each independent of ψ. Thus

$$\mathrm{ad}_G^2 H_0 = 0. \tag{5.60}$$

For the next term we have

$$\mathrm{ad}_G H_1 = \frac{\partial H_1}{\partial \psi}\frac{\partial G}{\partial J} - \frac{\partial H_1}{\partial J}\frac{\partial G}{\partial \psi} = \frac{\partial H_1}{\partial \psi}\frac{\partial G}{\partial J}. \tag{5.61}$$

From the form of G in (5.58) we have

$$\frac{\partial G}{\partial J} = \frac{1}{c_1(J)} \left(b\cos(\alpha) J_1(c_1(J)) - b\sin(\alpha + c_1(J)) - 1\right), \tag{5.62}$$

and from (5.51) we have

$$\begin{aligned}\frac{\partial H_1}{\partial \psi} = c_2(J) \Bigg[\cos\psi \left(\frac{b\sin(\alpha + c_1(J)\cos\psi) + M_{11}}{M_{11}}\right) \\ - \frac{b c_1(J)}{M_{11}} \sin^2\psi \cos(\alpha + c_1(J)\cos\psi)\Bigg]. \end{aligned} \tag{5.63}$$

Since the average $\langle\cdot\rangle$ is over the variable ψ, we simplify the notation by defining

$$A_1 = \frac{c_2(J)}{c_1(J) M_{11}} \left(b\cos\alpha J_1(c_1(J)) - b\sin(\alpha + c_1(J)) - 1\right). \tag{5.64}$$

With this definition we may write

$$\begin{aligned}\mathrm{ad}_G H_1 = A_1 \left[b\cos\psi \sin(\alpha + c_1(J)\cos\psi) + \cos\psi \right. \\ \left. - b c_1(J) \sin^2\psi \cos(\alpha + c_1(J)\cos\psi)\right]. \end{aligned} \tag{5.65}$$

We now calculate $\langle \mathrm{ad}_G H_1 \rangle$. Since the average of $\cos\psi$ over a full cycle of ψ is zero, we have

$$\langle \mathrm{ad}_G H_1 \rangle = A_1 b \langle \cos\psi \sin(\alpha + c_1(J)\cos\psi)\rangle - A_1 b c_1(J) \langle \sin^2\psi \cos(\alpha + c_1(J)\cos\psi)\rangle. \tag{5.66}$$

Using integration by parts, we have

$$\begin{aligned} \langle \sin^2\psi \cos(\alpha + c_1(J)\cos\psi)\rangle &= \frac{1}{2\pi}\int_0^{2\pi} \sin^2\psi \cos(\alpha + c_1(J)\cos\psi) d\psi \\ &= \frac{1}{2\pi c_1(J)}\int_0^{2\pi} \sin(\alpha + c_1(J)\cos\psi)\cos\psi d\psi \\ &= \frac{1}{c_1(J)}\langle \cos\psi \sin(\alpha + c_1(J)\cos\psi)\rangle, \end{aligned} \tag{5.67}$$

and using this result in (5.65) yields

$$\langle \mathrm{ad}_G H_1 \rangle = 0. \tag{5.68}$$

Finally consider

$$\begin{aligned} \langle H_2 \rangle &= -\frac{1}{2}\left\langle \frac{M_{11}^2 - b^2 \sin^2(\alpha + c_1(J)\cos\psi)}{M_{11}} \right\rangle \\ &= -\frac{M_{11}}{2} + \frac{b^2}{2M_{11}}\langle \sin^2(\alpha + c_1(J)\cos\psi)\rangle. \end{aligned} \tag{5.69}$$

Thus the second-order truncated normal-form Hamiltonian for the rotating four-bar is

$$\begin{aligned} H^{(2)} &= H_0 + \Omega^2 \langle H_2 \rangle \\ &= \sqrt{\frac{k}{M_{11}}} J + \frac{b^2\Omega^2}{2M_{11}} \left\langle \sin^2(\alpha + c_1(J)\cos\psi)\right\rangle, \end{aligned} \tag{5.70}$$

where a constant term has been dropped from the Hamiltonian. Since the system is now in second-order normal form, we have for the initial conditions $(\bar{J}, \bar{\psi})$

$$\phi_t^{H^{(2)}}(\bar{J}, \bar{\psi}) = \phi_t^{\Omega^2\langle H_2\rangle} \circ \phi_t^{H_0}(\bar{J}, \bar{\psi}). \tag{5.71}$$

From (5.41) the flow map for the nominal dynamics is

$$\phi_t^{H_0}(\bar{J}, \bar{\psi}) = \left(\bar{J}, \sqrt{\frac{k}{M_{11}}} t + \bar{\psi}\right). \tag{5.72}$$

To determine the flow map for the second-order correction we need to solve the Hamiltonian vector field of $\Omega^2\langle H_2\rangle$. We have

$$\begin{aligned} \dot{\psi} &= \Omega^2 \frac{\partial}{\partial J}\left\langle \sin^2(\alpha + c_1(J)\cos\psi)\right\rangle \\ &= \Omega^2 \left[\frac{2}{J\sqrt{kM_{11}}}\right]^{\frac{1}{2}} \left\langle \sin(\alpha + c_1(J)\cos\psi)\cos(\alpha + c_1(J)\cos\psi)\right\rangle, \end{aligned} \tag{5.73}$$

$$\dot{J} = 0. \tag{5.74}$$

Calculating the average, we find

$$\begin{aligned}
&\langle \sin(\alpha + c_1(J)\cos(\psi))\cos(\alpha + c_1(J)\cos(\psi))\rangle \\
&\quad = \frac{1}{2\pi}\int_0^{2\pi} \sin(\alpha + c_1(J)\cos\psi)\cos(\alpha + c_1(J)\cos\psi)d\psi \\
&\quad = \frac{1}{2\pi}\left[\int_0^{\pi} \sin(\alpha + c_1(J)\cos\psi)\cos(\alpha + c_1(J)\cos\psi)d\psi \right. \\
&\qquad\qquad \left. + \int_{\pi}^{2\pi} \sin(\alpha + c_1(J)\cos\psi)\cos(\alpha + c_1(J)\cos\psi)d\psi\right] \\
&\quad = \frac{1}{2\pi}\left[\int_0^{\pi} \sin(\alpha + c_1(J)\cos\psi)\cos(\alpha + c_1(J)\cos\psi) \right. \\
&\qquad\qquad \left. + \sin(\alpha - c_1(J)\cos\psi)\cos(\alpha - c_1(J)\cos\psi)d\psi\right]. \qquad (5.75)
\end{aligned}$$

The use of the standard sum-angle formulas for sin and cos together with an expansion of the products in (5.75) yields

$$\begin{aligned}
&\frac{1}{2\pi}\left[\int_0^{\pi} \left[\left(\sin\alpha\cos\alpha\cos^2(c_1(J)\cos\psi) - \sin^2\alpha\cos(c_1(J)\cos\psi)\sin(c_1(J)\cos\psi)\right.\right.\right. \\
&\qquad + \cos^2\alpha\cos(c_1(J)\cos\psi)\sin(c_1(J)\cos\psi) - \sin\alpha\cos\alpha\sin^2(c_1(J)\cos\psi)\big) \\
&\qquad + \big(\sin\alpha\cos\alpha\cos^2(c_1(J)\cos\psi) + \sin^2\alpha\cos(c_1(J)\cos\psi)\sin(c_1(J)\cos\psi) \\
&\qquad\quad \left.\left.\left. - \cos^2\alpha\cos(c_1(J)\cos\psi)\sin(c_1(J)\cos\psi) - \cos\alpha\sin\alpha\sin^2(c_1(J)\cos\psi)\big)\right] d\psi\right] \\
&\quad = \frac{\sin\alpha\cos\alpha}{\pi}\int_0^{\pi}\left(\cos^2(c_1(J)\cos\psi) - \sin^2(c_1(J)\cos\psi)\right)d\psi \\
&\quad = \frac{\sin\alpha\cos\alpha}{\pi}\int_0^{\pi}\left(2\cos^2(c_1(J)\cos\psi) - 1\right)d\psi \\
&\quad = \frac{\sin\alpha\cos\alpha}{\pi}\int_0^{\pi}\cos(2c_1(J)\cos\psi)d\psi \\
&\quad = \sin\alpha\cos\alpha J_0(2c_1(J)), \qquad (5.76)
\end{aligned}$$

and thus

$$\dot{\psi} = \frac{\Omega^2 b^2}{2M_{11}}\left[\frac{2}{J\sqrt{kM_{11}}}\right]^{\frac{1}{2}}\sin\alpha\cos\alpha J_0\left(2c_1(J)\right). \qquad (5.77)$$

Recall now that the equilibrium point is $\alpha = \frac{\pi}{2}$. Insertion of this value into (5.77) yields

$$\dot{\psi} = 0. \qquad (5.78)$$

The second-order correction is therefore the identity map, and we have

$$\phi_t^{H^{(2)}}(\bar{J},\bar{\psi}) = \phi_t^{H_0}(\bar{J},\bar{\psi}). \qquad (5.79)$$

Thus to second order the effect of a slow imposed rotation on the four-bar is zero.

6. Conclusions. The approach developed by Marsden, Montgomery, and Ratiu provides a unified setting for understanding the role of the Coriolis force in moving systems. In this work we have extended the method through the use of Hamiltonian normal form theory to account for the nonadiabatic nature of the imposed motion. In particular we have shown that the Hannay–Berry phase can be viewed as arising from a first-order normal form approximation to the moving system. The moving systems approach is then naturally understood as a perturbation approach in the rate of the imposed motion. More accurate models can be determined by carrying the perturbation series out to higher orders. The method was applied to a rotating free-floating spring-jointed equal-sided four-bar mechanism, and it was shown that the effect of the imposed motion is zero to second order.

It is important to note that the approach developed here is perturbative, and it therefore relies on the assumption that the imposed motion, while not adiabatic, is slow with respect to the nominal dynamics. The resulting correction terms are second-order in the rate of the imposed motion, and the technique is thus useful only when extremely accurate modeling of the moving system is required. In addition, the resulting approximation is valid only on the time scale $O(\frac{1}{\epsilon^2})$, where ϵ is the rate of the imposed motion. See [21] for more detailed comments along these lines.

Acknowledgment. The author gratefully acknowledges P. S. Krishnaprasad for fruitful discussion, critical comments, and invaluable aid during the course of this work.

REFERENCES

[1] R. Abraham and J. Marsden, *Foundations of Mechanics*, 2nd ed., Perseus Press, New York, 1985.
[2] A. Aczel, *Pendulum. Léon Foucault and the Triumph of Science*, Atria Books, New York, 2003.
[3] S. Andersson, *Geometric Phases in Sensing and Control*, Ph.D. thesis, University of Maryland, College Park, MD, 2003.
[4] S. Andersson and P. Krishnaprasad, *The Berry–Hannay phase of the equal-sided spring-jointed four-bar mechanism*, in Proceedings of the 40th IEEE Conference on Decision and Control, 2001, IEEE Press, Piscataway, NJ, pp. 3406–3407.
[5] V. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1988.
[6] V. Arnold, *Mathematical Methods of Classical Mechanics*, 2nd ed., Springer-Verlag, New York, 1989.
[7] M. Berry, *Quantal phase factors accompanying adiabatic changes*, Proc. Roy. Soc. London, A, 392 (1984), pp. 45–57.
[8] M. Berry, *Quantum phase corrections from adiabatic iteration*, Proc. Roy. Soc. London, A, 414 (1987), pp. 31–46.
[9] A. Bhattacharjee and T. Sen, *Geometric angles in cyclic evolutions of a classical system*, Phys. Rev. A, 38 (1988), pp. 4389–4394.
[10] G. Bryan, *On the beats in the vibrations of a revolving cylinder or bell*, Proc. Cambridge Philos. Soc., 7 (1890), pp. 101–111.
[11] E. Cartan, *Lécons sur la Méthode de la Répère Mobile*, Gauthier-Villars, Paris, 1952.
[12] J. Cary, *Lie transform perturbation theory for Hamiltonian systems*, Phys. Rep., 79 (1981), pp. 130–159.
[13] R. Cushman, *Normal form for Hamiltonian vectorfields with periodic flow*, in Differential Geometric Methods in Mathematical Physics, S. Sternberg, ed., Reidel, Boston, 1984, pp. 125–144.
[14] R. Cushman, *A survey of normalization techniques applied to perturbed Keplerian systems*, in Dynamics Reported, Expositions in Dynamical Systems, C. Jones, U. Kirchgraber, and H. Walther, eds., Springer-Verlag, New York, 1992, Vol. 1, pp. 54–112.
[15] P. de Sousa Gerbert, *A systematic derivative expansion of the adiabatic phase*, Ann. Phys., 189 (1989), pp. 155–173.

[16] R. Fox and P. Jung, *Quasiadiabatic time evolution, avoided level crossings, and Berry's phase*, Phys. Rev. A, 57 (1998), pp. 2339–2346.
[17] T. Frankel, *The Geometry of Physics*, Cambridge University Press, Cambridge, UK, 1997.
[18] S. Fujishima, T. Nakamura, and K. Fujimoto, *Piezoelectric vibratory gyroscope using flexural vibration of a triangular bar*, in Proceedings of the IEEE 45th Annual Symposium on Frequency Control, 1991, IEEE Press, Piscataway, NJ, pp. 261–265.
[19] F. Gaitan, *Berry's phase in the presence of a non-adiabatic environment with an application to magnetic resonance*, J. Magnetic Resonance, 139 (1999), pp. 152–164.
[20] I. Gjaja and A. Bhattacharjee, *Divergences in the iterative and perturbative methods for computing Hannay's angle*, Phys. Rev. A, 41 (1990), pp. 5650–5665.
[21] S. Golin and S. Marmi, *A class of systems with measurable Hannay angles*, Nonlinearity, 3 (1990), pp. 507–518.
[22] F. Grashof, *Theretische Maschinenlehre*, Verlag L. Voss, Leipzig, Germany, 1883.
[23] J. Henrard, *The adiabatic invariant in classical mechanics*, in Dynamics Reported, Springer-Verlag, New York, 1993, pp. 117–235.
[24] A. Lawrence, *Modern Inertial Technology*, Springer-Verlag, New York, 1993.
[25] J. Marsden, R. Montgomery, and T. Ratiu, *Reduction, symmetry, and phases in mechanics*, Mem. Amer. Math. Soc., 88 (1990), pp. 248–260.
[26] J. Marsden and T. Ratiu, *Introduction to Mechanics and Symmetry*, 2nd ed., Springer-Verlag, New York, 1999.
[27] G. Newton, *Theory and practice in vibratory rate gyroscopes*, Control Engineering, (1963), pp. 95–99.
[28] K. Nomizu, *Lie Groups and Differential Geometry*, The Mathematical Society of Japan, Tokyo, 1956.
[29] B. Paul, *Kinematics and Dynamics of Planar Machinery*, Prentice–Hall, Englewood Cliffs, NJ, 1979.
[30] M. Putty, *A Micromachined Vibrating Ring Gyroscope*, Ph.D. thesis, Department of Electrical Engineering, University of Michigan, Ann Arbor, MI, 1995.
[31] A. Shapere and F. Wilczek, eds., *Geometric Phases in Physics*, World Scientific, River Edge, NJ, 1989.
[32] N. Sreenath, *Modeling and Control of Multibody Systems*, Ph.D. thesis, University of Maryland, College Park, MD, 1987.
[33] C. Sun, *High-order quantum adiabatic approximation and Berry's phase factor*, J. Phys. A, 21 (1988), pp. 1595–1599.
[34] C. Sun, *High-order adiabatic approximations related to non-Abelian Berry's phase factors and nuclear quadropole resonance*, Phys. Rev. D, 41 (1990), pp. 1318–1323.
[35] A. Tomita and R. Chiao, *Observation of Berry's topological phase by use of an optical fiber*, Phys. Rev. Lett., 57 (1986), pp. 937–940.
[36] A. Weinstein, *Normal modes for nonlinear Hamiltonian systems*, Invent. Math., 20 (1973), pp. 47–57.
[37] R. Yang and P. Krishnaprasad, *On the geometry and dynamics of floating four-bar linkages*, Dyn. Stab. Syst., 9 (1994), pp. 19–45.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 122–152

CONVERGENCE OF THE GENERALIZED VOLUME AVERAGING METHOD ON A CONVECTION-DIFFUSION PROBLEM: A SPECTRAL PERSPECTIVE*

C. PIERRE†, F. PLOURABOUÉ‡, AND M. QUINTARD‡

Abstract. This paper proposes a thorough investigation of the convergence of the volume averaging method described by Whitaker [*The Method of Volume Averaging*, Kluwer Academic, Norwell, MA, 1999] as applied to convection-diffusion problems inside a cylinder. A spectral description of volume averaging brings to the fore new perspectives about the mathematical analysis of those approximations. This spectral point of view is complementary with the Lyapunov–Schmidt reduction technique and provides a precise framework for investigating convergence. It is shown for convection-diffusion inside a cylinder that the spectral convergence of the volume averaged description depends on the chosen averaging operator, as well as on the boundary conditions. A remarkable result states that only part of the eigenmodes among the infinite discrete spectrum of the full solution can be captured by averaging methods. This leads to a general convergence theorem (which was already examined with the use of the center manifold theorem [G. N. Mercer and A. J. Roberts, *SIAM J. Appl. Math.*, 50 (1990), pp. 1547–1565] and investigated with Lyapunov–Schmidt reduction techniques [S. Chakraborty and V. Balakotaiah, *Chem. Engrg. Sci.*, 57 (2002), pp. 2545–2564] in similar contexts). Moreover, a necessary and sufficient condition for an eigenvalue to be captured is given. We then investigate specific averaging operators, the convergence of which is found to be exponential.

Key words. volume averaging, homogenization, convection, diffusion, Sturm–Liouville, spectral theory, Picard's successive approximation method, spectral methods

AMS subject classifications. 74Q, 76M50, 76M22, 78M40, 35B30

DOI. 10.1137/040610015

1. Introduction.

Volume averaging techniques are widely used to model transport problems for which decoupled or separated scales can be identified. The first part of this introduction deals with the potential interest of volume averaging for convection-diffusion problems in different applications. In the second part we discuss the interest and the specificity of volume averaging as compared to other homogenization methods. This general discussion is developed in the paper in a specific case suitable for mathematical treatment: the problem of convection-diffusion in a circular tube.

Convection-diffusion inside a tube would seem to be a simple mathematical problem. It turns out that it is a nontrivial problem, well known in the history of applied mathematics. Starting from Graetz [17] and Lévêque [20] in the stationary case, it has more lately interested Taylor [38] and Aris [1] in the context of its transient nonstationary asymptotic behavior. These seminal works have inspired many others, some of which are discussed in the second part of the introduction when discussing the methodological point of view.

Many research areas such as chemical engineering, biomechanics, and porous media are interested by variants of such a simple generic convection-diffusion problem. For example, when the considered problem involves many tubes inside which con-

*Received by the editors June 15, 2004; accepted for publication (in revised form) April 11, 2005; published electronically October 17, 2005.
http://www.siam.org/journals/siap/66-1/61001.html
†Université de Nantes Laboratoire de mathématiques Jean Leray, UMR CNRS 6629, 44035 Nantesceder, France (Pierre.Charles@math.univ-nantes.fr).
‡IMFT UMR CNRS-INPT/UPS, 5502 Av. du Professeur Camille Soula, 31400 Toulouse, France (plourab@imft.fr, quintard@imft.fr).

vection occurs (such as heat exchangers or microvascular beds), transport equations have been sought in terms of cross-section averaged fields [7, 19, 26, 27, 41]. Recently, the design and optimization of microheater exchangers has stimulated the search for averaged equations governing averaged temperature either at the tube scale or at the scale of the whole exchanger [26, 41]. In the context of heat exchange in biological tissues, averaged descriptions have remained very useful models [28] since the pioneering Pennes model [2, 3, 30]. These investigations suggest that averaged temperature associated with "compartmental" domains such as tissues and blood flow in vessels are interesting quantities to consider in order to model heat exchanges inside bodies. In these cases it is crucial to understand how the microscale flow may be approximated by averaged models because, even if possible, a detailed description of the full stationary problem at the local scale of each tube is not of great interest. In the context of these applications, averaged models have proved to be useful and interesting for applied concerns. Nevertheless, even if the description of averaged quantities is useful in practice for obvious operational reasons, there are still numerous questions concerning the validity and the quality of the approximation given by these ad hoc models. As a matter of fact, even if the model predictions could be *in retrospect* tested numerically, it is always interesting to better understand what their mathematical foundations are. This allows one to better understand their limits and their possible extensions. In this paper, we investigate the model of stationary convection-diffusion inside a tube. This study shows that, in this particular context, an averaged description can capture only large scale features of the exact solution, the convergence of which can be made as precise as necessary.

From a methodological point of view, spatial averaging is at first used as an operational definition of macroscale quantities. From this, macroscale equations may be derived, and the reader is referred to the paper [13] for a review of the different perspectives and points of view. For example, macroscale equations are introduced by many authors from extensive use of irreversible thermodynamics [18] (this approach is also often called mixture theory). In this paper, we are interested in methods that provide a direct, deterministic link, through some *mapping variables*, between the microscale and the macroscale fields. Such a method has been applied to determine macroscale transport equations for porous media applications, as illustrated in [39], while concomitantly a very similar approach has been proposed by Brenner [9]. Many characteristics and assumptions of the cited methods are close to other macroscopization methods, such as homogenization theory [6, 36]. Indeed, the general agreement between both methods has been described for diffusion problems in [8]. The major features may be summarized in the following terms:

- The macroscopic characteristic scales are supposed to be decoupled from the microscopic ones, each level having its own variable description.
- The initial boundary value problem (IBVP) that determines the microscale fields is solved in an approximated manner in terms of the macroscale variables and some mapping variables. The approximation is materialized by microscale problems or *closure problems* that completely define these mapping variables.
- Having solved these microscale problems, the macroscopic mathematical description is essentially dependent on the estimation of macroscopic coefficients or *effective* coefficients that are explicitly given in terms of averages of the mapping variables.

One feature of the considered volume averaging method is, therefore, that some additional hypotheses are needed in order to simplify the original problem and relate

the microscale fields to the macroscale ones. These additional relations, which we called "closure relations," are problem dependent and must be consistent with the assumption made of separated scales. This feature is common to almost all homogenization methods. For example, asymptotic methods are based on regular asymptotic expansions for inner (microscale) and outer (macroscale) variables to be specified, the scaling of which has to be carefully evaluated by order of magnitude analysis of the relevant parameters [23]. Another method involving scales is the time-scale separation between master and slave modes based on center manifold description [33]. This method has been used to provide a general and rigorous treatment of Taylor dispersion [4, 5, 10, 21, 24, 31, 34, 40]. This method shares many features with the one examined in this paper, besides a more general background and different objectives. One important starting point for this method is to use steady state solutions as decomposed into a discrete and infinite set of eigenfunctions. Examining a linear problem, the temporal solution are then linearly decomposed into those stationary eigenmodes, i.e., each stationary eigenmode is associated with a nonstationary one. Among those, the one associated with the trivial zero eigenvalue is called the master mode because it is associated with slow temporal relaxations of interest for long-time asymptotic behavior. The other temporal modes fulfill fast temporal relaxations whose influence on the master mode can be recast into the master equation parameters. The coupling between slave and master modes is obtained from a linear decomposition strictly similar to the above-mentioned "closure relations." These closure relations are derived from a Lyapunov–Schmidt reduction [4, 5] associated with a small parameter which is the product between the Péclet number and the aspect ratio of the considered tube.

The general philosophy of this master/slave time separation method is then much similar to the one applied in this paper on the spatial level. In the case of volume averaged methods, far-field spatial asymptotic behavior (sometimes called "fully developed" spatial variations) is interesting in that it describes the evolution of a simple one-dimensional macroscopic field, without requiring of a precise description of supplementary spatial variations. There is nevertheless one major technical difference with the goal pursued in this paper. In the case of the master/slave time separation method, the invariant manifold theorem gives a nice framework for the validity of such slow/fast mode decomposition close to any trivial zero eigenvalue [24] (because the time scales separation is governed by the ratio of the fast to slow mode eigenvalues). This framework can be easily transposed for spatially decaying modes close to a trivial zero eigenvalue [4]. Those zero eigenvalue macroscopic modes might be interesting, especially when the problem has Neumann boundary conditions. In this case, direct Lyapunov–Schmidt reduction techniques have been used to assess the convergence of averaging models, for example, when chemical reaction occurs within the fluid [11]. In section 5.1.1 we will compare our results with those obtained in [11] that are rediscussed in the third section of [5]. Those zero eigenvalue macroscopic modes are nevertheless less interesting in transfer problems. In that case they are associated with a spatially uniform eigenmode whose contribution to the transfer between the tube wall and the fluid is zero. Other nontrivial spatially decaying eigenmodes should then be sought. This is especially true when boundary conditions are not of Neumann type, so that there is no trivial zero eigenmode. But, in this case, the invariant manifold theorem hardly guarantees the validity and accuracy of a slow/fast scale decoupling. One of the purposes of this paper is to reexamine the conditions for which a macro/micro decoupling is a sensible approach in the case of a simple convection-diffusion problem with general boundary conditions.

In this sense, the presented analysis extends previous works [5, 11] which have used Lyapunov–Schmidt reduction techniques close to a 0 eigenmode. Our analysis considers the approximation of nonzero eigenmodes with non-self-adjoint operators. While restricted to a given convection-diffusion problem, this paper examines the precise conditions for which a part of the exact solution can be captured by an averaged model. One important conclusion, for applications purposes, that is drawn from the proposed analysis is that, depending on the chosen averaging method (more precisely depending on the applied weighting function), the nontrivial, interesting eigenmodes cannot always be captured. It is, therefore, of great interest to know better what causes averaging for convection-diffusion problems to work and why.

Moreover, there is an additional interest in our analysis for those willing to use averaged models. Macroscale equations, as generally introduced in the literature [39], come from first order terms. The "quality" of the first order approximation is often checked through some comparison with direct simulations, or analytical solutions of the microscale equations, or by developing estimates for the higher order terms. It is often difficult to have a precise quantitative determination of those terms, and the first approach, if available, can offer valuable information. In a preliminary study of the tube problem, it was found that the approximation proposed by [32] would provide a reasonable estimate of the exchange term for the established regime in the case of diffusion/advection in a tube with constant temperature or concentration at the surface [16]. The objective of this paper is to exhibit a higher order analysis of the problem from which convergence proofs can be obtained so that a posteriori conditions are found for the definition of the macroscopic scale.

The paper is organized as follows. The second section reviews a convection-diffusion problem in the stationary case and describes its known solutions. A short review of the results obtained with the volume averaging method is also presented in this section to further document the general context of the study. The third section presents a generalization of the volume averaging method previously used to describe temporal variations [4, 5]. This leads to a precise formulation of the mathematical convergence to any eigenmode. The fourth section presents the convergence proof in a two-step procedure. Some numerical results associated with the the convergence of different averaging operators are presented at the end of this section.

2. General background.

2.1. Convection-diffusion problem. The material exposed in this section closely follows classical steps that may be found in textbooks; see, for instance, [14]. We first present the dimensionless formulation associated with convection-diffusion of a passive scalar inside a cylinder with radial coordinate r made dimensionless by the tube radius R. This passive scalar could be associated, for instance, with some heat or mass transfer problem, and we will refer to it as $T(r, \phi, z)$. Classically, the ratio of convection to diffusion characteristic times is associated with a dimensionless Péclet number $\mathrm{Pe} = \langle v \rangle R/D_m$, where D_m is the diffusion coefficient of the passive tracer in the liquid and $\langle v \rangle$ is the spatially averaged velocity field. The physical problem giving the convection velocity is supposed to be independent of the passive scalar, so that a translation-invariant fully developed flow $v(r)$ settles in the longitudinal direction z along the cylinder principal axis. Making dimensionless the longitudinal direction z by the tube radius R, the stationary governing equation expressing heat- or mass-conservation of the passive scalar $T(r, z)$ reads

$$\Delta T = Pe v(r) \partial_z T \quad \text{with} \quad v(r) \geq 0 \quad \text{analytical in 0}, \tag{2.1}$$

where Δ stands for the Laplace operator, which will be appropriately expressed in cylindrical coordinates. As discussed later, we will be mainly interested in the situation where $Pe \gg 1$. Nevertheless, it is important to note that other definitions of the dimensionless variable in the z direction could be adopted. As a matter of fact, the typical longitudinal variations are linearly increasing with the Péclet number when $Pe \gg 1$, and, furthermore, the longitudinal dimensions of the tube could be much larger than its radius. Hence, many authors [4, 5] prefer to introduce an additional parameter $pe = PeR/L$, where L is some longitudinal characteristic length associated with the axial variations. In this context, many studies such as the classical ones [1, 38] have been interested in the limit of $pe \ll 1$, while $Pe \gg 1$ so that longitudinal diffusion can be neglected in comparison with transverse diffusion. This choice is important when considering the averaged description of (2.1), which should then be written with a small parameter pe instead of a large parameter Pe on the right-hand side. In the following, we will keep using the Péclet number Pe parameter for the problem. Of course, this choice should give equivalent results as those obtained from the use of the small parameter pe, as will be explained in section 5.1.1.

In the case of a Newtonian fluid, the velocity field develops a parabolic Poiseuille flow $v(r) = 2(1 - r^2)$. Because of its particular importance, all the numerical results will be given in this case. However, all the theoretical results obtained in this paper still hold for general nonnegative velocity fields $v(r) \geq 0$ that are analytical in 0. General velocity profiles are of interest for applications associated with non-Newtonian fluid, such as, for example, blood, for which different analytical models have been proposed for the velocity profile in a tube [15]. This can also be useful in the treatment of turbulent dispersion in tubes, for which the Poiseuille solution is replaced by the turbulent average velocity field, following the double averaging procedure in Pedras and Lemos [29].

Because of its relevance to many research areas, this partial differential problem has received much attention. Three basic classes of boundary conditions are naturally associated with this cylindrical geometry: adiabatic Neumann boundary condition $\partial_r T(r = 1, \phi, z) = 0$ (we shall refer to it as $\mathcal{N}$ in the following), constant temperature Dirichlet boundary condition $T(r = 1, \phi, z) = 0$ (we shall refer to it as $\mathcal{D}$ in the following), or mixed Robin boundary condition $\partial_r T(r = 1, \phi, z) + \gamma T(r = 1, \phi, z) = 0$, where $\gamma > 0$ may be called Thiele modulus by reference to the case of heterogeneous reaction (we shall refer to it as $\mathcal{R}$ in the following). Furthermore, the passive scalar reference value is chosen so that, far away from the origin, it reaches its equilibrium state, $T(r, \infty) = 0$. The only missing boundary condition is the initial value of the scalar field at the cylinder origin $z = 0$, $T(r, 0) \equiv T_0(r)$, which has to be specified. It is easy to note that the PDE problem (2.1) is not tensorized, so that it does not independently factorize the radial coordinate r and the longitudinal one z. While very simple, the linear problem (2.1) does not have any explicit general solution. Hence, many authors have been interested in the special limit for which a variable separation can be found. In the limit of a large Péclet number, $Pe \gg 1$, when neglecting the longitudinal diffusion compared to the radial one, (2.1) degenerates to

$$\left(\Delta_c + \frac{1}{r^2}\partial_\phi^2\right) T = Pev(r)\partial_z T, \tag{2.2}$$

where Δ_c stands for the cylindrical part of the Laplace operator $\Delta_c \equiv 1/r\partial_r(r\partial_r)$ and ϕ is the azimuthal angle. It can be shown that such an approximation is $O(1/Pe^2)$, because in this limit, the longitudinal typical variations scale linearly with Pe [14].

Equation (2.2) associated with either Neumann $\mathcal{N}$, Dirichlet $\mathcal{D}$, or Robin $\mathcal{R}$ boundary conditions is then a separable problem for which the PDE degenerates into a Sturm–Liouville ODE problem. Graetz [17] has found that its general solution is associated with the discrete sets L_N, $N \in \mathbb{Z}$, of eigenvalues depending on the boundary condition

$$T(r,\phi,z) = \sum_{N\in\mathbb{Z}} \sum_{l\in L_N} c_{N,l} G_{N,l}(r) e^{iN\phi} e^{\frac{l}{Pe}z}. \tag{2.3}$$

We define the generalized Graetz functions $G_{N,l}$ as the functions of r that satisfy

$$\begin{cases} \left(\Delta_c - \frac{N^2}{r^2}\right) G_{N,l} = l v(r) G_{N,l} \quad \text{with} \quad \mathcal{D} : G_{N,l}(1) = 0, \quad \mathcal{N} : \partial_r G_{N,l}(1) = 0, \\ \dfrac{G_{N,l}(r)}{r^N}(r=0) = 1, \qquad\qquad\qquad\qquad \mathcal{R} : \ G_{N,l}(1) + \gamma \partial_r G_{N,l}(1) = 0. \end{cases} \tag{2.4}$$

For a general—analytical in 0—velocity field $v(r)$ one can use the Frobenius method (cf., e.g., [35]) to see that the equation

$$\left(\Delta_c - \frac{N^2}{r^2}\right) y = l v(r) y,$$

which is singular in 0, has two linearly independent solutions y_1 and y_2; the first one regular in 0 satisfies $y_1(r)/r^N(r=0) \neq 0$, and the second one being singular in 0: $y_2(0) = \pm\infty$. As a result, (2.4) with initial condition $G_{N,l}/r^N(r=0) = 1$ defines a unique function $G_{N,l}$—which we will call the generalized Graetz function—for each $l \in \mathbb{C}$ and $N \in \mathbb{Z}$. Thus the following conditions in (2.4), $G_{N,l}(1) = 0$ for $\mathcal{D}$, $\partial_r G_{N,l}(1) = 0$ for $\mathcal{N}$, or $G_{N,l}(1) + \gamma \partial_r G_{N,l}(1) = 0$ for $\mathcal{R}$, only select among these generalized Graetz functions those satisfying the correct boundary condition.

Historically, the cylindrical Graetz functions $G_{0,l}$ has been associated with a parabolic Poiseuille flow $v(r) = 2(1-r^2)$ and it is usually found in the literature that the function $G_{0,l}$ is the eigenfunction of $\sqrt{-l}$ rather than l. However, this notation will be kept for the sake of simplicity in the rest of the paper, and Appendix A gives a more detailed discussion of Graetz eigenfunctions and their relations with confluent hypergeometric functions—or Kummer's functions.

Because (2.4) defines a self-adjoint Sturm–Liouville problem, the eigenvalues associated either with the Dirichlet, Neumann, or Robin conditions are real. Moreover, the chosen far-field extinction boundary condition $T(r,\infty) = 0$ selects, among those, negative eigenvalues. L_N is, therefore, a discrete set $L_N \subset \mathbb{R}^-$ of ordered eigenvalues $L_N = \{\cdots < l_{i,N} < \cdots < l_{1,N} < l_{0,N} \leq 0\}$. For convenience, we will use specific notation for the sets associated with Dirichlet, Neumann, or Robin boundary conditions, i.e.,

$$\begin{aligned} L_N^{\mathcal{D}} &= \{l \in \mathbb{R}^-, G_{N,l}(1) = 0\}, \qquad L_N^{\mathcal{N}} = \{l \in \mathbb{R}^-, \partial_r G_{N,l}(1) = 0\}, \\ &\text{or} \qquad L_N^{\mathcal{R}} = \{l \in \mathbb{R}^-, G_{N,l}(1) + \gamma \partial_r G_{N,l}(1) = 0\}. \end{aligned} \tag{2.5}$$

Graetz computed the first eigenvalue with two-digit precision in [17]. Tables 2.1 and 2.2 give the numerical estimates of the first three eigenvalues associated with a parabolic flow, Dirichlet and Neumann boundary conditions. More complete and precise computations of the eigenvalues can be found, for example, in [37]. Solution given by (2.3) can be completed by the orthogonality properties of the eigenfunctions,

$$\int_0^{2\pi} \int_0^1 G_{N,l}(r) e^{iN\phi} \overline{G_{N',l'}(r)} e^{-iN'\phi} v(r) r \, dr \, d\phi = 0 \qquad \text{if} \quad N \neq N' \ \text{ or } \ l \neq l', \tag{2.6}$$

TABLE 2.1
First three elements $(i = 0, 1, 2)$ *of sets* $L_N^{\mathcal{D}}$ *for Dirichlet boundary conditions,* $N = 0, 1, 2, 3$ *and a parabolic velocity field* $v(r) = 2(1 - r^2)$.

$l_{i,N}^{\mathcal{D}}$	$i = 0$	$i = 1$	$i = 2$
$N = 0$	-3.656793458	-22.30473055	-56.96051540
$N = 1$	-10.69115115	-37.38965286	-80.07477640
$N = 2$	-21.24944651	-56.05580310	-106.8036412
$N = 3$	-35.46611328	-78.38573690	-137.2070675

TABLE 2.2
First three elements $(i = 0, 1, 2)$ *of sets* $L_N^{\mathcal{N}}$ *for Neumann boundary conditions,* $N = 0, 1, 2, 3$ *and a parabolic velocity field* $v(r) = 2(1 - r^2)$.

$l_{i,N}^{\mathcal{N}}$	$i = 0$	$i = 1$	$i = 2$
$N = 0$	0	-12.8398060	-41.93087773
$N = 1$	-4.160532810	-25.33493287	-62.48391850
$N = 2$	-12.83980600	-41.93087773	-87.08337035
$N = 3$	-26.13743028	-62.80555035	-115.8424000

where the overbar denotes a complex conjugate. Hence, using (2.6), the constant coefficients $c_{N,l}$ in decomposition (2.3) are directly related to the projection of the initial conditions over its corresponding eigenfunction $G_{N,l}$:

$$c_{N,l} = \frac{\displaystyle\int_0^{2\pi}\int_0^1 T_0(r,\phi)\overline{G_{N,l}(r)}e^{-iN\phi}v(r)r\ dr\ d\phi}{\displaystyle 2\pi\int_0^1 |G_{N,l}(r)|^2 v(r)r\ dr}. \tag{2.7}$$

Hence, using the eigenfunctions defined in (2.4) the complete solution of the high Péclet limit of the convection-diffusion problem (2.2) within a tube admits a complete spectral representation. Incidentally, the convergence of this representation is known to be rather slow [37]. This is especially true when describing the solution near the origin $z = 0$. In this limit, even if (2.3) and (2.7) describe the true mathematical solution, the Lévêque [20] asymptotic expansion should be preferred because of its simplicity.

Nevertheless, this spectral representation is very useful when only part of the solution is required, as, for example, for the far-field behavior when $z > Pe/(l_1 - l_0)$ for which the solution exponentially converges to the first eigenfunction. In the following, we will concentrate on the first eigenfunctions and their associated eigenvalues. We will be furthermore interested in the averaged description of the solution. It should be noted that a uniform averaging along the disk section of the cylinder only keeps axisymmetrical modes. A more detailed discussion about nonasymmetrical contributions to the averaged description will be discussed in section 5.1.3. The amplitude decomposition (2.7) nevertheless shows that every axisymmetrical eigenvalue $l_{i,0}$ contributes to uniformly averaged concentration solution. This should be kept in mind in the following because many results associated with averaged descriptions in the literature have neglected contributions from eigenvalue $l_{i,0}$, with $i \geq 1$. In the following, we will, for example, see (what is already obvious from directly averaging solutions (2.3) and (2.7), which lead to no contribution of $l \neq 0$ modes for which $\langle v(r)G_{N,l}\rangle = 0$)

that a uniform averaging does not permit one to capture any decaying eigenvalue associated with the Neumann boundary conditions.

2.2. Weighted volume averaging method. In this section we present an improved version of the volume averaging method introduced in [39] that nevertheless remains closely related to this first method—which we will call the *standard volume averaging method*. The improvement is based on the introduction of weighted averaging operators as proposed in [16] when the standard volume averaging method considers only averaging associated with the Lebesgue measure. The use of weighted averages had been considered long ago for averaging transport equations [12, 13, 22, 25]. The intentions were to correctly regularize the microscale fields with the objective of improving comparison with experiments. It is interesting to note that this paper emphasizes another important and fundamental role of weighted averages more related to the mathematical structure of the operator to be averaged.

2.2.1. Definition and notation. To introduce general weighted averaging operators we first introduce the standard averaging operator $\langle\ \rangle$ corresponding to the Lebesgue measure on each cylinder section for functions with radial symmetry,

$$\langle T\rangle(\phi,z) = 2\int_0^1 T(r,\phi,z)r\,dr,$$

and we now define a general weighted averaging operator $\langle\ \rangle^\star$, sometimes simply denoted $\star$, associated with any normalized weight function $w(r)$—i.e., such that $\langle w\rangle = 1$—in cylindrical coordinates as

$$\langle T\rangle^\star(\phi,z) \equiv T^\star(\phi,z) = \langle Tw\rangle(\phi,z) = 2\int_0^1 T(r,\phi,z)w(r)r\,dr. \tag{2.8}$$

In the next sections of this paper we will examine special cases of weight function w. First, a uniform weight $w = 1$ is associated with the standard volume averaging method [39]. Another interesting case, introduced in the preceding section is "mixing-cup" averaging, where the weight function has a dependence exactly similar to that of the velocity field $w(r) \equiv v(r)/\langle v\rangle$. The resulting averaged temperature is also often called bulk temperature. As mentioned in the previous section, this weight function is interesting considered in this context precisely because it corresponds exactly to the orthogonalization operator associated with the Graetz eigenfunctions, as illustrated in (2.6). In the following, the averaging operator is either defined using a specific weight function yet to be specified w, or, on the contrary, to simplify the notation, a generic $\star$ is used for averaging (2.8). Now, averaging the theoretical solution (2.3) leads to

$$T^\star(\phi,z) = \sum_{N\in\mathbb{Z}}\sum_{l\in L_N} C_{N,l}e^{iN\phi}e^{\frac{l}{Pe}z} \quad \text{with} \quad C_{N,l} = c_{N,l}G^\star_{N,l} \in \mathbb{R}. \tag{2.9}$$

It should be noted that a supplementary average along the azimuthal direction ϕ could be performed. If uniform along ϕ, such an average will only preserve the axisymmetric mode $N = 0$ in (2.9). If the azimuthal averaging is chosen nonuniform along ϕ, then the averaged solution could have contributions from nonaxisymmetric mode $N \neq 0$. In the following, we will be mainly interested in averaging along the radial coordinate. Thus the macroscopic field depends on the azimuthal angle ϕ. The results that are presented for the convergence of averaging models will be discussed for

any azimuthal mode N. Those averaged models could easily been averaged a second time along ϕ to find longitudinally varying averaged equations as finally discussed in section 5.1.3.

As mentioned in the introduction, the volume averaging method is a general technique whose purpose is to find a macroscopic description, i.e., an averaged description of a microscopic field that fulfills some PDE problem, without explicitly solving the complete problem, but solving some simplified version of it. Greek letters will be reserved for quantities associated with the volume averaging predictions. Prediction for the scalar field T is thus denoted Θ. In general, the prediction is decomposed into a macroscopic volume averaging prediction $\Theta^\star$ and some local deviation θ to this macroscopic behavior:

$$\Theta(r,\phi,z) = \Theta^\star(\phi,z) + \theta(r,\phi,z) \ = \sum_{N\in\mathbb{Z}} \left(\Theta_N^\star(z) + \theta_N(r,z)\right) e^{iN\phi} \tag{2.10}$$

with the associated condition $\langle\theta\rangle^\star = 0$. In the upscaling techniques considered in this paper, the derivation is sought generally under the form of a mapping onto the macroscopic variables and derivatives. The averaged of the microscale equation will be discussed in detail later. This macroscale equation can be used to show that $\Theta^\star$ also decomposes into a sum of exponential modes:

$$\Theta^\star(\phi,z) = \sum_{N\in\mathbb{Z}} \Theta_N^\star(z) e^{iN\phi} \tag{2.11}$$

with

$$\Theta_N^\star(z) = \sum_{\lambda\in\Lambda_N} C_{N,\lambda} e^{\frac{\lambda}{Pe}z} \quad \text{with} \quad C_{N,\lambda} = c_{N,\lambda}\Gamma^\star_{N,\lambda} \in \mathbb{R}, \tag{2.12}$$

where the corresponding Greek letters have been used to describe the approximated discrete spectrum Λ_N and its corresponding approximated eigenvalues λ, as well as the corresponding approximated eigenfunction $\Gamma_{N,\lambda}$, approximating $G_{N,l}$ with an approximated amplitude $c_{N,\lambda}$ that will be more explicitly defined in section 4.

The main purpose of section 4 is to find from which conditions it is possible to find intersections between Λ_N and the eigenvalue set L_N (2.5) of the theoretical problem (2.2). It will be found in section 4.1 that *only a part of the spectrum L_N can be approximated by elements of Λ_N*. It will, furthermore, be shown in section 4.2 that elements of Λ_N converges toward these elements of L_N that can be approximated when increasing the order of the averaging method. The rate of convergence is consequently studied in section 4.3.

2.2.2. Weighted volume averaging technique. In this section we present the principal steps of the weighted volume averaging technique. The next section will a posteriori justify the classical assumptions made in this section, from examining the weighted volume averaging method generalized to higher order. We will study here both Neumann and Dirichlet Graetz problems. The case of Dirichlet boundary conditions associated with the Graetz problem has been previously examined in the context of the standard volume averaging technique in [16]. The first step of the method is to average the governing equation (2.2), so that $\langle 2.2\rangle^\star$ is

$$Pe\partial_z \langle v\Theta\rangle^\star = \langle \Delta_c\Theta\rangle^\star + \frac{1}{r^2}\langle\partial_\phi^2\Theta\rangle^\star = \langle \Delta_c\Theta\rangle^\star + \frac{1}{r^2}\partial_\phi^2\langle\Theta\rangle^\star. \tag{2.13}$$

The next step is to use decomposition (2.10) and (2.11) in (2.13), so that a macroscopic equation is defined for $\Theta_N^\star$:

$$\langle \Delta_c \Theta_N \rangle^\star - N^2 \left\langle \frac{\Theta_N}{r^2} \right\rangle^\star = Pe\partial_z \langle v\Theta_N \rangle^\star . \tag{2.14}$$

The completeness of this macroscopic equation necessitates the knowledge of deviation θ_N. The problem associated with the deviation θ_N is now obtained from subtracting (2.14) from (2.2):

$$(v - \langle v \rangle^\star) Pe\partial_z \Theta_N^\star + Pe\partial_z (v\theta_N - \langle v\theta_N \rangle^\star) = \mathcal{L}_N^\star \Theta_N, \tag{2.15}$$

where $\mathcal{L}_N^\star$ stands for the nonlocal differential operator:

$$\begin{aligned} \mathcal{L}_N^\star \Theta_N &= \Delta_N \Theta_N - \langle \Delta_N \Theta_N \rangle^\star, \\ \Delta_N \Theta_N &= \Delta_c \Theta_N - \frac{N^2}{r^2} \Theta_N, \\ \langle \Delta_N \Theta \rangle^\star &= \langle \Delta_c \Theta_N \rangle^\star - N^2 \Big\langle \frac{1}{r^2} \Theta_N \Big\rangle^\star . \end{aligned} \tag{2.16}$$

This operator is neither local nor self-adjoint. It is nevertheless invertible, as shown in Appendix C. The first term of (2.15) is a macroscopic source term that enters in the microscopic problem defined for deviation θ_N. So far, no hypothesis has been made and the above equations are exact. These equations are nevertheless not closed because the coupling between the deviation and the macroscopic field still remains unsolved. Finding this coupling is in fact exactly identical to solving the original problem (2.2), the resolution of which we precisely want to avoid.

Hence, the key step is then to find a suitable hypothesis to close deviation problem (2.14) so that it should depend only on the macroscopic field $\Theta_N^\star$. First, it should be kept in mind that the governing equation (2.2) is linear. As a consequence, it is obvious that the deviation θ_N dependence with the macroscopic field $\Theta_N^\star$ has to be linear here. Such a linear dependence is in fact very generally admitted in most of the application of the method [39] and comes from the assumption of scale separation. Hence, one writes the "closure hypothesis" by introducing the additional closure field or mapping variables $\alpha_{0,1}(r)$ which relates the deviation $\theta_N(r,z)$ to the macroscopic field $\Theta_N^\star(z)$,

$$\theta_N(r,z) = \big(w(r)\alpha_{0,N}(r) - 1\big)\Theta_N^\star(z) + w(r)\alpha_{1,N}(r) Pe\partial_z \Theta_N^\star(z),$$

or, equivalently,

$$\Theta_N(r,z) = \alpha_{0,N}(r)\Theta_N^\star(z) + \alpha_{1,N}(r) Pe\partial_z \Theta_N^\star(z). \tag{2.17}$$

It is clear that additional terms are required to obtain an exact solution, and it is our objective to understand *what has been kept* in such an approximate solution. Using the closure hypothesis (2.17) in (2.15) we obtain

$$\begin{aligned} &(\mathcal{L}_N^\star \alpha_{0,N})\Theta_N^\star + (\mathcal{L}_N^\star \alpha_{1,N} - v(r)\alpha_{0,N} + \langle v\alpha_{0,N} \rangle^\star) Pe\partial_z \Theta_N^\star \\ &\quad -(v(r)\alpha_{1,N} - \langle v\alpha_{1,N} \rangle^\star) Pe^2 \partial_z^2 \Theta_N^\star = 0. \end{aligned}$$

The condition of this equality is that each coefficient multiplying the macroscopic field variations $\Theta^\star, \partial_z\Theta^\star, \partial_z^2\Theta^\star$ is equal to zero. Nevertheless, (2.17) has introduced

a closure hypothesis with only two terms, so that the first two terms should also be considered here self-consistently. This last point is further discussed in the next section. Hence, problems associated with the closure fields $\alpha_{0,N}$ and $\alpha_{1,N}$ are

$$\text{(2.18)}\quad \begin{cases} \left(\mathcal{L}_N^\star \alpha_{0,N}\right)(r) = 0, \\ \alpha_{0,N}^\star = 1, \end{cases} \quad \text{and} \quad \begin{cases} \left(\mathcal{L}_N^\star \alpha_{1,N}\right)(r) = v(r)\alpha_{0,N}(r) - \langle v\alpha_{0,N}\rangle^\star, \\ \alpha_{1,N}^\star = 0 \end{cases}$$

$$\text{with} \quad \alpha_{\mathrm{i},N}(1) = 0 \text{ for } \mathcal{D}, \qquad \partial_r \alpha_{i,N}(1) = 0 \text{ for } \mathcal{N} \quad \text{or}$$

$$\alpha_{i,N}(1) + \gamma \partial_r \alpha_{i,N}(1) = 0 \text{ for } \mathcal{R}, \quad i = 1, 2.$$

These problems can be solved analytically for a Neumann, Dirichlet, or Robin boundary condition, and their resolution is detailed in Appendix C. When introducing these solutions in the macroscopic problem (2.14), one finds the macroscopic problem

$$\text{(2.19)}\qquad K_{0,N}\Theta_N^\star + K_{1,N} Pe \partial_z \Theta_N^\star - \langle v\alpha_{1,N}\rangle Pe^2 \partial_z^2 \Theta_N^\star = 0,$$

which involves the *effective parameters*

$$\text{(2.20)}\qquad K_{0,N} = \langle \Delta_N \alpha_{0,N}\rangle^\star, \quad K_{1,N} = \langle \Delta_N \alpha_{1,N}\rangle^\star - \langle v\alpha_{0,N}\rangle^\star,$$

and the solution for $\Theta_N^\star$ decomposes to a sum of exponential modes with an associated characteristic length Pe/λ, which then defines the set $\Lambda_{1,N}$ of eigenvalues predicted by the volume averaging technique

$$\text{(2.21)}\qquad \Lambda_{1,N} = \{\ \lambda \ / \quad K_{0,N} + K_{1,N}\lambda - \langle v\alpha_{1,N}\rangle \lambda^2 = 0\ \}.$$

2.2.3. Explicit results. This section gives the solutions of problem (2.18), i.e., the mapping variables, and (2.19) obtained for different values of the weighted function w.

- *Standard volume averaging*, $w = 1$, axisymmetric mode $N = 0$.

The solution for the closure function has been found equal to the following:

$$\text{(2.22)}\quad \text{for } \mathcal{D}: \begin{cases} \alpha_{0,0}(r) = 2(1 - r^2), \\ \alpha_{1,0}(r) = \frac{r^6}{9} - \frac{r^4}{2} + \frac{r^2}{2} - \frac{1}{9}; \end{cases} \quad \text{for } \mathcal{N}: \begin{cases} \alpha_{0,0}(r) = 1, \\ \alpha_{1,0}(r) = -\frac{r^4}{8} + \frac{r^2}{4} - \frac{1}{12}. \end{cases}$$

Thus constants $K_{0,0}$ and $K_{1,0}$ can be computed:

$$\text{(2.23)}\qquad \text{for } \mathcal{D}: \begin{cases} K_{0,0} = -16, \\ K_{1,0} = -2; \end{cases} \qquad \text{for } \mathcal{N}: \begin{cases} K_{0,0} = 0, \\ K_{1,0} = -1. \end{cases}$$

These calculations permit us to compute the associated approximated eigenvalues by solving (2.21). As already observed in [16], the resulting Dirichlet eigenvalue $\lambda_{0,0}^{\mathcal{D}} \simeq -3.874877690$ gives a rather good approximation of the Graetz value $l_{0,0}^{\mathcal{D}} \simeq -3.656793458$ up to 6%. On the contrary, the Neumann eigenvalue $l_1^N \simeq -12.839806$ is completely missed by the volume averaging method, which nevertheless gives the trivial solution zero, $l_{0,0}^{\mathcal{N}} = 0$. This trivial solution is of course of great practical interest since it corresponds to the exact solution when the temperature at the origin is constant; it also gives the correct averaged temperature of the far-field solution.

- *Flow averaging*, $w = v/\langle v\rangle = 2(1 - r^2)$, axisymmetric mode $N = 0$.

The solution for the closure function has been found equal to the following:

$$\text{(2.24)}\quad \text{for } \mathcal{D}: \begin{cases} \alpha_{0,0}(r) = \frac{3}{2}(1 - r^2), \\ \alpha_{1,0}(r) = \frac{r^6}{12} - \frac{3r^4}{8} + \frac{57r^2}{160} - \frac{31}{480}; \end{cases} \quad \text{for } \mathcal{N}: \begin{cases} \alpha_{0,0}(r) = 1, \\ \alpha_{1,0}(r) = -\frac{r^4}{8} + \frac{r^2}{4} - \frac{1}{16}. \end{cases}$$

Thus constants $K_{0,0}$, $K_{1,0}$ can be computed:

$$\text{for } \mathcal{D}: \begin{cases} K_{0,0} = -3, \\ K_{1,0} = -\frac{63}{40}; \end{cases} \quad \text{for } \mathcal{N}: \begin{cases} K_{0,0} = 0, \\ K_{1,0} = -1. \end{cases} \tag{2.25}$$

The approximate Dirichlet eigenvalue in this case is found equal to $\lambda^{\mathcal{D}}_{0,0} \simeq -3.809523810$, which is 4% from the theoretical Graetz eigenvalue $l^{\mathcal{D}}_{0,0}$. The Neumann trivial solution $\lambda^{\mathcal{N}}_{0,0} = 0$ is also found and the first Neumann nontrivial eigenvalue $\lambda^{\mathcal{N}}_{1,0}$ is also totally missed in the case of a flow averaging.

The following section investigates the capacity of the method to find the correct answer to the problem while generalizing it by introducing higher order closure hypothesis.

3. Weighted volume averaging method of higher order. The notation and methodological steps in this section are closely following those previously presented in sections 2.2.1 and 2.2.2. More precisely, the solution we are looking for is decomposed as (2.10), and the same exact steps (2.13)–(2.15) are now considered again.

The improvement of the method consists in a generalization of the closure hypothesis (2.17). This is introduced in order to ameliorate the results previously obtained in section 2.2.3, with, for instance, the hope to capture the first nontrivial Neumann eigenvalue $l^{\mathcal{N}}_{1,0}$.

From the property (4.2) of the exact solution that will be studied in section 4.1.1, and from the previously examined closure relation (2.17) let us now introduce a generalized closure relation:

$$\Theta_N(r,z) = \sum_{n=0}^{p} \alpha_{n,N}(r) Pe^n \partial_z^n \Theta^\star_N(z) \tag{3.1}$$

with $p \geq 1$. The case $p = 1$ has been analyzed in section 2.2.2, and we now follow the same steps. Using (3.1) in the deviation equation (2.15) it is found, assuming $\alpha_{-1,N}(r) = 0$, that

$$\begin{aligned} &\sum_{n=0}^{p} (\mathcal{L}^\star_N \alpha_{n,N} - v\alpha_{n-1,N} + \langle v\alpha_{n-1,N}\rangle^\star) Pe^n \partial_z^n \Theta^\star_N(z) \\ &\quad - (v\alpha_{p,N} - \langle v\alpha_{p,N}\rangle^\star) Pe^{p+1} \partial_z^{p+1} \Theta^\star_N(z) = 0. \end{aligned}$$

The condition of this equality gives, at each order, the closure problem associated with the closure functions $\alpha_{n,N}(r)$, whose solvability is left to Appendix C, and which is to be solved recursively:

$$\begin{cases} \mathcal{L}^\star_N \alpha_{n,N} = v(r)\alpha_{n-1,N}(r) - \langle v\alpha_{n-1,N}\rangle^\star & \text{with} \quad \alpha_{-1,N}(r) = 0, \\ \alpha^\star_{0,N} = 1 \quad \text{or} \quad \alpha^\star_{n,N} = 0 & \text{for} \quad n \geq 1, \\ \alpha_{n,N}(1) = 0 & \text{for} \quad \mathcal{D}, \\ \partial_r \alpha_{n,N}(1) = 0 & \text{for} \quad \mathcal{N}. \end{cases} \tag{3.2}$$

The resolution of these problems is detailed in Appendix C.2.

From solving (3.2) it is possible to find the generalized macroscopic closed problem at order p:

$$\sum_{n=0}^{p} K_{n,N} Pe^n \partial_z^n \Theta^\star_N(z) - \langle v\alpha_{p,N}\rangle^\star Pe^{p+1} \partial_z^{p+1} \Theta^\star_N(z) = 0, \tag{3.3}$$

where the macroscopic coefficients $K_{n,N}$ are given by

$$K_{n,N} = \langle \Delta_N \alpha_{n,N} \rangle^\star - \langle v \alpha_{n-1,N} \rangle^\star, \quad K_{n,N} \in \mathbb{R}. \tag{3.4}$$

The predicted solutions of (3.3) then decompose into a sum of exponentials with modes λ/Pe for λ belonging to the set of predicted eigenvalues at order p, $\Lambda_{p,N}$, defined as the zeros set of a $p+1$ order polynomial:

$$\Lambda_{p,N} = \left\{ \lambda \Big/ \sum_{n=0}^{p} K_{n,N} \lambda^n - \langle v \alpha_{p,N} \rangle^\star \lambda^{p+1} = 0 \right\}. \tag{3.5}$$

As previously, $\Lambda_{p,N}$ is independent of Pe, but does depend on the chosen boundary conditions and the order p of the closure relation. This last point naturally leads to the concept of convergence.

DEFINITION 3.1 (convergence of the weighted volume averaging method). *The elements of all sets $\Lambda_{p,N}$ define sequences of predicted eigenvalues: $(\lambda_{i,p})_{p\geq 1, i\geq 0}$, $\lambda_{i,p} \in \Lambda_{p,N}$.*

We shall say that the method is convergent toward some eigenvalue $l_i \in L_N$ of the theoretical problem (2.2) *if there exists a sequence of predicted eigenvalues $(\lambda_{i,p})_{p\geq 1, i\geq 0}$ such that $\lambda_{i,p} \in \Lambda_{p,N}$ and $\lim_{p\to+\infty} \lambda_{i,p} = l_i \in L_N$.*

We will establish the convergence for a characterized part of the spectrum in section 4.3.

4. Convergence analysis. Previous sections have mainly considered the explicit application of the averaging method to Graetz problem. The necessary material and notation being now defined, this section considers the mathematical analysis of the convergence of these averaging methods. This convergence analysis requires two different steps. The first step introduces two necessary conditions over eigenvalues for convergence to hold. The second step gives the proof that these two necessary conditions are sufficient. In the two subsequent sections, the results are derived in a general context, and formally apply to any mode N, as well as any boundary conditions $\mathcal{D}$, $\mathcal{N}$, or $\mathcal{R}$ and any flow $v(r)$. Hence, in order to simplify notation, the analysis does not mention, unless necessary to avoid confusion, which azimuthal mode it refers to, nor the boundary conditions that are considered. Finally, specific situations will be considered in section 4.3 for analyzing the numerical convergence.

4.1. Restricted convergence of weighted averaging methods. We define in the two following sections two sets, the validity—$D^\star_{val}$—and the accessibility—$D^\star_{acc}$—domains, which are disks lying in the complex plane $\mathbb{C}$. As we will see, a necessary condition for the weighted averaging method to converge toward an eigenvalue $l \in L$ is that this eigenvalue belongs to both of these domains.

4.1.1. Validity domain $D^\star_{val}$. The variables of the initial problem (2.2) can be separated so that any solution $T(r,\phi,z)$ may be written as a product of functions of r, ϕ, and z only. Let us first show in this section that the exact solution of the problem can be formally written as a regular asymptotic expansion of the macroscopic field $T^\star(\phi,z)$. First let us decompose T as

$$T(r,\phi,z) = \sum_{N\in\mathbb{Z}} T_N(r,z) e^{iN\phi}. \tag{4.1}$$

The aim of this section is to analyze under which condition the Nth component $T_N(r,z)$ in decomposition (4.1) of the theoretical solution $T(r,\phi,z)$ can be written as

the following expansion of $T_N^\star(z)$:

$$T_N(r,z) = \sum_{n\geq 0} a_n(r) Pe^n \partial_z^n T_N^\star(z) \tag{4.2}$$

(where the index N on the closure functions $a_n(r)$ has been omitted for simplicity) to be compared with the general closure hypothesis (3.1) for $\Theta_N(r,z)$.

Let us recall the form of the original solution (2.3),

$$T_N(r,z) = \sum_{l\in L_N} c_l T_l(r,z) \quad \text{with} \quad c_l \in \mathbb{R}\,, \quad T_l(r,z) = G_{N,l}(r) e^{\frac{l}{Pe}z}, \tag{4.3}$$

so that (4.2) is true for $T_N(r,z)$ if and only if it holds for each function $T_l(r,z)$ standing in the decomposition (4.3) of $T_N(r,z)$. Comparing then the expression for T_l in (4.2) and (4.3), one can see that (4.2) holds for $T_l(r,z)$ if and only if the following equality over the Graetz eigenfunctions G_l holds:

$$\sum_{n\geq 0} a_n(r) l^n = \frac{G_{N,l}(r)}{G_{N,l}^\star}. \tag{4.4}$$

We will prove that both functions $G_{N,l}(r)$ and $G_{N,l}^\star$ are analytical with respect to l, so that the expansion of $G_{N,l}(r)/G_{N,l}^\star$ on the form (4.4) is only possible for l belonging to a disk $D_{val}^\star$ centered on zero whose radius R is equal to the smallest root of $G_l^\star$.

Definition 4.1. *Let us call validity domain the disk* $D_{val}^\star \subset \mathbb{C}$,

$$D_{val}^\star = \big\{l, \quad |l| < R\big\}, \quad \textit{where} \quad R = \inf\big\{|l| \,/\, G_{N,l}^\star = 0\big\},$$

depending only on the averaging operator $\star$ *and on* N.

Now, one can see that the decomposition (4.2) is not true in general. It is true only if all the eigenvalues l standing in the decomposition (4.3) of $T_N(r,z)$ belong to the validity domain $D_{val}^\star$. An important consequence is that a closure formulation (3.1) only makes sense for eigenvalues lying in $D_{val}^\star$. Hence, a necessary—but not sufficient—condition for an eigenvalue $l \in L$ to be predicted by the averaging method is to lie within $D_{val}^\star$. It is also interesting to note that $D_{val}^\star$ depends only on the averaging operator $\star$ and N, but not on the boundary conditions.

We summarize this condition, as well as the definition of the new functions $a_n(r)$, in the following lemma.

Lemma 4.2. *The base functions* $T_l(r,z) = G_{N,l}(r) e^{\frac{l}{Pe}z}$ *for problem* (2.2) *can be written*

$$T_l(r,z) = \sum_{n\geq 0} a_n(r) Pe^n \partial_z^n T_l^\star(z)$$

if and only if $l \in D_{val}^\star$ *defined by*

$$D_{val}^\star = \big\{l/|l| \leq R\big\}, \quad \textit{where} \quad R = \inf\big\{|l|/G_{N,l}^\star = 0\big\},$$

and the functions $a_n(r)$ *are the solution of the recursive scheme*

$$\begin{cases} \Delta_N a_n(r) \;=\; v(r) a_{n-1}(r) & \textit{with} \quad a_{-1}(r) = 0, \\ a_0^\star = 1 \quad \textit{and} \quad a_n^\star = 0 & \textit{for} \quad n \geq 1. \end{cases} \tag{4.5}$$

Proof. In Appendix B we give the proof that the functions $G_{N,l}(r)$, $\partial_r G_{N,l}(r)$, and $\Delta_N G_{N,l}(r)$ are analytical with respect to l on the whole complex plane $\mathbb{C}$. More precisely there exists a set of functions $(q_n(r))_{n\in\mathbb{N}}$ (depending also on N) defined by (B.4) such that for $r \in [0,1]$ and $l \in \mathbb{C}$ one has

$$G_{N,l}(r) = \sum_{n\geq 0} q_n(r) l^n \ , \quad \partial_r G_{N,l}(r) = \sum_{n\geq 0} \partial_r q_n(r) l^n \ , \quad \Delta_N G_{N,l}(r) = \sum_{n\geq 0} \Delta_N q_n(r) l^n.$$

As a result the three functions $\frac{G_{N,l}(r)}{G^\star_{N,l}}$, $\frac{\partial_r G_{N,l}(r)}{G^\star_{N,l}}$, $\frac{\Delta_N G_{N,l}(r)}{G^\star_{N,l}}$ are analytical with respect to l for $l \in D^\star_{val}$ and $r \in [0,1]$ and there exist three sets of functions $(a_n(r))_{n\in\mathbb{N}}$, $(b_n(r))_{n\in\mathbb{N}}$, and $(c_n(r))_{n\in\mathbb{N}}$ such that for $l \in D^\star_{val}$ and $r \in [0,1]$,

$$\frac{G_{N,l}}{G^\star_{N,l}}(r) = \sum_{n\geq 0} a_n(r) l^n \ , \quad \frac{\partial_r G_{N,l}}{G^\star_{N,l}}(r) = \sum_{n\geq 0} b_n(r) l^n \ , \quad \frac{\Delta_N G_{N,l}}{G^\star_{N,l}}(r) = \sum_{n\geq 0} c_n(r) l^n. \tag{4.6}$$

Using the integration theorem on these series one gets

$$b_n(r) = \partial_r a_n(r) \quad \text{and} \quad c_n(r) = \Delta_N a_n(r) \quad \forall\, n \in \mathbb{N},$$

so that

$$\frac{\Delta_N G_{N,l}}{G^\star_{N,l}}(r) = \sum_{n\geq 0} \Delta_N a_n(r) l^n. \tag{4.7}$$

Now the function $\frac{G_{N,l}}{G^\star_{N,l}}(r)$ is the unique solution of the ODE:

$$\Delta_N f = l v(r) f \quad \text{and} \quad f^\star = 1. \tag{4.8}$$

Rewriting (4.8) with (4.6) and (4.7) gives that the functions $(a_n(r))_{n\in\mathbb{N}}$ are exactly given by the recursive scheme (4.5). ∎

4.1.2. Accessibility domain $D^\star_{acc}$. The eigenvalues predicted by the averaging method are the roots of the polynomial equation (3.5). Let us consider—as $p \to \infty$—the limit set of predicted eigenvalues Λ_∞ defined as the zeros set of the series $\sum_{n\geq 0} K_n \lambda^n$:

$$\Lambda_\infty = \left\{ \lambda \Big/ \sum_{n\geq 0} K_n \lambda^n = 0 \right\}, \tag{4.9}$$

where the index N on the macroscopic coefficient, assumed to be fixed, has been omitted. Among the eigenvalues predicted with the averaging method at order p, the only ones that make sense are those approximating some $\lambda \in \Lambda_\infty$, and by increasing the order p of the method one can only improve the computation on these modes $\lambda \in \Lambda_\infty$. As a result, for an eigenvalue $l \in L$ of the theoretical problem (2.2) to be approximated by the averaging method, and for this method to be convergent as $p \to \infty$ to this eigenvalue $l \in L$, it is necessary—but not sufficient—that the series $\sum_{n\geq 0} K_n l^n$ be convergent.

With definition (3.4) of the macroscopic coefficient K_n, the series $\sum_{n\geq 0} K_n l^n$ make sense for $l \in D^\star_{acc}$ defined as follows.

DEFINITION 4.3. *Let us call accessibility domain* $D^\star_{acc} \subset \mathbb{C}$ *the disk of all the complex* $\lambda \in \mathbb{C}$ *such that the series*

$$\sum_{n\geq 0} \alpha_n(r)\lambda^n, \quad \sum_{n\geq 0} \Delta_N\alpha_n(r)\lambda^n, \quad \sum_{n\geq 0} \partial_r\alpha_n(r)\lambda^n \tag{4.10}$$

are convergent for $r \in [0,1]$. *If* $\lambda \in D^\star_{acc}$, *we say that* λ *is accessible by the averaging method.*

Contrary to the validity domain $D^\star_{val}$, the accessibility domain $D^\star_{acc}$ does not depend only on the averaging operator $\star$ and on N, but also on the boundary conditions that influences the computation of functions α_n.

4.1.3. Evaluation of $D^\star_{val}$ and $D^\star_{ac}$. We here focus on the numerical evaluation of the two previously introduced domains $D^\star_{val}$ and $D^\star_{ac}$.

To compute the radius of the validity domain $D^\star_{val}$, we need to compute the smallest root of the function of l, $G^\star_{N,l}$. For this, we give in Appendix B an expansion of the generalized Graetz functions $G_{N,l}(r)$ with the help of a set of functions $q_n(r)$ defined in (B.4): $G_{N,l}(r) = \sum_{n\geq 0} q_n(r)l^n$. The computation of these functions $q_n(r)$ make it possible to compute the radius of $D^\star_{val}$ as the smallest root of the polynomial $\sum_{n\geq 0} q^\star_n l^n$.

To compute the radius of the accessibility domain $D^\star_{acc}$ one needs an upper bound on the three functions $\alpha_n(r)$, $\partial_r\alpha_n(r)$, and $\Delta_N\alpha_n(r)$ for $r \in [0,1]$. Experiments based on the computation of these functions showed that $\Delta_N\alpha_n(r=1)$ is a good upper bound for these functions and the radius of $D^\star_{acc}$ is equal to the convergence radius of the series $\sum_{n\geq 0} \Delta_N\alpha_n(r=1)\lambda^n$.

Radii for $D^\star_{val}$ and $D^\star_{acc}$ for some chosen weight functions $w(r)$ are given in Table 4.1. Comparing Table 4.1 with Tables 2.1 and 2.2 shows that the standard and the flow averaging method can only capture $l^{\mathcal{D}}_{0,0}$ for $\mathcal{D}$ and $l^{\mathcal{N}}_{0,0} = 0$ for $\mathcal{N}$. This result is self-consistent with the computations previously examined in section 2.2.3. To capture the first nontrivial eigenvalue for $\mathcal{N}$ one needs to use other averaging operators. Moreover, it will be shown in section 4.2 that the two necessary conditions introduced in the previous sections are actually sufficient for the convergence to hold. In addition, it will appear that the first nontrivial eigenvalue for $\mathcal{N}$, $l^{\mathcal{N}}_{1,0} \simeq -12.8398060$, and even the second eigenvalue for $\mathcal{D}$, $l^{\mathcal{D}}_{1,0} \simeq -22.30473055$, can be captured when using adapted averaging operators.

4.2. Convergence theorem. We introduced in the previous subsection two necessary conditions associated with any eigenvalue $l \in L$ to be captured by the averaging method. We prove here that these conditions are actually sufficient for the convergence to hold. More precisely, the eigenvalues predicted by the averaging method when $p \to \infty$ are exactly the eigenvalues of the theoretical problem (2.2) that both belong to the validity and accessibility domains.

TABLE 4.1
Radius of $D^\star_{val}$ *and* $D^\star_{acc}$ *for different weights and for* $N = 0$.

$w(r)$	$D^\star_{val}$	$D^\star_{acc}$, $\mathcal{D}$	$D^\star_{acc}$, $\mathcal{N}$
1	7.84	15.899	10.568
$2(1-r^2)$	$-l^{\mathcal{N}}_1 \simeq 12.839$	18.632	12.839
$1/(2r)$	354.75	24.789	14.665
$10(1-r)^3$	>500	29.82	23.33

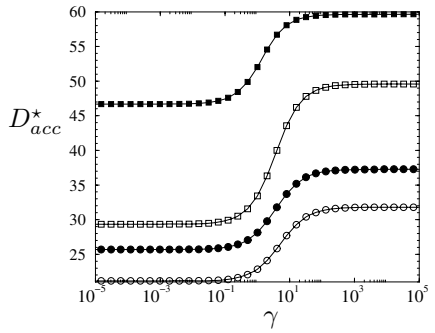


FIG. 4.1. *Computation of the accessibility domain $D^\star_{acc}$ for a Robin boundary condition versus the parameter γ and for $N = 0$. The four weighting functions w considered on Table* 4.1 *have been analyzed. Black circles stands for the flow-averaging method with $w = v/\langle v\rangle$, and white circles for the classical uniform volume averaging $w = 1$. White squares are for $w = 1/(2r)$ and black squares are for $w = 10(1-r)^3$.*

THEOREM 4.4. *Between the set of eigenvalues L of theoretical problem* (2.2) *and the three following sets: the validity domain $D^\star_{val}$ defined in Definition* 4.1*, the limit set of predicted eigenvalues Λ_∞ in* (4.9)*, and the accessibility domain $D^\star_{acc}$ defined in Definition* 4.3*, one has the relation (for any azimuthal mode N, any boundary condition $\mathcal{D}$, $\mathcal{N}$, or $\mathcal{R}$, and any averaging operator $\star$)*

$$\Lambda_\infty \cap D^\star_{val} = L \cap D^\star_{val} \cap D^\star_{acc}, \tag{4.11}$$

which means that the eigenvalues predicted by the averaging method inside $D^\star_{val}$ exactly converge toward the theoretical eigenvalues of (2.2) *being inside $D^\star_{val} \cap D^\star_{acc}$.*

Proof. We recall that we defined in (2.4) the generalized Graetz functions $G_{N,l}(r)$ for each $l \in \mathbb{C}$ and each $N \in \mathbb{Z}$ as the unique solution for the ODE,

$$\Delta_N G_{N,l} = l v(r) G_{N,l}(r), \quad \frac{G_{N,l}(r)}{r^N}(r=1) = 1, \tag{4.12}$$

and that for Neumann, Dirichlet, or Robin boundary conditions, the associated sets of theoretical eigenvalues are given by (2.5).

Using Lemma 4.2, one has that

$$L^{\mathcal{N}} \cap D^\star_{val} = \left\{ l \in \mathbb{C},\ \sum_{n\geq 0} \partial_r a_n(1) l^n = 0 \right\},$$

$$L^{\mathcal{D}} \cap D^\star_{val} = \left\{ l \in \mathbb{C},\ \sum_{n\geq 0} a_n(1) l^n = 0 \right\}, \quad \text{or}$$

$$L^{\mathcal{R}} \cap D^\star_{val} = \left\{ l \in \mathbb{C},\ \sum_{n\geq 0} (a_n(1) + \gamma \partial_r a_n(1)) l^n = 0 \right\},$$

where the functions $(a_n(r))_{n\in\mathbb{N}}$ are those defined by the recursive scheme (4.5). For simplicity, one introduces the quantities A_n defined as follows:

$$\text{for } \mathcal{D} \ :\ A_n = a_n(1); \quad \text{for } \mathcal{N} \ :\ A_n = \partial_r a_n(1); \quad \text{for } \mathcal{R} \ :\ A_n = a_n(1) + \gamma \partial_r a_n(1). \tag{4.13}$$

Thus for $\mathcal{D}$, $\mathcal{N}$, or $\mathcal{R}$ cases one has

$$L \cap D^{\star}_{val} = \left\{ l \in \mathbb{C},\ \sum_{n\geq 0} A_n l^n = 0 \right\}.$$

Let us consider the two functions of l as the sum of the following series in l:

$$A_l = \sum_{n\geq 0} A_n l^n, \qquad K_l = \sum_{n\geq 0} K_n l^n,$$

which are convergent for $l \in D^{\star}_{val} \cap D^{\star}_{acc}$.

Then, to prove (4.11) one exactly has to show that

$$\forall\, l \in D^{\star}_{val} \cap D^{\star}_{acc}\ :\ A_l = 0 \quad \text{if and only if} \quad K_l = 0. \tag{4.14}$$

To prove this, one has to find a relation between A_n and the macroscopic coefficient K_n. For this, one introduces the set of functions $\big(e_n(r)\big)$ associated with the difference between functions $a_n(r)$ and $\alpha_n(r)$ defined in (4.5) and (3.2):

$$e_n(r) = \alpha_n(r) - a_n(r)\ . \tag{4.15}$$

These functions, by subtracting (4.5) from (3.2), are exactly defined by the following recursive scheme:

$$\begin{cases} \Delta_N e_n(r) = K_n + v(r) e_{n-1}(r) & \text{with } e_{-1}(r) = 0, \\ e^{\star}_n = 0, & \\ e_n(1) = -A_n & \text{for } \mathcal{D}, \\ \partial_r e_n(1) = -A_n & \text{for } \mathcal{N}, \\ e_n(1) + \gamma \partial_r e_n(1) = -A_n & \text{for } \mathcal{R}. \end{cases} \tag{4.16}$$

This recursive formula does depend on both macroscopic coefficients K_n and A_n. Let us finally define the macroscopic difference function $E_l(r)$ by

$$E_l(r) = \sum_{n\geq 0} e_n(r) l^n = \sum_{n\geq 0} \alpha_n(r) l^n - \sum_{n\geq 0} a_n(r) l^n,$$

which is well defined for $l \in D^{\star}_{val} \cap D^{\star}_{acc}$.

We search a differential problem satisfied by $E_l(r)$.

Thanks to Lemma 4.9 on $D^{\star}_{val}$ and to Definition 4.3 of $D^{\star}_{val}$, the series

$$\sum_{n\geq 0} a_n(r) l^n, \quad \sum_{n\geq 0} \partial_r a_n(r) l^n, \quad \sum_{n\geq 0} \Delta_N a_n(r) l^n \quad \text{and}$$
$$\sum_{n\geq 0} \alpha_n(r) l^n, \quad \sum_{n\geq 0} \partial_r \alpha_n(r) l^n, \quad \sum_{n\geq 0} \Delta_N \alpha_n(r) l^n$$

converge for all $l \in D^{\star}_{val} \cap D^{\star}_{acc}$ and all $r \in [0,1]$. Then, using the integration theorem and the properties (4.16) of functions $e_n(r)$ one has

$$\forall\, l \in D^{\star}_{val} \cap D^{\star}_{acc} \quad \forall\, r \in [0,1] \quad : \begin{cases} \Delta_N E_l(r) = K_l + v(r) E_l(r), & \\ E^{\star}_l = 0, & \\ E_l(1) = -A_l & \text{for } \mathcal{D}, \\ \partial_r E_l(1) = -A_l & \text{for } \mathcal{N}, \\ E_l(1) + \gamma \partial_r E_l(1) = -A_l & \text{for } \mathcal{R}. \end{cases}$$

K_l or A_l being fixed, this problem has one and only one solution so that A_l is a function of K_l and vice versa. Now, it is easy to check that the solution associated with $A_l = 0$ is $E_l = 0$, which eventually fixes $K_l = 0$. Conversely, and for the same reason, $K_l = 0$ fixes $A_l = 0$. This ensures (4.14), which proves Theorem 4.4. ∎

It is interesting to note that because $E_l = 0$ is the solution associated with a converging eigenvalue $\lambda_\infty \in \Lambda_\infty = l \in L$, the ratio between the predicted eigenfunction and its value at $r = 0$ also converges to the theoretical Graetz eigenfunction. This leads to the following important corollary.

COROLLARY 4.5. *For an eigenvalue $l \in L \cap D^\star_{val} \cap D^\star_{acc}$, with an associated set of approximated eigenvalues $(\lambda_p)_{p\in\mathbb{N}}$ such that $\lim_{p\to\infty} \lambda_p = l$, let us define the approximated eigenfunction Γ_{λ_p} as*

$$\Gamma_{\lambda_p}(r) = \frac{1}{\rho}\sum_{n=0}^{p} \alpha_n(r)\lambda_p^n \quad \textit{with} \quad \rho = \sum_{n=0}^{p} \frac{\alpha_n(r)}{r^N}(r=0)\lambda_p^n; \tag{4.17}$$

then Γ_{λ_p} converges toward the generalized Graetz function $G_{N,l}$,

$$\lim_{p\to\infty} \|\Gamma_{\lambda_p} - G_{N,l}\| = 0. \tag{4.18}$$

Moreover, defining the amplitude c_λ, in the same way as c_l defined in (2.7),

$$c_{\lambda_p} = \frac{\displaystyle\int_0^{2\pi}\int_0^1 T_0(r)\overline{\Gamma_{\lambda_p}(r)}e^{-iN\phi}v(r)r\,dr}{\displaystyle 2\pi\int_0^1 |\Gamma_{\lambda_p}(r)|^2 v(r)r\,dr}; \quad \textit{then} \quad \lim_{p\to\infty} |c_{\lambda_p} - c_l| = 0. \tag{4.19}$$

Hence, not only does Theorem 4.4 give a necessary and sufficient condition for an eigenvalue to converge, but also the eigenmode will converge to the corresponding theoretical solution. We now numerically study the convergence of various eigenmodes for different averaging operator w.

4.3. Convergence evaluation. This section studies the numerical evaluation of the convergence to either the eigenvalue, the eigenfunction, or the eigenmode amplitude for a Poiseuille parabolic velocity profile $v(r) = 2(1 - r^2)$.

We calculate the closure functions α_n from the recursive scheme (3.2), so that the coefficients K_n defined in (3.4) of the eigenvalues polynomial (3.5) can be computed. From the obtained solution leading to $p + 1$ eigenvalues, we select the larger one in $\mathbb{R}^-$. Figure 4.2 displays the relative error of this approximated eigenvalue for different weighting functions w. For the first Dirichlet eigenvalue, Figure 4.2(a) displays exponential convergence rates. Moreover, when comparing the results of Figure 4.2(a) with Table 4.1, it is not surprising to observe that a larger radius of convergence $D^\star_{acc}$ gives rise to a faster convergence rate. As demonstrated in the previous section, the second eigenvalue for the Dirichlet or Neumann boundary condition is not accessible to the standard volume averaging methods—$w = 1$—or the kinematic volume averaging—$w = v/\langle v\rangle$. On the contrary, two other weighting functions w have been proposed in Table 4.1, the convergence of which has been established for the second eigenvalue in the previous section. Figures 4.2(b) and (c) study their convergence on the second eigenvalue in the Dirichlet and Neumann cases. It is interesting to observe on these figures that the convergence rate still looks exponential, even if the convergence rate is much slower than those observed in Figure 4.2(a). More modes should indeed be needed for an acceptable precision to be obtained.

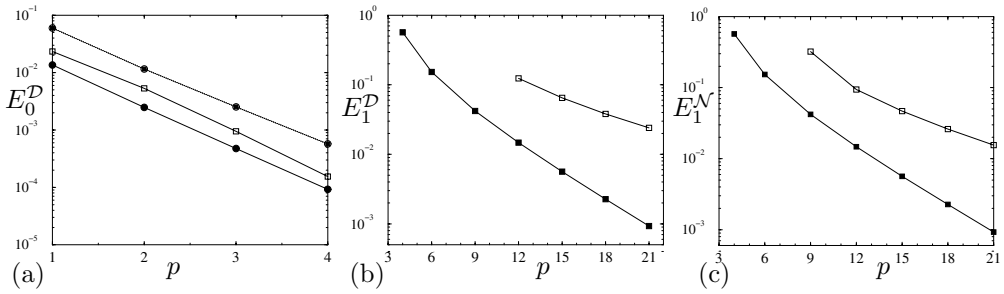


FIG. 4.2. *Relative error for axisymmetrical $N = 0$ eigenvalues $l_{0,0}$ and $l_{0,1}$.* (a) *Relative error $E_0^{\mathcal{D}} = |\lambda_{0,p}^{\mathcal{D}} - l_{0,0}^{\mathcal{D}}|/l_{0,0}^{\mathcal{D}}$ between the predicted eigenvalue and the theoretical one $l_{0,0}^{\mathcal{D}} = -3.656793458$, versus the order p of the approximation. Black circles stand for the flow-averaging method with $w = v/\langle v \rangle$, and white circles for the classical uniform volume averaging $w = 1$. White squares are for $w = 1/(2r)$ and black squares are for $w = 10(1-r)^3$. In every case the convergence is exponential, as indicated by the observed semilog linear behavior.* (b) *We use the same conventions for the second Dirichlet eigenvalue $l_{0,1}^{\mathcal{D}} = -22.30473055$ convergence $E_1^{\mathcal{D}} = |\lambda_{1,p}^{\mathcal{D}} - l_{0,1}^{\mathcal{D}}|/l_{0,1}^{\mathcal{D}}$.* (c) *We use the same conventions for the second Neumann eigenvalue convergence $l_{0,1}^{\mathcal{N}} = -12.8398060$ with $E_1^{\mathcal{N}} = |\lambda_{1,p}^{\mathcal{N}} - l_{0,1}^{\mathcal{N}}|/l_{0,1}^{\mathcal{N}}$.*

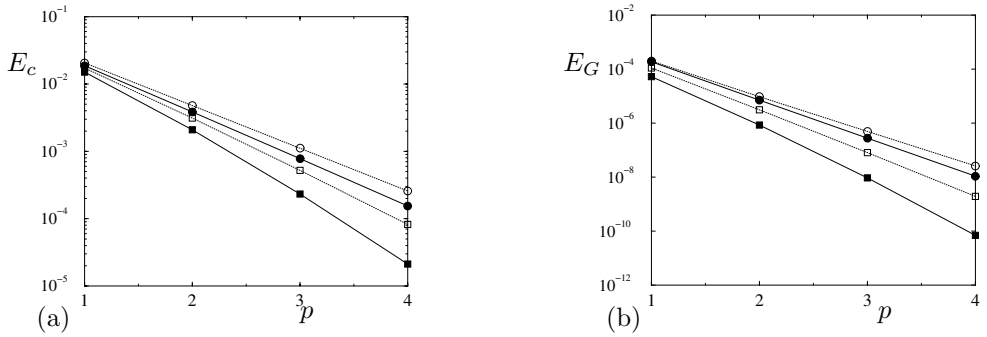


FIG. 4.3. *Relative error for axisymmetrical $N = 0$ eigenmode.* (a) *Relative error $E_c = |c_{\lambda_{0,p}^{\mathcal{D}}} - c_{l_{0,0}^{\mathcal{D}}}|/c_{l_{0,0}^{\mathcal{D}}}$ between the predicted amplitude and its theoretical value associated with a uniform initial temperature $T_0 = 1$ at $z = 0$ for the first Dirichlet eigenmode, versus the order p of the approximation.* (b) *Absolute error $E_G = ||\Gamma_{\lambda_{0,p}^{\mathcal{D}}} - G_{l_{0,0}^{\mathcal{D}}}|| = \langle w(\Gamma_{\lambda_{0,p}^{\mathcal{D}}} - G_{l_{0,0}^{\mathcal{D}}})^2 \rangle$ on the predicted eigenfunction for the first Dirichlet eigenmode.*

Moreover, for finite values of the spectral cut-off p, the second eigenvalue could not always be captured. For example, this can be observed in Figure 4.2(b) in the case of weighting function $w = 10(1-r)^3$, for which the eigenvalue becomes real, so that it is considered to be captured by the approximation for $p \geq 12$ only. This example also illustrates that an empirical test of the convergence is not always successful. If one would have guessed, ignoring the convergence proof, from the computation of the first 10 mapping variables α_p, $p < 10$, that the first eigenvalue computed in Figure 4.2(b) is captured by the weighting function $w = 10(1-r)^3$, it would have found the wrong answer. Figure 4.3 displays the convergence of the amplitude and the eigenfunction defined in Corollary 4.5 for the first Dirichlet mode. It is interesting to note that even the first approximation $p = 1$ that has been detailed in section 2.2.3 permits a rather precise amplitude and eigenmode estimate for every tested weighting function w. The convergence rate displayed on Figure 4.3 is also found to be exponential,

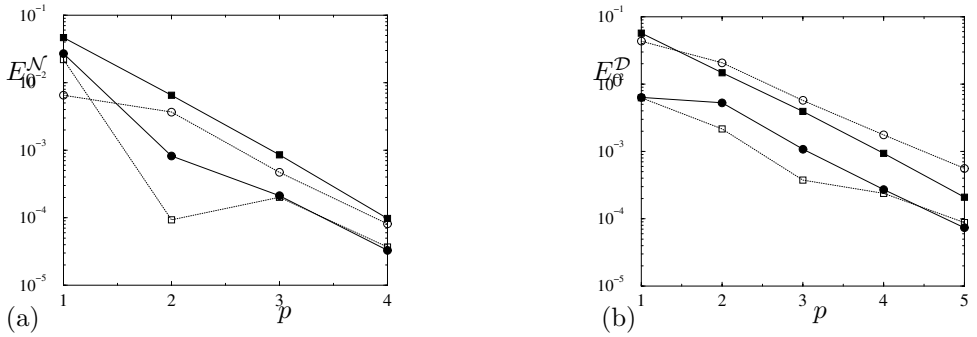


FIG. 4.4. *We use the same conventions as in Figure* 4.2 *for the relative error for nonaxisymmetrical* $N = 1$ *eigenvalues* $l_{1,0}$. (a) *Relative error* $E_0^{\mathcal{N}} = |\lambda_{0,p}^{\mathcal{N}} - l_{1,0}^{\mathcal{N}}|/l_{1,0}^{\mathcal{N}}$ *between the predicted eigenvalue and the theoretical Neumann one* $l_{1,0}^{\mathcal{N}} = -4.160532810$, *versus the order* p *of the approximation.* (b) *We use the same conventions for the first Dirichlet eigenvalue convergence* $l_{1,0}^{\mathcal{D}} = -10.69115115$ *with* $E_0^{\mathcal{D}} = |\lambda_{1,p}^{\mathcal{D}} - l_{1,0}^{\mathcal{D}}|/l_{1,0}^{\mathcal{D}}$.

as already observed for the eigenvalue convergence. This result does not seem very surprising, for the generalized averaging method has many characteristics in common with a spectral discrete method.

Finally, nonaxisymmetrical mode convergence have been investigated. The convergence of the leading order $N = 1$ eigenvalue is represented on Figure 4.4. It is interesting to observe that low order approximation (e.g., $p < 5$) gives rise to a rather precise estimation of this first nonaxisymmetric mode. It should then be noted that for both Neumann and Dirichlet boundary conditions, $|l_{1,0}|$ is smaller than $|l_{0,1}|$. Hence, the better convergence of Figures 4.4(a) and (b) compared to Figures 4.2(b) and (c) can be qualitatively understood. Neumann and Dirichlet situations give lower and upper bounds for the convergence of the more general Robin boundary condition when varying γ from 0 to infinity. Hence, the Robin case should converge the same way as it is observed in the above figures.

5. Discussion and conclusion.

5.1. Discussion. This section discusses the results obtained in the previous sections in the light of previous analysis found in the literature.

5.1.1. Context of the presented analysis. As already mentioned in section 2.1 after defining the convection diffusion problem, (2.1), different characteristic lengths can be chosen for making dimensionless the longitudinal dimension z, and this leads to different Péclet numbers $Pe \gg 1$ or $pe \ll 1$. Any choice should lead to consistent results. When choosing the Péclet number $Pe \gg 1$, it is known that longitudinal variations along z scale linearly with Pe. This result holds as an asymptotic expansion discarding $O(1/Pe^2)$ terms [1], and leads to the simplified constitutive equation (2.2). Balakotaiah and Chang [5] mention that the condition $Pe \gg 6.93$ is necessary for neglecting axial diffusion. The linear scaling of z variations with Pe is described in solution (2.3) and used in the average description of the problem (2.12). From this nondimensionalization choice, it appears that standard [39] "ad hoc" closure relations used in (2.17) and (3.1) do not depend on the Péclet number because each z derivative cancels the corresponding algebraic dependence in Pe. It then appears that closure relations (2.17) and (3.1) are in fact an asymptotic expansion that in-

volves the eigenvalue l of the problem as a small parameter. The validity range of this asymptotic expansion, which should better be described as an analytic expansion of the eigenfunction with the eigenvalue, is investigated in section 4, while in the meantime the "ad hoc"closure relations are a posteriori justified by the convergence proof obtained in the same section. All the validity range results for eigenvalues are obtained independently from the value of the Péclet number, but are valid for $Pe \gg 1$ since the starting constitutive equation (2.2) derives from (2.1) discarding $O(1/Pe^2)$ terms [1].

5.1.2. Comparison with other convergence results. It is now interesting to more clearly compare our analysis with previously obtained convergence results. For example, some convergence criteria have been discussed in the context of center manifold approximations of the convection-diffusion problem (2.1) by Balakotaiah and Chang in [4]. In the case of spatially varying solutions, the solution is projected over Graetz eigenfunctions, and a criterion has been found by summing the expansion series. The convergence criterion can be expressed in the same framework by considering the smallest longitudinal variations associated with a critical λ_c. In the case of Dirichlet boundary conditions, it was found $\lambda_c^{\mathcal{D}} = 13.80$ in [4], whereas $\lambda_c^{\mathcal{N}} = 37.7$ was obtained in the case of Neumann boundary conditions. These values have to be compared with the results in Table 4.1. One has to note, that, in our case, the convergence radius D_{acc} obtained from computing the expansion series is not the only relevant parameter for convergence. D_{val}, which comes from the analyticity condition on the averaged Graetz eigenfunction that we wish to approximate, must also be considered. The convergence radius is the minimum of D_{acc} and D_{val}.

The Lyapunov–Schmidt reduction technique such as used in [5, 11] is also another method that should be compared to our analysis. As mentioned in the introduction, this approximation shares much similarities with ours, and the results are also quite similar. In this case, the considered equation (2.1) is written by making dimensionless the longitudinal direction z by L so that the Péclet number is replaced by the small parameter $pe = PeR/L$, as already indicated in section 2.1. The first step of the Lyapunov–Schmidt reduction approximation is to look for a regular asymptotic expansion solution of (2.1) in terms of the small parameter pe. The solution is then decomposed into two parts similar to (2.10) and (3.1) (but based on a splitting of the linear operator eigenfunctions into "master" eigenfunctions of the kernel of the adjoint operator and "slave" eigenfunctions of the image of the adjoint operator; see, for example, [11]). In the case of the Neumann boundary condition and weighting function $w = 1$, the first closure field solutions that we have obtained are exactly the same as those previously obtained in [5, 11]. More precisely, the first slave mode computed in equation (31) of [11] or equation (3.8) of [5] corresponds to the Neumann solution $\alpha_{1,0}$ found in (2.22). Nevertheless, higher expansion closure fields differ from those of Chakraborty and Balakotaiah [11]. By summing the expansion series, those authors have been able to find a convergence radius for the approximation. Following criteria equation (73) of [11], and the discussion in section 3 of [5], the convergence radius of the Neumann boundary conditions with uniform averaging is $\lambda_c^{\mathcal{N}} = 48 \times 0.288 = 13.8$, which should be compared with the value 10.56 in Table 4.1.

This comparison shows that some of our convergence results are very similar to those previously obtained in the literature with other approaches.

5.1.3. Azimuthal averaging. In this section we discuss the possibility and the interest of azimuthal averaging. First, it should be noted that relation (2.7) gives the amplitude of each nonaxisymmetrical mode of the theoretical solution. If, for example,

an initial condition is chosen with a pulse at a given location (r_0, ϕ_0), i.e., $T_0(r,\phi) = \delta(r-r_0)\delta(\phi-\phi_0)$, then all nonaxisymmetrical modes, $N \neq 0$ will be represented with a weight $c_{N,l} = G_{N,l}(r_0)v(r_0)r_0/\int_0^1 |G_{N,l}(r)|^2 v(r)r\,dr$ because the Fourier transform of the Dirac distribution is uniform. In that case, if one averages the theoretical solution with a uniform weight function along the azimuthal angle ϕ, all nonaxisymmetrical mode,s $N \neq 0$ will not contribute to the averaged concentration because $\langle e^{iN\phi}\rangle_\phi = 0$ for $N \neq 0$. This is not true when using a nonuniform averaging operator w_ϕ along the azimuthal angle ϕ. In this case, there should be some contribution to the averaged concentration coming from nonaxisymmetrical mode $N \neq 0$, summing $c_{N,l}\langle e^{iN\phi}w_\phi\rangle_\phi\langle G_{N,l}\rangle^\star$ contributions.

Some of these nonaxisymmetrical contributions to the true averaged concentration solution could indeed be captured by an averaging method, as shown in the previous sections. Hence, for each nonaxisymmetrical eigenvalue l, one can obtain the appropriate averaging approximation $c_{\lambda_p}\langle e^{iN\phi}w_\phi\rangle_\phi\langle\Gamma_{\lambda_p}\rangle^\star$ of its contribution to the averaged solution.

5.2. Conclusion. This paper analyzes the convergence of volume averaging methods on unidirectional convection-diffusion problems. Neumann, Robin, and Dirichlet boundary conditions have been considered. The last problem is of a great interest in the case of local nonequilibrium conditions, i.e., averaged temperature not equal to the value at the boundary for which approximate solutions are more difficult to obtain.

Concentrating on the stationary solution associated with large Péclet numbers, it has been found that volume averaging methods converge toward the exact solution. A necessary and sufficient condition for this convergence to occur has been found for any unidirectional velocity field, which depends on the averaging operator w. This condition has been obtained in a general form as related to the analytical character of the averaged eigenfunction with the eigenvalue λ. This condition has in fact a general scope, because it is the basis for writing "closure relations" as a power series of the eigenvalue.

It is interesting to note that the convergence also depends, obviously, on the eigenvalue to be captured. In the case of a parabolic velocity profile, the convergence to the Graetz solution has been studied in more detail. In the case of Dirichlet boundary conditions, "natural" operators $w = 1$ or $w = v$ allow the convergence to the first nontrivial eigenvalue. In the case of a Neumann boundary condition, these usual weighting operators do not capture the first nontrivial eigenvalue of the Graetz problem. In this case, it is necessary to use other averaging operators w to get the first spatially decaying mode, some of which have been proposed in this paper.

This result shows that averaging over some spatial volume unavoidably degenerates the space of mathematically accessible solutions. Nevertheless, despite smoothing out the small scales—the large eigenvalues—the averaged solution can lead to an asymptotically exact representation of the large scale structure—the small eigenvalues —of the solution. It is expected that this conclusion could be of some general scope when decreasing the dimension number of a problem by averaging along part of its dimensions.

Moreover, the mathematical proof presented in this paper has been complemented in the case of a parabolic Poiseuille flow by some numerical computation of convergence rates. They have been found to be exponential, as expected from a spectral discrete method. It should also be of some general scope when averaging linear problems. It is interesting to note that the convergence toward nontrivial eigenvalues is

directly related to a correct evaluation of the heat transfer between the fluid and the solid boundary. As a matter of fact it should be kept in mind that the Nusselt number Nu, defined as usual as the dimensionless number associated with the heat (or mass) transfer [14] scales asymptotically, when $z \gg Pe/(l_1 - l_0)$, as $Nu = l_1^2/2$. Hence, convergence toward the eigenvalue of the averaged model is also directly related to a correct evaluation of the asymptotic transfer between the flow and the solid.

Different extensions of this work could be considered. First, a direct transposition of the convergence proof in the case of a plane geometry, with transverse velocity field, should be easily obtained. The quantitative results on the accessibility domain as well as on the convergence accuracy could nevertheless be different in that case. The second extension of interest should be related to more complicated situations associated with a coupling with conduction in some external solid domain.

Appendix A. Graetz functions and Kummer's functions. The generalized Graetz functions are the eigenfunctions of the operator $\frac{1}{1-r^2}\Delta_N$,

$$\frac{1}{1-r^2}\Delta_N \equiv \frac{1}{1-r^2}\left(\partial_r^2 + \frac{1}{r}\partial_r - \frac{N^2}{r^2}\right). \tag{A.1}$$

One wants to solve the self-adjoint Sturm–Liouville problem

$$\frac{1}{1-r^2}\Delta_N f = -\ell^2 f, \tag{A.2}$$

where we have introduced the positive eigenvalue $\ell^2 = -l$ to compare to (2.4). Defining a new function y, from $f(r) = r^N e^{-\frac{\ell}{2}r^2} y(\ell r^2)$, y is then a solution of the hypergeometric equation

$$z\partial_z^2 y + (1 + N - z)\,\partial_z y - \left(\frac{1+N}{2} - \frac{\ell}{4}\right) y = 0. \tag{A.3}$$

In its more general form, the hypergeometric equation reads

$$z\partial_z^2 y + (c - z)\,\partial_z y - ay = 0, \tag{A.4}$$

which possesses two solutions called confluent hypergeometric functions, and when $c = 1$,

- the first one is singular at $z = 0$ and is not considered here;
- the other one is regular, convergent, and denoted $\Phi(a, c, z)$, It is defined by the Kummer's series (with infinite radius of convergence)

$$\Phi(a, c, z) = 1 + \frac{a}{c}z + \frac{a(a+1)}{c(c+1)}\frac{z^2}{2} + \cdots + \frac{a\cdots(a+n-1)}{c\cdots(c+n-1)}\frac{z^n}{n!} + \cdots. \tag{A.5}$$

f is proportional to the Graetz function $G_{N,\ell}$,

$$G_{N,\ell}(r) = r^N e^{-\ell r^2/2}\Phi\left(\frac{1+N}{2} - \frac{\ell}{4}, 1 + N, \ell r^2\right). \tag{A.6}$$

Appendix B. Analyticity in l of the Graetz functions. In this appendix we prove that the generalized Graetz functions defined in (2.4) $G_{N,l}(r)$ are analytical

in l on the whole complex field $\mathbb{C}$. More precisely, for the closure functions $q_{N,n}(r)$ defined in (B.4) one has for each $l \in \mathbb{C}$

$$G_{N,l}(r) = \sum_{n\geq 0} q_{N,n}(r) l^n, \quad \partial_r G_{N,l}(r) = \sum_{n\geq 0} \partial_r q_{N,n}(r) l^n. \tag{B.1}$$

We point out that this result is true for any $N \in \mathbb{Z}$ and for any flow $v(r)$ that is nonnegative and analytical in 0.

We shall prove this result in two steps:

- in section B.1 we prove that (B.1) is true when l belongs to a disk $D \subset \mathbb{C}$ which we characterize;
- in section B.2 we prove that $D = \mathbb{C}$.

We first recall the following definitions:

For a given value $N \in \mathbb{Z}$ of the axisymmetric parameter, the operator Δ_N is defined as

$$\Delta_N \equiv \partial_r^2 + \frac{1}{r}\partial_r - \frac{N^2}{r^2},$$

so that $\Delta_{-N} = \Delta_N$. Hence, we will consider the case $N \geq 0$ only.

The operator Δ_N can be written under a divergence form

$$\Delta_N f = \frac{1}{r^{N+1}} \partial_r \left(r^{2N+1} \partial_r \left(\frac{f}{r^N} \right) \right). \tag{B.2}$$

For each $l \in \mathbb{C}$ the Graetz function $G_{l,N}$ is the only solution for the following ODE:

$$\begin{cases} \Delta_N G_{N,l} = l v(r) G_{N,l}(r), \\ \dfrac{G_{N,l}}{r^N}(0) = 1. \end{cases} \tag{B.3}$$

We define the set of closure functions $q_{N,n}$, for $n \geq 0$, as follows:

$$\begin{cases} \Delta_N q_{N,n} = v(r) q_{N,n-1}(r) \quad \text{with } q_{N,-1} = 0, \\ \dfrac{q_{N,0}}{r^N}(0) = 1 \quad \text{and} \quad \dfrac{q_{N,n}}{r^N}(0) = 0 \quad \text{for} \quad n \geq 1. \end{cases} \tag{B.4}$$

B.1. A criterion for the Graetz function to be analytical in l.

THEOREM B.1. *Let D be the convergence disk of the series*

$$\sum_{n\geq 0} q_{N,n}(1) l^n, \tag{B.5}$$

where the closure functions $q_{N,n}$ are defined in (B.4)*. Then for all $l \in D$,*

$$\begin{gathered} G_{N,l}(r) = \sum_{n\geq 0} q_{N,n}(r) l^n, \quad \partial_r G_{N,l}(r) = \sum_{n\geq 0} \partial_r q_{N,n}(r) l^n, \\ \text{and} \quad \Delta_N G_{N,l}(r) = \sum_{n\geq 0} \Delta_N q_{N,n}(r) l^n. \end{gathered} \tag{B.6}$$

Proof. We begin by proving that for a fixed $l \in D$ the three series $\sum_{n\geq 0} q_{N,n}(r)l^n$, $\sum_{n\geq 0} \partial_r q_{N,n}(r)l^n$, and $\sum_{n\geq 0} \Delta_N q_{N,n}(r)l^n$ are uniformly convergent for $r \in [0,1]$.

First of all the recursive definition (B.4) of the functions $q_{N,n}$ implies that, for all $n \geq 0$, $q_{N,n}(r) = r^N \psi_n(r)$, where ψ_n is a nonnegative, nondecreasing, continuous function on $[0,1]$,

(B.7)
$$q_{N,0} = r^N \quad \text{and} \quad q_{N,n}(r) = r^N \int_0^r \frac{1}{y^{2N+1}} \int_0^y x^{N+1} v(x) q_{N,n-1}(x)\, dx\, dy \quad \text{for} \quad n \geq 1,$$

so that $0 \leq q_{N,n}(r) \leq q_{N,n}(1)$ and the series $\sum_{n\geq 0} q_{N,n}(r)l^n$ is uniformly converging on $[0,1]$ for $l \in D$.

In the same way $\Delta_N q_{N,n}(r) = v(r) q_{N,n-1}(r)$, and so one has $0 \leq \Delta_N q_{N,n}(r) \leq \|v\| q_{N,n-1}(1)$ and the series $\sum_{n\geq 0} \Delta_N q_{N,n}(r)l^n$ is uniformly converging on $[0,1]$ for $l \in D$.

Now one has

$$\begin{aligned} 0 \leq \partial_r q_{N,n}(r) &= N r^{N-1} \int_0^r \frac{1}{y^{2N+1}} \int_0^y x^{N+1} v(x) q_{N,n-1}(x)\, dx\, dy \\ &\quad + \frac{1}{r^{N+1}} \int_0^r x^{N+1} v(x) q_{N,n-1}(x) dx \\ &\leq N r^{N-1} q_{N,n}(1) + \frac{\|v\|}{N+2} r q_{N,n-1}(1) \\ &\leq C \left(q_{N,n}(1) + q_{N,n-1}(1) \right), \end{aligned}$$

where the constant C depends only on N and v so that the series $\sum_{n\geq 0} \partial_r q_{N,n}(r)l^n$ is uniformly converging on $[0,1]$ for $l \in D$.

Now, for a given value $l \in D$ we introduce the two functions defined on $[0,1]$,

$$F(r) = \sum_{n\geq 0} q_{N,n}(r)l^n, \quad H(r) = \sum_{n\geq 0} \Delta_N q_{N,n}(r)l^n,$$

since these are uniformly converging series, and since $\sum_{n\geq 0} \partial_r q_{N,n}(r)l^n$ is also a uniformly converging series for $r \in [0,1]$, one can use the integration theorem, which implies that

$$H(r) = \Delta_N F(r) \quad \text{for } r \in [0,1],$$

and at the same time one has with (B.4) that $H(r) = l v(r) F(r)$ and that $\frac{F}{r^N}(0) = 1$. The unicity of the solutions of (B.3) ensures then that $F(r) = G_{N,l}(r)$, and this ends the proof. ∎

B.2. Analyticity on the whole complex field $\mathbb{C}$.

LEMMA B.2. *The series*

$$\sum_{n\geq 0} q_{N,n}(1)l^n$$

is convergent on the whole complex plane $\mathbb{C}$ *and so* (B.1) *is true for all* $l \in \mathbb{C}$.

Proof. With the integral formulation (B.7) on the closure functions $q_{N,n}$ one has

$$\begin{aligned} q_{N,n+m}(1) &= \int_0^1 \frac{1}{y_1^{2N+1}} \int_0^{y_1} x_1^{N+1} v(x_1) q_{N,n+m-1}(x_1)\, dx_1\, dy_1 \\ &= \int_0^1 \frac{1}{y_1^{2N+1}} \int_0^{y_1} x_1^{2N+1} v(x_1) \cdots \int_0^{x_{m-1}} \frac{1}{y_m^{2N+1}} \\ &\quad \times \int_0^{y_m} x_m^{N+1} v(x_m) q_{N,n}(x_m) dx_m dy_m \cdots dx_1 dy_1, \end{aligned}$$

and since $0 \leq q_{N,n}(r) \leq r^N q_{N,n}(1)$ (see (B.7)), we have

$$\begin{aligned} \frac{q_{N,n+m}(1)}{q_{N,n}(1)} &\leq \int_0^1 \frac{1}{y_1^{2N+1}} \int_0^{y_1} x_1^{2N+1} v(x_1) \cdots \int_0^{x_{m-1}} \frac{1}{y_m^{2N+1}} \\ &\quad \times \int_0^{y_m} x_m^{2N+1} v(x_m) dx_m\, dy_m \cdots dx_1\, dy_1 \\ &\leq \|v\|^m \int_0^1 \frac{1}{y_1^{2N+1}} \int_0^{y_1} x_1^{2N+1} \cdots \int_0^{x_{m-1}} \frac{1}{y_m^{2N+1}} \\ &\quad \times \int_0^{y_m} x_m^{2N+1} dx_m\, dy_m \cdots dx_1\, dy_1, \end{aligned}$$

where $\|v\| = \sup v(r)$.

This upper bound can be computed explicitly,

$$\frac{q_{N,n+m}(1)}{q_{N,n}(1)} \leq \|v\|^m \frac{1}{2(2N+2)} \cdots \frac{1}{2m(2N+2m)} := \alpha_m,$$

and $\alpha_m^{-1/m}$ is a lower bound for the radius of convergence of the series (B.2). One can easily check that

$$\alpha^{-1/m} \geq 2 \frac{2N+2}{\|v\|} (m!)^{1/m},$$

and so $\alpha^{-1/m}$ grows up to infinity. As a result the series (B.2) is convergent on the whole complex plane $\mathbb{C}$. ∎

Appendix C. Invertibility of the operator $\mathcal{L}_N^\star$ and resolution of the closure problems. In this appendix we prove that the closure problems

$$\text{(C.1)} \quad \begin{cases} \mathcal{L}_N^\star \alpha_n = v(r)\alpha_{n-1}(r) - \langle v\alpha_{n-1}\rangle^\star & \text{with } \alpha_{-1}(r) = 0, \\ \alpha_0^\star = 1 \quad \text{or} \qquad \alpha_n^\star = 0 & \text{for } n \geq 1 + \text{boundary condition} \end{cases}$$

for a boundary condition either of a homogeneous Dirichlet, homogeneous Neumann, or Robin type,

$$\text{(C.2)} \qquad \alpha_n(1) = 0, \quad \partial\alpha_n(1) = 0, \quad \text{or} \quad \partial\alpha_n(1) + \gamma\alpha_n(1) = 0,$$

has one and only one bounded solution for each $n \in \mathbb{N}$.

The operator $\mathcal{L}_N^\star$ is defined for $N \in \mathbb{Z}$ and for a normalized averaging operator $\star$ (i.e., such that $\langle 1\rangle^\star = 1$) by

$$\text{(C.3)} \qquad \mathcal{L}_N^\star f = \Delta_N f - \langle \Delta_N f\rangle^\star;$$

for the operator Δ_N,

$$\Delta_N f = \partial_r^2 f + \frac{1}{r}\partial_r f - \frac{N^2}{r^2} f.$$

Because $\Delta_N = \Delta_{-N}$ we will only consider here the proof for $N \geq 0$.

We proceed in two steps: in section C.1 we prove a lemma on the general solution of $\Delta_N f = g$ and in section C.2 we apply that lemma to the problems (C.1) for every boundary condition (C.2).

C.1. A technical lemma.

LEMMA C.1. *Let g be a continuous function defined on $[0,1]$ and such that $g^\star = 0$. Then for all $A \in \mathbb{R}$ the ODE*

$$\Delta_N f - A = g, \tag{C.4}$$

$$f^\star = M \in \mathbb{R} \tag{C.5}$$

has one and only one bounded solution on $[0,1]$.

Moreover, this solution fulfills

$$\langle \Delta_N f \rangle^\star = A$$

and then is a solution of

$$\begin{cases} \mathcal{L}_N^\star f = g, \\ f^\star = M. \end{cases}$$

Proof. We define the function $\psi_1(r)$,

$$\psi_1(r) = -r^N \int_r^1 \frac{1}{y^{2N+1}} \int_0^y x^{N+1} g(x)\, dx\, dy, \tag{C.6}$$

which is well defined since g is continuous in 0 for $N \geq 0$, and the function $\psi_2(r)$,

$$\psi_2(r) = \frac{r^N - r^2}{N^2 - 4} \quad \text{if } N \neq 2 \quad \text{and} \quad \psi_2(r) = \frac{r^2}{4}\ln(r) \quad \text{for } N = 2. \tag{C.7}$$

Any solution of (C.4) is of the form

$$f(r) = \lambda r^N + \mu r^{-N} + A\psi_2(r) + \psi_1(r) \quad \text{if } N \neq 0, \quad \text{or}$$
$$f(r) = \lambda r^N + \mu \ln(r) + A\psi_2(r) + \psi_1(r) \quad \text{if } N = 0.$$

Then all bounded solutions of (C.4) on $[0,1]$ are on the form

$$f(r) = \lambda r^N + A\psi_2(r) + \psi_1(r), \tag{C.8}$$

and (C.5) gives

$$\lambda = \frac{M - A\psi_2^\star - \psi_1^\star}{\langle r^N \rangle^\star}.$$

Thus (C.4) and (C.5) have only one bounded solution.

Since $g^\star = 0$ one also has $\langle r^N \rangle^\star = A$. ∎

C.2. Resolution of the closure problems.

Homogeneous Dirichlet case. We consider the solution f as in (C.8) of (C.4) and (C.5) and search for a value of A such that $f(1) = 0$.

We have

$$f(1) = \frac{M - A\psi_2^\star - \psi_1^\star}{\langle r^N \rangle^\star}. \tag{C.9}$$

So there is only one bounded solution f of (C.4) and (C.5) such that $f(1) = 0$; it is defined as

$$f(r) = \frac{M - A\psi_2^\star - \psi_1^\star}{\langle r^N \rangle^\star} r^N + A\psi_2(r) + \psi_1(r),$$
$$A = \frac{M - \psi_1^\star}{\psi_2^\star},$$

and A is well defined because ψ_2 is negative and so $\psi_2^\star \neq 0$.

Consequently, the closure problems (C.1) for an homogeneous Dirichlet boundary condition are well posed.

Homogeneous Neumann case. We consider the solution f as in (C.8) of (C.4) and (C.5) and search for a value of A such that $\partial_r f(1) = 0$.

By multiplying (C.4) by r^{N+1} and integrating over $[0, 1]$ one gets

$$\partial_r f(1) = N f(1) + \frac{A}{N+2} + \int_0^1 r^{N+1} g(r)\, dr, \tag{C.10}$$

and since

$$f(1) = \frac{M - A\psi_2^\star - \psi_1^\star}{\langle r^N \rangle^\star},$$

there is only one solution defined as

$$f(r) = \frac{M - A\psi_2^\star - \psi_1^\star}{\langle r^N \rangle^\star} r^N + A\psi_2(r) + \psi_1(r),$$
$$A\left(\frac{1}{N+2} - N\frac{\psi_2^\star}{\langle r^N \rangle^\star}\right) = N\frac{\psi_1^\star - M}{\langle r^N \rangle^\star} - \int_0^1 r^{N+1} g(r)\, dr,$$

where A is well defined because ψ_2 is negative and so $\frac{1}{N+2} - N\frac{\psi_2^\star}{\langle r^N \rangle^\star} \neq 0$.

Consequently the closure problems (C.1) for an homogeneous Neumann boundary condition are well posed.

Robin case. We consider the solution f as in (C.8) of (C.4) and (C.5) and search for a value of A such that $\partial_r f(1) + \gamma f(1) = 0$ for $\gamma > 0$.

With (C.9) and (C.10) we have

$$\partial_r f(1) + \gamma f(1) = A\left(\frac{1}{N+2} - (N+\gamma)\frac{\psi_2^\star}{\langle r^N \rangle^\star}\right) - (N+\gamma)\frac{\psi_1^\star - M}{\langle r^N \rangle^\star} + \int_0^1 r^{N+1} g(r)\, dr,$$

and so there is only one solution defined as

$$f(r) = \frac{M - A\psi_2^\star - \psi_1^\star}{\langle r^N \rangle^\star} r^N + A\psi_2(r) + \psi_1(r),$$
$$A\left(\frac{1}{N+2} - (N+\gamma)\frac{\psi_2^\star}{\langle r^N \rangle^\star}\right) = (N+\gamma)\frac{\psi_1^\star - M}{\langle r^N \rangle^\star} - \int_0^1 r^{N+1} g(r)\, dr,$$

where A is well defined for $\gamma \geq 0$.

Consequently the closure problems (C.1) for an homogeneous Neumann boundary condition are well posed.

REFERENCES

[1] R. Aris, *On the dispersion of a solute in a fluid flowing through a tube*, Proc. Roy. Soc. Ser. A, 235 (1956), pp. 65–77.

[2] H. Arkin, L. X. Xu, and K. R. Holmes, *Recent developments in modeling heat transfer in blood perfused tissues*, IEEE Trans. Biomed. Engrg., 41 (1994), pp. 97–107.

[3] J. W. Baish, P. S. Ayyaswamy, and K. R. Foster, *Heat transport mechanisms in vascular tissues: A model comparison*, J. Biomech. Engrg., 108 (1986), pp. 324–331.

[4] V. Balakotaiah and H. C. Chang, *Dispersion of chemical solutes in chromatographs and reactors*, Proc. Trans. Roy. Soc. Lond. Ser. A, 351 (1995), pp. 39–75.

[5] V. Balakotaiah and H. C. Chang, *Hyperbolic homogenized models for thermal and solutal dispersion*, SIAM J. Appl. Math., 63 (2003), pp. 1231–1258.

[6] J. Bensoussan, L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structure*, North-Holland, Amsterdam, 1978.

[7] A. Berman, *Laminar flow in channels with porous walls*, J. Appl. Phys., 24 (1953), pp. 1232–1235.

[8] A. Bourgeat, M. Quintard, and S. Whitaker, *Eléments de comparaison entre la méthode d'homogénéisation et la méthode de prise de moyenne avec fermeture*, C. R. Acad. Sci. Paris Sér. II Méc. Phys. Chim. Sci. Univers Sci. Terre, 306 (1988), pp. 463–466.

[9] H. Brenner, *Dispersion resulting from flow through spatially periodic porous media*, Philos. Trans. Roy. Soc. London Ser. A, 297 (1980), pp. 81–133.

[10] M. D. Bryden and H. Brenner, *Multiple-timescale analysis of taylor dispersion in converging and diverging flows*, J. Fluid Mech., 311 (1996), pp. 343–359.

[11] S. Chakraborty and V. Balakotaiah, *Low-dimensional models for describing mixing effects in laminar flow turbulent reactors*, Chem. Engrg. Sci., 57 (2002), pp. 2545–2564.

[12] J. H. Cushman, *On unifying the concepts of scale, instrumentation and stochastics in the development of multiple phase transport theory*, Water Res. Resour., 20 (1984), pp. 1668–1676.

[13] J. H. Cushman, L. S. Bennethum, and B. X. Hu, *A primer on tools for porous media*, Adv. Water Resour., 25 (2002), pp. 1043–1067.

[14] W. Deen, *Analysis of Transport Phenomena*, Oxford University Press, London, 1947.

[15] Y. C. Fung, *Biomechanics: Mechanical Properties of Living Tissues*, 2nd ed., Springer, New York, 1996.

[16] F. Golfier, M. Quintard, and S. Whitaker, *Heat and mass transfer in tubes: An analysis using the method of volume averaging*, J. Por. Media, 5 (2002), pp. 169–185.

[17] L. Graetz, *On the thermal conductivity of liquids*, Ann. Phys. Chem., 18 (1883), pp. 79–94.

[18] S. Hassanizadeh and W. Gray, *General conservation equations for multi-phase systems* 1 *averaging procedure*, Adv. Water Resour., 2 (1979), pp. 131–144.

[19] S. Kakac, R. K. Shah, and A. E. Bergles, *Low Reynolds Number Flow Heat Exchangers*, Hemisphere Publishing, Washington, D.C., 1983.

[20] M. A. Lévêque, *Les lois de transmission de la chaleur par convection*, Annales des Mines, Paris, 13 (1928), pp. 201–409.

[21] E. M. Lungu and H. K. Moffat, *The effect of wall conductance on heat diffusion in duct flow*, J. Engrg. Math., 16 (1982), pp. 121–136.

[22] C. M. Marle, *On macroscopic equations governing multiphase flow with diffusion and chemical reactions in porous media*, Internat. J. Engrg. Sci., 20 (1982), pp. 643–662.

[23] C. Mei, J. L. Auriault, and C. Ng, *Some applications of the homogenization theory*, Adv. Appl. Mech., 32 (1996), pp. 278–348.

[24] G. N. Mercer and A. J. Roberts, *A center manifold description of contaminant dispersion in channels with varying flow properties*, SIAM J. Appl. Math., 50 (1990), pp. 1547–1565.

[25] M. Quintard and S. Whitaker, *Transport in ordered and disordered porous media: Volume-averaged equations, closure problems and comparison with experiment*, Chem. Engrg. Sci., 48 (1993), pp. 2537–2564.

[26] A. Nakayama, F. Kuwahara, A. Naoki, and G. Xu, *A volume averaging theory and its sub-control-volume model for analyzing heat and fluid flow within complex heat transfer equipment*, in Proceedings of the 12th International Heat Transfer Conference, Vol. 2, J. Taine, ed., Grenoble, Elsevier, Paris, 2002, pp. 851–856.

[27] A. Nakayama, F. Kuwahara, M. Sugiyama, and G. Xu, *A two-energy equation model for conduction and convection in porous media*, Int. J. Heat Mass Transfer, 44 (2001), pp. 4375–4379.
[28] D. A. Nelson, *Invited editorial on "Pennes'* 1948 *paper revisited,"* J. Appl. Physiol., 85 (1998), pp. 2–3.
[29] M. Pedras and M. D. Lemos, *Macroscopic turbulence modeling for incompressible flow through undeformable porous media*, Int. J. Heat Mass Transfer, 44 (2001), pp. 1081–1093.
[30] H. H. Pennes, *Analysis of tissue and arterial blood temperatures in the resting human forearm*, J. Appl. Physiol., 1 (1948), pp. 93–122.
[31] C. G. Phillips, S. R. Kaye, and C. D. Robinson, *Time-dependent transport by convection and diffusion with exchange between two phases*, J. Fluid Mech., 297 (1995), pp. 373–401.
[32] M. Quintard and S. Whitaker, *Convection, dispersion, and interfacial transport of contaminants: Homogeneous porous media*, Adv. Water Resour., 17 (1994), pp. 221–239.
[33] A. J. Roberts, *The utility of an invariant manifold description of the evolution of a dynamical system*, SIAM J. Appl. Math., 20 (1989), pp. 1447–1458.
[34] S. Rosencrans, *Taylor dispersion in curved channels*, SIAM J. Appl. Math., 57 (1997), pp. 1216–1241.
[35] S. L. Ross, *Differential Equations*, Blaisdell, London, 1964.
[36] E. Sanchez-Palencia, *Nonhomogeneous Media and Vibration Theory*, Lecture Notes in Phys. 127, Springer, New York, 1980.
[37] R. K. Shah and A. L. London, *Laminar flow forced convection in ducts*, Adv. Heat Trans., Suppl. 1 (1978).
[38] G. I. Taylor, *Dispersion of solute matter in solvent flowing slowly through a tube*, Proc. Roy. Soc. Ser. A., 219 (1953), pp. 186–203.
[39] S. Whitaker, *The Method of Volume Averaging*, Kluwer Academic, Norwell, MA, 1999.
[40] W. R. Young and S. Jones, *Shear dispersion*, Phys. Fluid, 3 (1991), pp. 1087–1101.
[41] Z.-G. Yuan, W. H. Somerton, and K. S. Udell, *Thermal dispersion in thick-walled tubes as a model of porous media*, Int. J. Heat Mass Transfer, 34 (1991), pp. 2715–2726.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 153–173

SHAPE ANALYSIS OF AN ADAPTIVE ELASTIC ROD MODEL*

ISABEL N. FIGUEIREDO†, CARLOS F. LEAL†, AND CECÍLIA S. PINTO‡

Abstract. We analyze the shape semiderivative of the solution to an asymptotic nonlinear adaptive elastic rod model, derived in Figueiredo and Trabucho [*Math. Mech. Solids*, 9 (2004), pp. 331–354], with respect to small perturbations of the cross section. The rod model is defined by generalized Bernoulli–Navier elastic equilibrium equations and an ordinary differential equation with respect to time. Taking advantage of the model's special structure and the regularity of its solution, we compute and completely identify, in an appropriate functional space involving time, the weak shape semiderivative.

Key words. adaptive elasticity, rod, shape derivative

AMS subject classifications. 74B20, 74K10, 74L15, 90C31

DOI. 10.1137/040604443

1. Introduction.

In this paper we consider a sensitivity analysis problem in shape optimization: the calculus of the derivative of the solution to an asymptotic nonlinear adaptive elastic rod model, with respect to shape variations of the cross section of the rod. More precisely, for each small parameter $s \in [0, \delta]$ we define a perturbed adaptive elastic rod $\overline{\Omega}_s = \overline{\omega}_s \times [0, L]$. The scalar $L > 0$ is its length, and $\omega_s = \omega + s\,\theta(\omega)$ is a perturbation of a fixed cross section ω, in the direction of the vector field $\theta = (\theta_1, \theta_2)$, that realizes the shape variation. To each rod $\overline{\Omega}_s$ we associate the corresponding unique solution (u^s, d^s) of the asymptotic adaptive elastic rod model, derived in Figueiredo and Trabucho [7]. The purpose of this paper is to compute the limit $(\frac{u^s-u}{s}, \frac{d^s-d}{s})$, when $s \to 0^+$, where (u, d) is the solution's rod model for the case $s = 0$. This limit is the semiderivative of the shape function $J : s \in [0, \delta] \to J(\Omega_s) = (u^s, d^s)$, at $s = 0$ in the direction of the vector field θ (in the sense of Delfour and Zolésio [5, p. 289]), or equivalently, the material derivative of the map J at $s = 0$ (in the sense of Haslinger and Mäkinen [8, p. 111]).

The difficulties that arise in the computation of the limit $(\frac{u^s-u}{s}, \frac{d^s-d}{s})$, when $s \to 0^+$, are caused by the complicated form of the asymptotic adaptive elastic rod model derived in Figueiredo and Trabucho [7]. In fact, this is a simplified adaptive elastic model, proposed for the mathematical modeling of the physiological process of bone remodeling. It couples the generalized Bernoulli–Navier elastic equilibrium equations with an ordinary differential equation with respect to time, which is the remodeling rate equation. This latter equation expresses the process of absorption and deposition of bone material due to external stimulus (cf. Cowin and Hegedus [3] and Hegedus and Cowin [9] for a description of the theory of adaptive elasticity, Cowin and Nachlinger [4] for uniqueness results, and Monnier and Trabucho [10] for existence results of three-dimensional solutions). For each $s \in [0, \delta]$, the pair (u^s, d^s) is the unique solution of this asympotic adaptive elastic rod model, where u^s is the displacement vector field

*Received by the editors February 25, 2004; accepted for publication (in revised form) May 10, 2005; published electronically October 17, 2005. This work is part of the European project HRN-CT-2002-00284 and is partially supported by the project FCT–POCTI/MAT/59502/2004 of Portugal.
http://www.siam.org/journals/siap/66-1/60444.html

†Departamento de Matemática, Universidade de Coimbra, Apartado 3008, 3001-454 Coimbra, Portugal (Isabel.Figueiredo@mat.uc.pt, http://www.mat.uc.pt/∼isabelf, carlosl@mat.uc.pt).

‡Departamento de Matemática, Escola Superior de Tecnologia de Viseu, Campus Politécnico, 3504-510 Viseu, Portugal (cagostinho@mat.estv.ipv.pt).

of the rod $\overline{\Omega}_s$ and d^s is a scalar field that represents the change in volume fraction of the elastic material (from a reference volume fraction) in the rod $\overline{\Omega}_s$. Moreover, u^s is the solution of the generalized Bernoulli–Navier equilibrium equations, and d^s is the solution of the remodeling rate equation. In addition, u^s and d^s are coupled in the model because the material coefficients depend on d^s and the remodeling rate equation depends on u^s.

In spite of this complex structure, we are able to compute the limit $(\frac{u^s-u}{s}, \frac{d^s-d}{s})$ when $s \to 0^+$. There are two main results in this paper, which lead to this limit's computation. The first principal result states that, for each time t, the sequence $(\frac{u^s-u}{s}, \frac{d^s-d}{s})(.,t)$ converges weakly to $(\bar{u}, \bar{d})(.,t)$, when $s \to 0^+$, in an appropriate functional space of Sobolev type. (We denote by (u^s, d^s) the solution of the perturbed rod model, formulated in the unperturbed fixed domain $\overline{\Omega}$, which is the domain of (u, d).) The second main result identifies the weak shape semiderivative denoted by $(\bar{u}, \bar{d})$; it is the unique solution of a nonlinear problem which couples a variational equation (whose solution is $\bar{u}$) and depends on (u, d) and $\bar{d}$, and an ordinary differential equation with respect to time (whose solution is $\bar{d}$) that depends on (u, d) and $\bar{u}$.

The reasonings that we have used to achieve these two results are next summarized. We show that the sequences (u^s, d^s) and $(\frac{u^s-u}{s}, \frac{d^s-d}{s})$ are bounded in appropriate functional spaces, involving time; we use the continuity, the ellipticity, the regularity properties, and the special structure of the asymptotic adaptive elastic rod model. In order to identify the weak shape semiderivative we also apply the weak and/or strong convergence of the sequences $\{u^s\}$ and $\{d^s\}$, when $s \to 0^+$, and again the special structure of the asymptotic adaptive elastic rod model. In particular, due to the form of the remodeling rate equation, we are able to use the integral Gronwall's inequality, which is the key to obtaining the estimates for the sequences $\{d^s\}$ and $\{\frac{d^s-d}{s}\}$ and to identifying the ordinary differential equation with respect to time, whose solution is $\bar{d}$.

Finally let us briefly explain the contents of the paper. After this introduction, in section 2, we describe the problem P_s, which is the asymptotic nonlinear adaptive elastic model for the perturbed rod $\overline{\Omega}_s$; we also prove a regularity property of its solution, and finally we describe the shape problem that we want to solve. In section 3 we reformulate the problem P_s in the unperturbed domain $\overline{\Omega}$; this reformulation is necessary because, in order to compute the limit of the quotient sequence $(\frac{u^s-u}{s}, \frac{d^s-d}{s})$, the vector fields u^s, u, d^s, and d must be defined in the same fixed domain, independent of s. In section 4 we prove that all the sequences $\{u^s\}$, $\{d^s\}$, $\{\frac{u^s-u}{s}\}$, and $\{\frac{d^s-d}{s}\}$ are bounded in appropriate functional spaces involving time; we determine, for each time t, the weak limit of the quotient sequence $(\frac{u^s-u}{s}, \frac{d^s-d}{s})(.,t)$ when $s \to 0^+$; and we identify the weak shape semiderivative (this identification is summarized in theorem 4.11). Finally we present some conclusions and future work.

2. Description of the problem. In this section we first introduce the notation used in this paper; namely, we consider a family of rods $\overline{\Omega}_s = \overline{\omega}_s \times [0, L]$, with length L and cross section ω_s, parameterized by $s \in [0, \delta]$, which is a small parameter. Next, for each s, we describe the adaptive elastic rod model denoted by P_s, derived by Figueiredo and Trabucho [7]. We prove a regularity result for the displacement vector field u_s, the first component of the solution (u_s, d_s) of P_s. Finally, we describe the shape problem under consideration in this paper.

2.1. Notation. Let $\delta > 0$ be a small parameter, and for each $s \in [0, \delta]$ we consider the perturbation I_s of the identity operator I in $\mathbb{R}^2$, defined by $I_s(x_1, x_2) = (I + s\theta)(x_1, x_2) = (x_{s1}, x_{s2})$, for all $(x_1, x_2) \in \mathbb{R}^2$, where $\theta = (\theta_1, \theta_2) : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is a

vector field regular enough (at least $\theta \in [W^{2,\infty}(\mathbb{R}^2)]^2$). Let ω be an open, bounded, and connected subset of $\mathbb{R}^2$, with a boundary $\partial\omega$ regular enough. For each $s \in [0,\delta]$ we define $\omega_s = I_s(\omega)$, which is the perturbation of ω in the direction of the vector field θ. We also denote by $\overline{\Omega}_s$ the set occupied by a cylindrical adaptive elastic rod, in its reference configuration, with length $L > 0$ and cross section ω_s, that is, $\overline{\Omega}_s = \overline{\omega}_s \times [0,L] = I_s(\overline{\omega}) \times [0,L] \subset \mathbb{R}^3$. Moreover, we denote by $x_s = (x_{s1}, x_{s2}, x_3)$ a generic element of $\overline{\Omega}_s$ and define the sets $\Gamma_s = \partial\omega_s \times]0,L[$, $\Gamma_{s0} = \overline{\omega}_s \times \{0\}$, $\Gamma_{sL} = \overline{\omega}_s \times \{L\}$, and $\Gamma_{s0,L} = \Gamma_{s0} \cup \Gamma_{sL}$, where $\partial\omega_s$ is the boundary of ω_s. These last four sets represent, respectively, the lateral boundary of the rod $\overline{\Omega}_s$ and its extremities. We assume that, for each $s \in [0,\delta]$, the coordinate system (O, x_{s1}, x_{s2}, x_3) is a principal system of inertia associated with the rod Ω_s. Consequently, axis Ox_3 passes through the centroid of each section $\omega_s \times \{x_3\}$, and we have $\int_{\omega_s} x_{s1}\, d\omega_s = \int_{\omega_s} x_{s2}\, d\omega_s = \int_{\omega_s} x_{s1}x_{s2}\, d\omega_s = 0$. (We observe that the choice of the vector field θ, which realizes the shape variation of the cross section ω, must be admissible with this condition.)

The set $C^m(\overline{\Omega}_s)$ is the space of real functions m times continuously differentiable in $\overline{\Omega}_s$. The spaces $W^{m,q}(\Omega_s)$ and $W^{0,q}(\Omega_s) = L^q(\Omega_s)$ are the usual Sobolev spaces, where q is a real number satisfying $1 \le q \le \infty$ and m is a positive integer. The norms in these Sobolev spaces are denoted by $\|.\|_{W^{m,q}(\Omega_s)}$. The set $\mathcal{R}_s = \{v_s \in \mathbb{R}^3 :\ v_s = a + b \wedge x_s,\ a, b \in \mathbb{R}^3\}$, where $\wedge$ is the exterior product in $\mathbb{R}^3$, is the set of infinitesimal rigid displacements. We denote by $[W^{m,q}(\Omega_s)]^3/\mathcal{R}_s$ the quotient space induced by the set $\mathcal{R}_s$ in the Sobolev space $[W^{m,q}(\Omega_s)]^3$.

Throughout the paper, the Latin indices i, j, k, $l, \ldots$ belong to the set $\{1,2,3\}$; the Greek indices α, β, $\mu, \ldots$ vary in the set $\{1,2\}$; and the summation convention with respect to repeated indices is employed, that is, for example, $a_i b_i = \sum_{i=1}^3 a_i b_i$.

Let $T > 0$ be a real parameter, and denote by t the time variable in the interval $[0,T]$. If V is a topological vectorial space, the set $C^m([0,T];V)$ is the space of functions $g : t \in [0,T] \to g(t) \in V$ such that g is m times continuously differentiable with respect to t. If V is a Banach space, we denote by $\|.\|_{C^m([0,T];V)}$ the usual norm in $C^m([0,T];V)$. Moreover, given a function $g_s(x_s,t)$ defined in $\overline{\Omega}_s \times [0,T]$, we denote by $\dot{g}_s$ its partial derivative with respect to time and by $\partial_{s\alpha} g_s$ and $\partial_3 g_s$ its partial derivatives with respect to $x_{s\alpha}$ and x_3; that is, $\dot{g}_s = \frac{\partial g_s}{\partial t}$, $\partial_{s\alpha} g_s = \frac{\partial g_s}{\partial x_{s\alpha}}$, and $\partial_3 g_s = \frac{\partial g_s}{\partial x_3}$.

2.2. The adaptive elastic rod model. Figueiredo and Trabucho [7] have applied the asymptotic expansion method to the three-dimensional adaptive elasticity model derived by Cowin and Hegedus [3, 9], with the modifications proposed by Monnier and Trabucho [10], for a thin rod whose cross section is a function of a small parameter and for a three-dimensional remodeling rate equation depending nonlinearly or linearly on the strain tensor field (cf. also Trabucho and Viaño [11] for an explanation of the mathematical modeling of rods with the asymptotic expansion method). They have obtained a simplified adaptive elastic rod model, which is designated in what follows by the asymptotic adaptive elastic rod model. This is a system of nonlinear coupled equations, which includes generalized Bernoulli–Navier equilibrium equations and a simplified remodeling rate equation. For any perturbed rod $\overline{\Omega}_s$ (with $s \in [0,\delta]$) and for the case where the original three-dimensional remodeling rate equation depends linearly on the strain tensor field, this system is defined as follows:

$$
\left[\begin{array}{l}
u_s = (u_{s1}, u_{s2}, u_{s3}) : \overline{\Omega}_s \times [0,T] \to \mathbb{R}^3, \quad d_s : \overline{\Omega}_s \times [0,T] \to \mathbb{R}, \\
u_{s\alpha} : [0,L] \times [0,T] \to \mathbb{R}, \\
u_{s3} = \underline{u}_{s3} - x_{s\alpha}\partial_3 u_{s\alpha} \quad \text{and} \quad \underline{u}_{s3} : [0,L] \times [0,T] \to \mathbb{R}, \\
(u_s, d_s) \quad \text{satisfies the following:}
\end{array}\right. \tag{2.1}
$$

$$
\left[\begin{array}{l}
\textit{Equilibrium equations} \text{ in } (0,L)\times(0,T)\\
\left[\begin{array}{l}
-\partial_3\big(l_s(d_s)\partial_3\underline{u}_{s3} - e_{s\alpha}(d_s)\partial_{33}u_{s\alpha}\big)\\
\quad = \int_{\omega_s}\gamma_s(\xi_{s0}+P_\eta(d_s))\,f_{s3}\,d\omega_s + \int_{\partial\omega_s} g_{s3}\,d\partial\omega_s,
\end{array}\right.\\
\left[\begin{array}{l}
\partial_{33}\big(-e_{s\beta}(d_s)\partial_3\underline{u}_{s3} + h_{s\alpha\beta}(d_s)\partial_{33}u_{s\alpha}\big)\\
= \int_{\omega_s}\gamma_s(\xi_{s0}+P_\eta(d_s))\,f_{s\beta}\,d\omega_s + \int_{\partial\omega_s} g_{s\beta}\,d\partial\omega_s\\
+\int_{\omega_s} x_{s\beta}\,\partial_3[\gamma_s(\xi_{s0}+P_\eta(d_s))\,f_{s3}]\,d\omega_s + \int_{\partial\omega_s} x_{s\beta}\,\partial_3 g_{s3}\,d\partial\omega_s, \qquad \beta=1,2,
\end{array}\right.
\end{array}\right. \tag{2.2}
$$

$$
\left[\begin{array}{l}
\textit{Boundary conditions} \text{ for } \{\overline{x}_3\}\times(0,T), \text{ with } \overline{x}_3 = 0, L\\
\big(l_s(d_s)\partial_3\underline{u}_{s3} - e_{s\alpha}(d_s)\partial_{33}u_{s\alpha}\big)(\overline{x}_3) = \int_{\omega_s} h_{s3}(\overline{x}_3)\,d\omega_s,\\
\big(e_{s\beta}(d_s)\partial_3\underline{u}_{s3} - h_{s\alpha\beta}(d_s)\partial_{33}u_{s\alpha}\big)(\overline{x}_3) = \int_{\omega_s} x_{s\beta}\,h_{s3}(\overline{x}_3)\,d\omega_s,\\
\left[\begin{array}{l}
\partial_3\big(e_{s\beta}(d_s)\partial_3\underline{u}_{s3} - h_{s\alpha\beta}(d_s)\partial_{33}u_{s\alpha}\big)(\overline{x}_3)\\
\quad = \int_{\omega_s} h_{s\beta}(\overline{x}_3)\,d\omega_s - \int_{\partial\omega_s} x_{s\beta}g_{s3}(\overline{x}_3)\,d\partial\omega_s\\
\qquad -\int_{\omega_s} x_{s\beta}\,\gamma_s(\xi_{s0}+P_\eta(d_s))f_{s3}(\overline{x}_3)\,d\omega_s,
\end{array}\right.
\end{array}\right. \tag{2.3}
$$

$$
\left[\begin{array}{ll}
\textit{Remodeling rate equation} & \\
\dot{d}_s = c(d_s)e_{33}(u_s) + a(d_s) & \text{in} \quad \Omega_s\times(0,T),\\
c(d_s) = A_{\alpha\beta}(d_s)\frac{b_{\alpha\beta33}(d_s)}{b_{3333}(d_s)} + A_{33}(d_s), & \\
d_s(x_s,0) = \hat{d}_s(x_s) & \text{in} \quad \overline{\Omega}_s.
\end{array}\right. \tag{2.4}
$$

The unknowns of the model (2.1)–(2.4) are the displacement vector field $u_s(x_s,t)$, corresponding to the displacement of the point x_s of the rod $\overline{\Omega}_s$ at time t, and the measure of change in volume fraction of the elastic material (from the reference volume fraction ξ_{s0}) $d_s(x_s,t)$ at (x_s,t). In particular, $e_{33}(u_s) = \partial_3 u_{s3} = \partial_3\underline{u}_{s3} - x_{s\alpha}\partial_{33}u_{s\alpha}$ is an element of the linear strain tensor field $(e_{ij}(u_s))$, which depends on u_s.

On the other hand, the data of the model (2.1)–(2.4) are the following: the open set $\Omega_s\times(0,T)$; the density $\gamma_s = \gamma$ of the full elastic material, which is supposed to be a constant independent of s; the reference volume fraction of the elastic material ξ_{s0}, which belongs to $C^1(\overline{\Omega}_s)$ for each s; the body load $f_s = (f_{si})$ such that $f_{si}\in C^1([0,T])$ and depends only on t; the normal tractions on the boundary $g_s = (g_{si})$ and $h_s = (h_{si})$; the initial value of the change in volume fraction $\hat{d}_s$, which belongs to $C^0(\overline{\Omega}_s)$; the truncation operator $\mathcal{P}_\eta(.)$; and the coefficients $l_s(d_s)$, $e_{s\alpha}(d_s)$, $h_{s\alpha\beta}(d_s)$, $c(d_s)$, $a(d_s)$, $A_{\alpha\beta}(d_s)$, $A_{33}(d_s)$, $b_{\alpha\beta33}(d_s)$, and $b_{3333}(d_s)$, which are all material coefficients depending upon the change in volume fraction d_s.

On these data we also suppose further conditions, which we will describe next. We assume that, for each $s\in[0,\delta]$, $0<\xi_{s0}^{min}\le\xi_{s0}(x_s)\le\xi_{s0}^{max}<1$ and the normal tractions verify

$$
g_{si}\in C^1([0,T];W^{1-1/p,p}(\Gamma_s)), \quad h_{si}\in C^1([0,T];W^{1-1/p,p}(\Gamma_{s0}\cup\Gamma_{sL})), \tag{2.5}
$$

with $p>3$. In addition, we assume that the resultant of the system of applied forces is null for rigid displacements; this means that, for any $v_s = (v_{si})$ in $\mathcal{R}_s$ and for all $t\in[0,T]$,

$$
\int_{\Omega_s}\gamma(\xi_{s0}+\mathcal{P}_\eta(d_s))f_{si}v_{si}dx_s + \int_{\Gamma_s} g_{si}v_{si}d\Gamma_s + \int_{\Gamma_{s0,L}} h_{si}v_{si}d\Gamma_{s0,L} = 0. \tag{2.6}
$$

The truncation operator $\mathcal{P}_\eta$ is of class C^1 and satisfies $0 < \frac{\eta}{2} \leq (\xi_{s0} + \mathcal{P}_\eta(d_s))(x_s) \leq 1$ for all $x_s \in \overline{\Omega}_s$, where $\eta > 0$ is a small parameter.

The coefficients $b_{\alpha\beta 33}(d_s)$ and $b_{3333}(d_s)$ are continuously differentiable with respect to d_s and are elements of the matrix $(b_{ijkl}(d_s))$, which is the inverse of the matrix defined by the three-dimensional elastic coefficients $(c_{ijkl}(d_s))$ of the rod $\overline{\Omega}_s$, that depend on d_s through truncation and mollification (cf. formulas (47)–(48), Figueiredo and Trabucho [7]). The coefficients $A_{\alpha\beta}(d_s)$, $A_{33}(d_s)$, $c(d_s)$, and $a(d_s)$ are remodeling rate coefficients and are continuously differentiable with respect to d_s.

Moreover, $b_{\alpha\beta 33}(d_s)$ and $b_{3333}(d_s)$ belong to the space $C^1([0,T]; C^1(\mathbb{R}^3))$ when $d_s \in C^1([0,T]; C^0(\overline{\Omega}_s))$ (cf. Monnier and Trabucho [10, p. 542] and also formulas (47)–(48) of Figueiredo and Trabucho [7]). In addition we also assume that there exist strictly positive constants C_1, C_2, C_3, C_4, C_5, and C_6 independent of s and t such that for any $(x_s, t) \in \overline{\Omega}_s \times [0,T]$

$$0 < C_1 \leq \frac{1}{b_{3333}(d_s)} \leq C_2 \quad \forall s \in [0,\delta], \tag{2.7}$$

$$|c(d_s)| \leq C_3,\ |a(d_s)| \leq C_4,\ |c'(d_s)| \leq C_5,\ |a'(d_s)| \leq C_6 \quad \forall s \in [0,\delta], \tag{2.8}$$

where $c'(.)$ and $a'(.)$ are the derivatives of the scalar functions $c(.)$ and $a(.)$. We remark that the assumption (2.7) is a direct consequence of the definition of b_{3333}, and also a consequence of [10, Lemma 1, p. 542]. The assumption (2.8) can be proven using exactly the same arguments of this Lemma 1, supposing that $c(d_s)$ and $a(d_s)$ depend on d_s through truncation and mollification.

The coefficients $l_s(d_s)$, $e_{s\alpha}(d_s)$, and $h_{s\alpha\beta}(d_s)$, which depend on $b_{3333}(d_s)$ (cf. formula (49), Figueiredo and Trabucho [7]), are functions of x_3 and t, and are defined by

$$l_s = \int_{\omega_s} \frac{1}{b_{3333}(d_s)} d\omega_s, \quad e_{s\alpha} = \int_{\omega_s} \frac{x_{s\alpha}}{b_{3333}(d_s)} d\omega_s, \quad h_{s\alpha\beta} = \int_{\omega_s} \frac{x_{s\alpha}x_{s\beta}}{b_{3333}(d_s)} d\omega_s. \tag{2.9}$$

The variational formulation of the equilibrium equations (2.2) is obtained by multiplying the first equilibrium equation (2.2) by $\underline{v}_{s3} \in W^{1,2}(0,L)$ and the second and third equilibrium equations by $x_{s1}\, v_{s1}$ and $x_{s2}\, v_{s2}$, respectively, with $v_{s\beta} \in W^{2,2}(0,L)$, for $\beta = 1, 2$, and subsequently integrating in $(0,L)$ and using the boundary conditions (2.3). Thus, the asymptotic adaptive elastic rod model (2.1)–(2.4) is equivalent to the following nonlinear (variational and differential) system (P_s) (cf. formula (56) of Figueiredo and Trabucho [7]):

$$(\mathrm{P}_s) \quad \left[\begin{array}{l} \text{Find} \quad u_s : \overline{\Omega}_s \times [0,T] \to \mathbb{R}^3 \quad \text{and} \quad d_s : \overline{\Omega}_s \times [0,T] \to \mathbb{R} : \\ u_s(.,t) \in V(\Omega_s)/\mathcal{R}_s, \\ a_s(u_s, v_s) = L_s(v_s) \qquad \forall v_s \in V(\Omega_s)/\mathcal{R}_s, \\ \dot{d}_s = c(d_s) e_{33}(u_s) + a(d_s) \quad \text{in} \quad \Omega_s \times (0,T), \\ d_s(x_s, 0) = \hat{d}_s(x_s) \quad \text{in} \quad \overline{\Omega}_s. \end{array}\right.$$

The space $V(\Omega_s) = \{v_s \in [W^{1,2}(\Omega_s)]^3 :\ e_{\alpha\beta}(v_s) = e_{3\beta}(v_s) = 0\}$ is identified with

$$\begin{aligned} \Big\{ v_s = (v_{s1}, v_{s2}, v_{s3}) \in [W^{2,2}(0,L)]^2 \times W^{1,2}(\Omega_s) :\quad & v_{s\alpha}(x_s) = v_{s\alpha}(x_3), \\ & v_{s3}(x_s) = \underline{v}_{s3}(x_3) - x_{s\alpha}\partial_3 v_{s\alpha}(x_3), \quad \underline{v}_{s3} \in W^{1,2}(0,L) \Big\}, \end{aligned} \tag{2.10}$$

and the quotient space $V(\Omega_s)/\mathcal{R}_s$ is the following set:

$$\Big\{v_s = z_s + a + b \wedge x_s : \;\; z_s \in V(\Omega_s),\; a \in \mathbb{R}^3,\;\; b = (b_1, b_2, 0) \in \mathbb{R}^3\Big\}. \tag{2.11}$$

The bilinear form $a_s(.,.)$, depending on the unknown d_s, is defined in $V(\Omega_s)/\mathcal{R}_s$ by

$$a_s(z_s, v_s) = \int_{\Omega_s} \frac{1}{b_{3333}(d_s)} e_{33}(z_s) e_{33}(v_s) d\Omega_s \quad \forall z_s,\; v_s \in V(\Omega_s)/\mathcal{R}_s, \tag{2.12}$$

and $L_s(.)$ is a linear form also defined in $V(\Omega_s)/\mathcal{R}_s$ such that $L_s(v_s)$ is equal to

$$\int_{\Omega_s} \gamma(\xi_{s0} + P_\eta(d_s))\, f_{si}\, v_{si}\, d\Omega_s + \int_{\Gamma_s} g_{si}\, v_{si}\, d\Gamma_s + \int_{\Gamma_{s0,L}} h_{si}\, v_{si}\, d\Gamma_{s0,L}. \tag{2.13}$$

We remark that in (2.11) we must have $b_3 = 0$, because otherwise the quotient space $V(\Omega_s)/\mathcal{R}_s$ would not be contained in $V(\Omega_s)$. In fact, developing $v_s = z_s + a + b \wedge x_s$, we have for the first component $v_{s1} = z_{s1} + a_1 + b_2 x_3 - b_3 x_{s2}$, for the second component $v_{s2} = z_{s2} + a_2 - b_1 x_3 + b_3 x_{s1}$, and finally for the third component $v_{s3} = \underline{z}_{s3} - x_{s\alpha}\partial_3 z_{s\alpha} + a_3 + b_1 x_{s2} - b_2 x_{s1}$. Therefore if $b_3 \neq 0$, then $v_s \notin V(\Omega_s)$, and if $b_3 = 0$, then $(b_1, b_2, 0) \wedge x_s = \big(b_2 x_3, -b_1 x_3, b_1 x_{s2} - b_2 x_{s1}\big)$ and we obtain $v_{s1} = z_{s1} + a_1 + b_2 x_3$, which depends only on x_3, $v_{s2} = z_{s2} + a_2 - b_1 x_3$, which depends only on x_3, and $v_{s3} = \underline{v}_{s3} - x_{s\alpha}\partial_3 v_{s\alpha}$ with $\underline{v}_{s3} = \underline{z}_{s3} + a_3$, which depends only on x_3.

By the following Korn-type inequality in the quotient space $V(\Omega_s)/\mathcal{R}_s$ (cf. Ciarlet [1] or Valent [12]) we have

$$\exists c > 0 : \quad \|v_s\|^2_{[W^{1,2}(\Omega_s)]^3} \leq c \|e_{33}(v_s)\|^2_{L^2(\Omega_s)}, \tag{2.14}$$

where

$$\|e_{33}(v_s)\|^2_{L^2(\Omega_s)} = c_s \|\partial_3 \underline{v}_{s3}\|^2_{L^2(0,L)} + \left(\int_{\omega_s} x^2_{s\alpha} d\omega_s\right) \|\partial_{33} v_{s\alpha}\|^2_{L^2(0,L)}, \tag{2.15}$$

with $c_s = [\mathrm{meas}(\omega_s)]^{\frac{1}{2}}$, since $e_{33}(v_s) = \partial_3 v_{s3} = \partial_3 \underline{v}_{s3} - x_{s\alpha}\partial_{33} v_{s\alpha}$. Hence, we conclude that $\|e_{33}(.)\|_{L^2(\Omega_s)}$ is a norm in the space $V(\Omega_s)/\mathcal{R}_s$, equivalent to the usual norm induced in the quotient space by $\|.\|_{[W^{1,2}(\Omega_s)]^3}$. Moreover, $V(\Omega_s)/\mathcal{R}_s$ is a Hilbert space with the norm $\|e_{33}(.)\|_{L^2(\Omega_s)}$, and the bilinear form $a_s(.,.)$ is elliptic in $V(\Omega_s)/\mathcal{R}_s$. In fact, there exists a constant $C > 0$ such that

$$\begin{cases} a_s(v_s, v_s) = \int_{\Omega_s} \frac{1}{b_{3333}(d_s)}\, e_{33}(v_s)\, e_{33}(v_s)\, d\Omega_s \geq C_1 \|e_{33}(v_s)\|^2_{L^2(\Omega_s)} \\ = C_1 \|v_s\|^2_{V(\Omega_s)/\mathcal{R}_s} \geq C \|v_s\|^2_{[W^{1,2}(\Omega_s)]^3} \quad \forall v_s \in V(\Omega_s)/\mathcal{R}_s, \end{cases} \tag{2.16}$$

where C_1 is the constant defined in condition (2.7).

For each $s \in [0, \delta]$, there exists a unique pair (u_s, d_s) solution of the asymptotic adaptive elastic rod model P_s, which verifies $u_s \in C^1([0,T]; V(\Omega_s)/\mathcal{R}_s)$ and $d_s \in C^1([0,T]; C^0(\overline{\Omega}_s))$ (cf. Theorem 6, Figueiredo and Trabucho [7]). The next theorem states a regularity result, concerning the component solution u_s, that will be important in section 4. In order to prove it, we introduce the following notation:

$$\begin{aligned} & z_{s3} = l_s \partial_3 \underline{u}_{s3} - e_{s\alpha}\partial_{33} u_{s\alpha}, \quad z_{s\beta} = h_{s\alpha\beta}\partial_{33} u_{s\alpha} - e_{s\beta}\partial_3 \underline{u}_{s3}, \\ & F_{s3} = \int_{\omega_s} \gamma(\xi_{s0} + P_\eta(d_s))\, f_{s3}\, d\omega_s + \int_{\partial\omega_s} g_{s3}\, d\partial\omega_s, \\ & F_{s\beta} = \begin{cases} \int_{\omega_s} \gamma(\xi_{s0} + P_\eta(d_s))\, f_{s\beta}\, d\omega_s + \int_{\partial\omega_s} g_{s\beta}\, d\partial\omega_s \\ \quad + \int_{\omega_s} x_{s\beta}\, \partial_3[\gamma(\xi_{s0} + P_\eta(d_s))\, f_{s3}]\, d\omega_s + \int_{\partial\omega_s} x_{s\beta}\, \partial_3 g_{s3}\, d\partial\omega_s, \end{cases} \end{aligned} \tag{2.17}$$

where $z_{si} \in C^1([0,T];L^2(0,L))$, $F_{si} \in C^1([0,T];L^2(0,L))$, for $i = 1,2,3$, and the matrix M_s,

$$M_s = \begin{bmatrix} l_s & -e_{s1} & -e_{s2} \\ -e_{s1} & h_{s11} & h_{s12} \\ -e_{s2} & h_{s21} & h_{s22} \end{bmatrix} \in C^1\big([0,T];[C^1(\mathbb{R}^3)]^9\big). \tag{2.18}$$

THEOREM 2.1 (regularity of u_s). *Let (u_s, d_s) be the unique solution of problem (P_s). We assume that the determinant of matrix M_s is not zero, $\det M_s \neq 0$ (for example, if $b_{3333}(d^s) = c$, where c is a constant, then $\det M_s > 0$). Then, for each $t \in [0,T]$, $u_{s\beta}(.,t) \in W^{3,2}(0,L)$, $\underline{u}_{s3}(.,t) \in W^{2,2}(0,L)$, and consequently $u_s(.,t) \in [W^{2,2}(\Omega_s)]^3$.*

Proof. We first remark that the equilibrium equations (2.2) can be written in the form

$$-\partial_3 z_{s3} = F_{s3}, \qquad \partial_{33} z_{s\beta} = F_{s\beta}, \quad \text{for } \beta = 1,2, \tag{2.19}$$

and for each t, $F_{s3}(.,t)$ and $F_{s\beta}(.,t)$ belong to the space $L^2(0,L)$.

Since $z_{s3}(.,t)$ belongs to $L^2(0,L)$, and because of the first equilibrium equation in (2.19), $\partial_3 z_{s3}(.,t)$ also belongs to $L^2(0,L)$, so we conclude that, for each t, $z_{s3}(.,t) \in W^{1,2}(0,L)$. For each t, $z_{s\beta}(.,t) \in L^2(0,L)$ and consequently $\partial_3 z_{s\beta}(.,t) \in [W^{1,2}(0,L)]'$, where $[W^{1,2}(0,L)]'$ is the dual of $W^{1,2}(0,L)$. But from the second equilibrium equation in (2.19) we have that $\partial_{33} z_{s\beta}(.,t) \in L^2(0,L)$ and also $\partial_{33} z_{s\beta}(.,t) \in [W^{1,2}(0,L)]'$ because $L^2(0,L) \subset [W^{1,2}(0,L)]'$. Thus, as a consequence of a lemma of Lions (cf. Ciarlet [2, p. 39]) we have $\partial_3 z_{s\beta}(.,t) \in L^2(0,L)$. Hence the elements $z_{s\beta}(.,t)$, $\partial_3 z_{s\beta}(.,t)$, and $\partial_{33} z_{s\beta}(.,t)$ belong to the space $L^2(0,L)$, which means that $z_{s\beta}(.,t) \in W^{2,2}(0,L)$.

Therefore, assembling these properties, we obtain for each t and for $\beta = 1,2$

$$z_{s3}(.,t) = p_{s3}(.,t) \in W^{1,2}(0,L), \qquad z_{s\beta}(.,t) = p_{s\beta}(.,t) \in W^{2,2}(0,L), \tag{2.20}$$

where p_{si} is a primitive of F_{si}, in the distribution's sense in $W^{1,2}(0,L)$, for $i = 1,2,3$. Replacing z_{si} by its definition (2.17), the system (2.20) is equivalent to the following system:

$$\begin{bmatrix} l_s & -e_{s1} & -e_{s2} \\ -e_{s1} & h_{s11} & h_{s12} \\ -e_{s2} & h_{s21} & h_{s22} \end{bmatrix} \begin{bmatrix} \partial_3 \underline{u}_{s3} \\ \partial_{33} u_{s1} \\ \partial_{33} u_{s2} \end{bmatrix} = \begin{bmatrix} p_{s3} \\ p_{s1} \\ p_{s2} \end{bmatrix}. \tag{2.21}$$

With the assumption $\det M_s \neq 0$, we clearly obtain, by solving (2.21), that

$$\partial_3 \underline{u}_{s3}(.,t) \in W^{1,2}(0,L), \qquad \partial_{33} u_{s\beta}(.,t) \in W^{1,2}(0,L), \quad \text{for } \beta = 1,2. \tag{2.22}$$

We remark that the regularity indicated in (2.22) depends also on the regularity of the elements of M_s that belong to the space $C^1([0,T];C^1(\mathbb{R}^3))$.

Thus we conclude that the components $u_{s\beta}(.,t) \in W^{3,2}(0,L)$, for $\beta = 1,2$ and $\underline{u}_{s3}(.,t) \in W^{2,2}(0,L)$, and consequently, because $u_s = (u_{s1}, u_{s2}, \underline{u}_{s3} - x_{s\beta}\partial_3 u_{s\beta})$, we have $u_s(.,t) \in [W^{2,2}(\Omega_s)]^3$. ∎

2.3. The shape problem. We now consider the shape map J defined by

$$\begin{array}{rccl} J: & [0,\delta] & \longrightarrow & C^1([0,T];V(\Omega_s)/\mathcal{R}_s) \times C^1([0,T];C^0(\overline{\Omega}_s)) \\ & s & \longrightarrow & J(\Omega_s) = (u_s, d_s), \end{array} \tag{2.23}$$

where (u_s, d_s) is the unique solution of the nonlinear asymptotic adaptive rod model P_s (cf. (P_s)), defined in the perturbed rod Ω_s. As remarked before, the unknowns u_s and d_s are coupled in the model P_s and depend on (x_s, t).

We recall that $I_s(\omega)$ is a shape perturbation of the cross section ω of the rod $\overline{\Omega} = \overline{\omega} \times [0, L]$, so $\overline{\Omega}_s$ is a shape perturbation of the rod $\overline{\Omega}$; that is, $\overline{\Omega}_s = I_s(\overline{\omega}) \times [0, L] = (I + s\,\theta)(\overline{\omega}) \times [0, L]$ and $\overline{\Omega} = \overline{\omega} \times [0, L] = I_0(\overline{\omega}) \times [0, L] = \overline{\Omega}_0$.

The aim of this paper is to compute the shape semiderivative $dJ(\Omega; \theta)$ at $s = 0$ in the direction of the vector field θ. This semiderivative is defined by (cf. Delfour and Zolésio [5, p. 289])

$$dJ(\Omega;\theta) = \lim_{s\to 0^+} \frac{J(\Omega_s) - J(\Omega)}{s} = \left(\lim_{s\to 0^+} \frac{u_s - u}{s},\ \lim_{s\to 0^+} \frac{d_s - d}{s} \right), \tag{2.24}$$

where (u, d) is the solution of problem (P_s) but for the unperturbed rod $\overline{\Omega}_0 = \overline{\Omega} = \overline{\omega} \times [0, L]$. We also remark that the semiderivative $dJ(\Omega; \theta)$ is equivalent to the definition of the material derivative of the map J at $s = 0$, in the sense of Haslinger and Mäkinen [8, p. 111].

As explained in section 4, Theorem 4.11, it is possible to compute and to identify, in a weak sense and in an appropriate product space, this shape semiderivative.

3. Equivalent formulation of the adaptive rod model. In order to be able to calculate the shape semiderivative (2.24) we must reformulate, for each s, the problem P_s (cf. (P_s)) in the domain $\overline{\Omega} \times [0, T]$, independent of s. Therefore we first formulate in Ω all the forms involved in the definition of problem P_s. Afterwards, we describe the resulting rod model, denoted by P^s (with upper index s) and formulated in the fixed domain $\overline{\Omega} \times [0, T]$, that is equivalent to P_s (with lower index s).

3.1. Reformulation of the forms defining $\mathbf{P_s}$. We define the map

$$Q_s(x_1, x_2, x_3) = \big(x_1 + s\theta_1(x_1, x_2), x_2 + s\theta_2(x_1, x_2), x_3\big), \tag{3.1}$$

which verifies

$$\Omega_s = Q_s(\Omega) \quad \text{and} \quad \det \nabla Q_s = 1 + s \operatorname{div} \theta + s^2 \det \nabla\theta, \tag{3.2}$$

where the matrices ∇Q_s and $\nabla\theta$ are the gradients of Q_s and θ, respectively, and $\operatorname{div}\theta = \partial_\alpha \theta_\alpha$ is the divergence of θ.

To each function v_s defined in Ω_s we associate the corresponding function v^s (with upper index s) defined on Ω by $v^s = v_s \circ Q_s$. Hence, for any $v_s \in V(\Omega_s)/\mathcal{R}_s$, the correspondent v^s is in $V(\Omega)/\mathcal{R}$ (whose definition is (2.11), with $s = 0$). Moreover,

$$\begin{aligned} e_{33}(v_s) &= \partial_3 \underline{v}_{s3} - x_{s\alpha}\partial_{33} v_{s\alpha} = \partial_3 \underline{v}_{s3} - (x_\alpha + s\,\theta_\alpha)\partial_{33} v^s_\alpha \\ &= \partial_3 \underline{v}^s_3 - x_\alpha \partial_{33} v^s_\alpha - s\,\theta_\alpha \partial_{33} v^s_\alpha = e_{33}(v^s) - s\,\theta_\alpha \partial_{33} v^s_\alpha. \end{aligned} \tag{3.3}$$

Using (3.2)–(3.3) and the change of variables formula for domain and boundary integrals (cf. Delfour and Zolésio [5, pp. 351–353]), we get the next expression for $a_s(u_s, v_s)$,

$$\int_\Omega \frac{1}{b_{3333}(d^s)} \big(e_{33}(u^s) - s\,\theta_\alpha \partial_{33} u^s_\alpha\big)\big(e_{33}(v^s) - s\,\theta_\alpha \partial_{33} v^s_\alpha\big)\, \det \nabla Q_s d\Omega, \tag{3.4}$$

and for $L_s(v_s)$ the expression

$$(3.5)\qquad \begin{cases} \int_\Omega \gamma(\xi_0^s + P_\eta(d^s))\, f_i^s\, v_i^s\, (\det \nabla Q_s)\, d\Omega + \int_\Gamma g_i^s\, v_i^s\, |(Cof\, \nabla Q_s)^T n|_{\mathbb{R}^3}\, d\Gamma \\ \qquad + \int_{\Gamma_0 \cup \Gamma_L} h_i^s\, v_i^s\, |(Cof\, \nabla Q_s)^T n|_{\mathbb{R}^3}\, d(\Gamma_0 \cup \Gamma_L). \end{cases}$$

In (3.5), $|.|_{\mathbb{R}^3}$ is the Euclidean norm in $\mathbb{R}^3$, $n = (n_1, n_2, n_3)$ is the unit outer normal vector along the boundary $\partial\Omega$ of Ω, and $(Cof\, \nabla Q_s)^T$ is the transpose of the cofactor matrix of ∇Q_s, that is, $(Cof\, \nabla Q_s)^T = (\det \nabla Q_s)(\nabla Q_s)^{-T}$, whose definition depends only on s and the partial derivatives of θ. Developing (3.4)–(3.5), we obtain the following decomposition for the equation $a_s(u_s, v_s) = L_s(v_s)$:

(3.6)
$$\left[\begin{array}{l} a_0^s(u^s, v^s) + s a_1^s(u^s, v^s) + s^2 a_2^s(u^s, v^s) + s^3 a_3^s(u^s, v^s) + s^4 a_4^s(u^s, v^s) \\ \\ = \left\{ \begin{array}{l} F_0^s(v^s) + G_0^s(v^s) + H_0^s(v^s) + s\left(F_1^s(v^s) + G_1^s(v^s) + H_1^s(v^s)\right) \\ + s^2\left(F_2^s(v^s) + G_2^s(v^s) + H_2^s(v^s)\right) + s^3\left(F_3^s(v^s) + G_3^s(v^s) + H_3^s(v^s)\right). \end{array} \right. \end{array} \right.$$

The bilinear forms $a_i^s(.,.)$ for $i = 0, 1, 2, 3, 4$ depend on θ and d^s and are defined by the formulas

$$a_0^s(u, v) = \int_\Omega \frac{1}{b_{3333}(d^s)} e_{33}(u) e_{33}(v) d\Omega,$$

$$a_1^s(u, v) = \left\{ \begin{array}{l} \displaystyle\int_\Omega \frac{1}{b_{3333}}(d^s)\Big[- \theta_\alpha\big(e_{33}(u)\partial_{33} v_\alpha + e_{33}(v)\partial_{33} u_\alpha\big) \\ \qquad\qquad + e_{33}(u) e_{33}(v) \mathrm{div}\, \theta\Big] d\Omega, \end{array} \right.$$

$$(3.7)\qquad a_2^s(u, v) = \left\{ \begin{array}{l} \displaystyle\int_\Omega \frac{1}{b_{3333}(d^s)}\Big[e_{33}(u) e_{33}(v) \det \nabla\theta + \theta_\alpha\, \theta_\beta\, \partial_{33} u_\alpha\, \partial_{33} v_\beta \\ \qquad\qquad - (\mathrm{div}\, \theta)\, \theta_\alpha\big(e_{33}(u)\partial_{33} v_\alpha + e_{33}(v)\partial_{33} u_\alpha\big)\Big] d\Omega, \end{array} \right.$$

$$a_3^s(u, v) = \left\{ \begin{array}{l} \displaystyle\int_\Omega \frac{1}{b_{3333}}(d^s)\Big[\theta_\alpha \theta_\beta \partial_{33} u_\alpha \partial_{33} v_\beta\, \mathrm{div}\, \theta \\ \qquad\qquad - (\det \nabla\theta)\, \theta_\alpha\big(e_{33}(u)\partial_{33} v_\alpha + e_{33}(v)\partial_{33} u_\alpha\big)\Big] d\Omega, \end{array} \right.$$

$$a_4^s(u, v) = \int_\Omega \frac{1}{b_{3333}(d^s)}\Big[\theta_\alpha\, \theta_\beta\, \partial_{33} u_\alpha\, \partial_{33} v_\beta\, \det \nabla\theta\Big] d\Omega$$

for any pair (u, v) in the space $V(\Omega)/\mathcal{R}$. The linear forms F_0^s, F_1^s, F_2^s, and F_3^s depend on θ and d^s and are also defined in the same quotient space $V(\Omega)/\mathcal{R}$ by the following expressions:

$$(3.8)\qquad \begin{array}{l} F_0^s(v) = \displaystyle\int_\Omega \gamma(\xi_0^s + P_\eta(d^s))\, (f_\alpha^s\, v_\alpha + f_3^s\, \underline{v}_3)\, d\Omega, \\ F_1^s(v) = \displaystyle\int_\Omega \gamma(\xi_0^s + P_\eta(d^s))\, \Big[(f_\alpha^s\, v_\alpha + f_3^s\, \underline{v}_3)\, \mathrm{div}\, \theta - f_3^s\, \theta_\alpha\, \partial_3 v_\alpha\Big]\, d\Omega, \\ F_2^s(v) = \displaystyle\int_\Omega \gamma(\xi_0^s + P_\eta(d^s))\, \Big[(f_\alpha^s\, v_\alpha + f_3^s\, \underline{v}_3)\, \det \nabla\theta - f_3^s\, \theta_\alpha\, \partial_3 v_\alpha\, \mathrm{div}\, \theta\Big]\, d\Omega, \\ F_3^s(v) = -\displaystyle\int_\Omega \gamma(\xi_0^s + P_\eta(d^s))\, f_3^s\, \theta_\alpha\, \partial_3 v_\alpha\, \det \nabla\theta\, d\Omega. \end{array}$$

The linear forms G_0^s, G_1^s, G_2^s, G_3^s and H_0^s, H_1^s, H_2^s, H_3^s result from the change of variable in the boundary integrals (defined in Γ_s and in $\Gamma_{s0} \cup \Gamma_{sL}$, respectively) and depend on θ and n (the unit outer normal vector) but are independent of d^s. These forms are defined by the following expressions, for any v in the space $V(\Omega)/\mathcal{R}$:

$$
\begin{aligned}
G_0^s(v) &= \int_\Gamma \left(g_\alpha^s\, v_\alpha + g_3^s\, \underline{v}_3\right) d\Gamma,\\
G_1^s(v) &= \int_\Gamma \Big[(g_\alpha^s\, v_\alpha + g_3^s\, \underline{v}_3)\, G_1(\theta,n) - g_3^s\, \theta_\alpha\, \partial_3 v_\alpha\Big]\, d\Gamma,\\
G_2^s(v) &= \int_\Gamma \Big[(g_\alpha^s\, v_\alpha + g_3^s\, \underline{v}_3)\, G_2(\theta,n) - g_3^s\, \theta_\alpha\, \partial_3 v_\alpha\, G_1(\theta,n)\Big]\, d\Gamma,\\
G_3^s(v) &= -\int_\Gamma g_3^s\, \theta_\alpha\, \partial_3 v_\alpha\, G_3(\theta,n)\, d\Gamma,
\end{aligned}
\tag{3.9}
$$

where $G_1(\theta,n)$, $G_2(\theta,n)$, and $G_3(\theta,n)$ are bounded scalar functions of θ and n and

$$
\begin{aligned}
H_0^s(v) &= \int_{\Gamma_0\cup\Gamma_L} \left(h_\alpha^s\, v_\alpha + h_3^s\, \underline{v}_3\right) d(\Gamma_0 \cup \Gamma_L),\\
H_1^s(v) &= \int_{\Gamma_0\cup\Gamma_L} \Big[(h_\alpha^s\, v_\alpha + h_3^s\, \underline{v}_3)\, H_1(\theta) - h_3^s\, \theta_\alpha\, \partial_3 v_\alpha\Big]\, d(\Gamma_0 \cup \Gamma_L),\\
H_2^s(v) &= \int_{\Gamma_0\cup\Gamma_L} \Big[(h_\alpha^s\, v_\alpha + h_3^s\, \underline{v}_3)\, H_2(\theta) - h_3^s\, \theta_\alpha\, \partial_3 v_\alpha\, H_1(\theta)\Big]\, d(\Gamma_0 \cup \Gamma_L),\\
H_3^s(v) &= -\int_{\Gamma_0\cup\Gamma_L} h_3^s\, \theta_\alpha\, \partial_3 v_\alpha\, H_2(\theta)\, d(\Gamma_0 \cup \Gamma_L),
\end{aligned}
\tag{3.10}
$$

where $H_1(\theta)$ and $H_2(\theta)$ are bounded scalar functions of θ.

3.2. The problem $\mathbf{P_s}$ formulated in $\overline{\Omega} \times [0, T]$. As a direct consequence of (3.6) we can formulate, for each $s \in [0, \delta]$, the problem (P_s) in the fixed domain $\overline{\Omega} \times [0, T]$, as explained in the following theorem. The new equivalent problem is denoted by (P^s), with upper index s.

THEOREM 3.1 (problem (P^s)). *For each $s \in [0, \delta]$, we assume that $d^s(x, 0) = \hat{d}(x)$ in $\overline{\Omega}$, and $\hat{d}$ is independent of s. Then, the problem (P^s), for $s \neq 0$, is equivalent to the following problem (P^s) posed in the domain $\overline{\Omega} \times [0, T]$ independent of s:*

$$
\left[\begin{array}{l}
\text{Find } u^s : \overline{\Omega} \times [0,T] \to \mathbb{R}^3 \quad \text{and} \quad d^s : \overline{\Omega} \times [0,T] \to \mathbb{R} \quad \text{such that}\\
u^s(.,t) \in V(\Omega)/\mathcal{R},\\
\left[\begin{array}{l}
a_0^s(u^s,v) + s\, a_1^s(u^s,v) + s^2\, a_2^s(u^s,v) + s^3\, a_3^s(u^s,v) + s^4\, a_4^s(u^s,v)\\
= F_0^s(v) + sF_1^s(v) + s^2\, F_2^s(v) + s^3\, F_3^s(v)\\
\quad + G_0^s(v) + s\, G_1^s(v) + s^2\, G_2^s(v) + s^3\, G_3^s(v)\\
\qquad + H_0^s(v) + s\, H_1^s(v) + s^2\, H_2^s(v) + s^3\, H_3^s(v) \qquad \forall v \in V(\Omega)/\mathcal{R},
\end{array}\right.\\
\dot{d}^s = c(d^s) e_{33}(u^s) + a(d^s) - s\, c(d^s) \theta_\alpha \partial_{33} u_\alpha^s \quad \text{in} \quad \Omega \times (0,T),\\
d^s(x,0) = \hat{d}(x) \quad \text{in} \quad \overline{\Omega},
\end{array}\right.
\tag{3.11}
$$

where the sets $V(\Omega)$ and $V(\Omega)/\mathcal{R}$ are defined by (2.10) *and* (2.11), *for $s = 0$. We also denote by (u, d) the unique solution of problem (P^s) for $s = 0$; that is, (u, d) is*

the solution of the following problem (P^0) (*cf.* (P_s), *for the case* $s = 0$), *formulated in* $\overline{\Omega} \times [0,T]$:

$$
\left\lfloor
\begin{array}{l}
\text{Find } u : \overline{\Omega} \times [0,T] \to \mathbb{R}^3 \quad \text{and} \quad d : \overline{\Omega} \times [0,T] \to \mathbb{R} \quad \text{such that} \\
u(.,t) \in V(\Omega)/\mathcal{R}, \\
a_0(u,v) = L_0(v) \equiv F_0(v) + G_0(v) + H_0(v) \qquad \forall v \in V(\Omega)/\mathcal{R}, \\
\dot{d} = c(d) e_{33}(u) + a(d) \quad \text{in} \quad \Omega \times (0,T), \\
d(x,0) = \hat{d}(x) \quad \text{in} \quad \overline{\Omega},
\end{array}
\right. \tag{3.12}
$$

where $a_0(.,.)$, $F_0(.)$, $G_0(.)$, *and* $H_0(.)$ *are independent of* s *and are defined by*

$$
\begin{aligned}
a_0(z,v) &= \int_\Omega \frac{1}{b_{3333}(d)}\, e_{33}(z)\, e_{33}(v)\, d\Omega \quad (a_0(.,.) \text{ depends on } d), \\
F_0(v) &= \int_\Omega \gamma(\xi_0 + P_\eta(d))\, (f_\alpha\, v_\alpha + f_3\, \underline{v}_3)\, d\Omega \quad (F_0(.) \text{ depends on } d), \\
G_0(v) &= \int_\Gamma \big(g_\alpha\, v_\alpha + g_3\, \underline{v}_3\big)\, d\Gamma, \\
H_0(v) &= \int_{\Gamma_0 \cup \Gamma_L} \big(h_\alpha\ v_\alpha + h_3\, \underline{v}_3\big)\, d(\Gamma_0 \cup \Gamma_L) \quad \forall z,\, v \in V(\Omega)/\mathcal{R}.
\end{aligned} \tag{3.13}
$$

4. Calculus and identification of the shape semiderivative. In this section we first present some preliminary estimates, which prove that the sequences $\{(u^s, d^s)\}$, $\{e_{33}(u^s)\}$, and $\{(\frac{u^s-u}{s}, \frac{d^s-d}{s})\}$ are bounded, independently of s, in appropriate functional spaces involving time. These results guarantee the existence, for each t, of a pair $(\bar{u}, \bar{d})(.,t)$, which is the weak limit of a subsequence of $\{(\frac{u^s-u}{s}, \frac{d^s-d}{s})(.,t)\}$ when $s \to 0^+$. Moreover, using the regularity of u^s, we also show that the sequences $\{e_{33}(u^s - u)\}$ and $\{d^s - d\}$ converge strongly to 0, in $C^0([0,T]; C^0(\overline{\Omega}))$, when $s \to 0^+$. These two strong convergences and the preliminary estimates are the key results that enable us to prove that the weak shape semiderivative $(\bar{u}, \bar{d})$ exists and is the unique solution of a nonlinear problem.

4.1. Preliminary estimates. We present several estimates that are needed for the identification of the shape semiderivative.

THEOREM 4.1 (first estimates for the sequences u^s and d^s). *We suppose that the conditions* (2.7)–(2.8) *are verified and* $\hat{d}(x) \in L^2(\Omega)$; *then*

$$
\exists c_1 > 0 : \quad \|u^s\|_{C^0([0,T];V(\Omega)/\mathcal{R})} \le c_1 \quad \forall s \in [0,\delta], \tag{4.1}
$$

$$
\exists c_2 > 0 : \quad \|d^s\|_{C^0([0,T];L^2(\Omega))} \le c_2 \quad \forall s \in [0,\delta], \tag{4.2}
$$

where c_1 *and* c_2 *are constants independent of* s.

Proof. The pair (u^s, d^s) is the solution of problem (P^s) (cf. (3.11)); thus for each time t,

$$
a_0^s(u^s, u^s) = L^s(u^s) - \sum_{i=1}^{4} s^i\, a_i^s(u^s, u^s). \tag{4.3}
$$

By the ellipticity of $a_0^s(.,.)$ in $V(\Omega)/\mathcal{R}$ (cf. (2.16)) we have for each t

$$a_0^s(u^s,u^s) \geq c\,\|u^s(.,t)\|^2_{V(\Omega)/\mathcal{R}}, \tag{4.4}$$

where c is a constant independent of s and t. From (2.15), for $s=0$, we obtain

$$\begin{aligned} &\|\partial_3 \underline{u}_3^s(.,t)\|_{L^2(0,L)} \leq \|e_{33}(u^s(.,t))\|_{L^2(\Omega)} = \|u^s(.,t)\|_{V(\Omega)/\mathcal{R}},\\ &\|\partial_{33} u_\alpha^s(.,t)\|_{L^2(0,L)} \leq c\,\|e_{33}(u^s(.,t))\|_{L^2(\Omega)} = c\,\|u^s(.,t)\|_{V(\Omega)/\mathcal{R}}, \end{aligned} \tag{4.5}$$

where c is a constant independent of s and t. Thus using (4.5), we easily check that $a_i^s(.,.)$, for $i=1,2,3,4$, are continuous bilinear forms that verify, for each t,

$$a_i^s(u^s,u^s) \leq c_i\,\|u^s(.,t)\|^2_{V(\Omega)/\mathcal{R}} \tag{4.6}$$

with c_i strictly positive constants, independent of s and t, and depending on θ. Using the definition of the linear form $L^s(.)$ and (4.5), we also have, for each t,

$$L^s(u^s) \leq c_L\,\|u^s(.,t)\|_{V(\Omega)/\mathcal{R}}, \tag{4.7}$$

where c_L is another strictly positive constant independent of s and t. So applying (4.4), (4.6), and (4.7), we conclude that, for each time t,

$$\left(c - \sum_{i=1}^{4} s^i\,c_i\right)\,\|u^s(.,t)\|^2_{V(\Omega)/\mathcal{R}} \leq c_L\,\|u^s(.,t)\|_{V(\Omega)/\mathcal{R}}, \tag{4.8}$$

and we obtain the estimate (4.1), since s is a very small parameter.

Taking now the integral in time in the remodeling rate equation of problem (3.11), we get

$$d^s(x,t) = \int_0^t \Big[c(d^s)e_{33}(u^s) + a(d^s) - s\,c(d^s)\theta_\alpha \partial_{33} u_\alpha^s\Big]\,dr + \hat{d}(x). \tag{4.9}$$

But as the material and remodeling coefficients $c(d^s)$ and $a(d^s)$ appearing in (4.9) are bounded (cf. (2.8)), we deduce that

$$\|d^s(.,t)\|_{L^2(\Omega)} \leq \int_0^T \Big[c_1\,\|u^s(.,r)\|_{V(\Omega)/\mathcal{R}} + c_2\Big]\,dr + \|\hat{d}(x)\|_{L^2(\Omega)}, \tag{4.10}$$

with c_1 and c_2 two strictly positive constants independent of s. Therefore and because of (4.1) we have the inequality (4.2). □

THEOREM 4.2 (second estimate for the sequence u^s). *We assume that the hypotheses of Theorem* 2.1 *are satisfied, and* $\frac{1}{b_{3333}(d^s)} = b + \mathcal{O}(s)$, *where* $b:\mathbb{R}\to\mathbb{R}$ *is a scalar function independent of* s, $0<|b|\leq c$ *with* $c>0$ *a constant, and* $\mathcal{O}(s)$ *is a term of order* s *(cf. Monnier and Trabucho* [10, formulas (6) and (2)], *for a justification of this latter condition on the material coefficient* $b_{3333}(d^s)$*). Then*

$$\exists c_1>0: \quad \|u^s\|_{C^0([0,T];W^{2,2}(\Omega))} \leq c_1, \tag{4.11}$$

$$\exists c_2>0: \quad \|e_{33}(u^s)\|_{C^0([0,T];C^0(\overline{\Omega}))} \leq c_2, \tag{4.12}$$

where c_1 *and* c_2 *are constants independent of* $s\in[0,\delta]$.

Proof. Using (2.21)–(2.22) (in the proof of Theorem 2.1), we infer that, for each t, $\|u^s(.,t)\|_{W^{2,2}(\Omega)} \leq C(M^s; p_i^s)$, where $M^s = M_s \circ Q_s$, $p_i^s = p_{si} \circ Q_s$, and $C(M^s; p_i^s)$ is a strictly positive constant depending on the $W^{1,2}(0,L)$ norms of the elements of M^s and p_i^s. As p_i^s are data of the problem, related to the forces (cf. (2.20)), for $i = 1, 2, 3$, and due to the definition of M^s and the additional hypothesis for $\frac{1}{b_{3333}(d^s)}$, we easily deduce that there exists a constant $c > 0$, independent of s and t, such that $C(M^s; p_i^s) \leq c$ for all $s \in [0, \delta]$, and therefore we have (4.11).

Also from the regularity Theorem 2.1 we have, for each t,

$$e_{33}(u^s)(.,t) = \partial_3 \underline{u}_3^s(.,t) - x_\alpha \partial_{33} u_\alpha^s(.,t) \in W^{1,2}(\Omega) \cap C^0(\overline{\Omega}) \tag{4.13}$$

because $\partial_3 \underline{u}_3^s(.,t)$ and $\partial_{33} u_\alpha^s(.,t)$ belong to $W^{1,2}(0,L)$, which is compactly embedded in the space $C^0([0,L])$. Hence we get

$$\|e_{33}(u^s)(.,t)\|_{C^0(\overline{\Omega})} \leq c_1 \|e_{33}(u^s)(.,t)\|_{W^{1,2}(\Omega)} \leq c_2 \|u^s(.,t)\|_{W^{2,2}(\Omega)} \leq c_3, \tag{4.14}$$

where c_1, c_2, and c_3 are constants independent of s and t, and consequently we have (4.12). □

THEOREM 4.3 (estimate for the sequence $\frac{u^s-u}{s}$). *Let (u^s, d^s) and (u, d) be the solutions of problems* (P^s) *(cf.* (3.11)*) and* (P^0) *(cf.* (3.12)*), respectively. We assume that conditions* (2.7)–(2.8) *are verified, and, for each s, $\xi_0^s = \xi_0$, $f_i^s = f_i$, $g_i^s = g_i$, $h_i^s = h_i$, where ξ_0, f_i, g_i, and h_i are independent of s. Then,*

$$\left\| \frac{u^s - u}{s} \right\|_{C^0([0,T];V(\Omega)/\mathcal{R})} \leq c_1 \left\| \frac{d^s - d}{s} \right\|_{C^0([0,T];L^2(\Omega))} + c_2, \tag{4.15}$$

where c_1 and c_2 are strictly positive constants independent of s and t.

Proof. In this proof we sometimes write u^s instead of $u^s(.,t)$ in order to simplify the notations. For each $t \in [0,T]$, we have

$$\frac{1}{s}[a^s(u^s, v) - a_0(u, v)] = \frac{L^s(v) - L_0(v)}{s} \qquad \forall v \in V(\Omega)/\mathcal{R}. \tag{4.16}$$

Developing this last equation for the choice $v = \frac{u^s-u}{s}$, we obtain that, for each t,

$$\left[\begin{array}{l} a_0^s(\frac{u^s-u}{s}, \frac{u^s-u}{s}) = \frac{1}{s}\Big[F_0^s(\frac{u^s-u}{s}) - F_0(\frac{u^s-u}{s})\Big] - s^3\, a_4^s(u^s, \frac{u^s-u}{s}) \\ - \int_\Omega \frac{b_{3333}(d) - b_{3333}(d^s)}{s} \big(b_{3333}(d^s)\, b_{3333}(d)\big)^{-1} e_{33}(u) e_{33}(\frac{u^s-u}{s})\, d\Omega \\ + \frac{1}{s}\Big[G_0^s(\frac{u^s-u}{s}) - G_0(\frac{u^s-u}{s}) + H_0^s(\frac{u^s-u}{s}) - H_0(\frac{u^s-u}{s})\Big] \\ - a_1^s(u^s, \frac{u^s-u}{s}) + F_1^s(\frac{u^s-u}{s}) + G_1^s(\frac{u^s-u}{s}) + H_1^s(\frac{u^s-u}{s}) \\ + s\Big[- a_2^s(u^s, \frac{u^s-u}{s}) + F_2^s(\frac{u^s-u}{s}) + G_2^s(\frac{u^s-u}{s}) + H_2^s(\frac{u^s-u}{s})\Big] \\ + s^2\Big[- a_3^s(u^s, \frac{u^s-u}{s}) + F_3^s(\frac{u^s-u}{s}) + G_3^s(\frac{u^s-u}{s}) + H_3^s(\frac{u^s-u}{s})\Big]. \end{array}\right. \tag{4.17}$$

Using this last equation, the ellipticity of $a_0^s(.,.)$, and the properties of continuity of all the other remaining terms in (4.17), we obtain the estimate (4.15). We next explain these calculations in detail, analyzing (4.17) for each t.

Because of condition (2.7), we have for each t,

$$\left| a_0^s\left(\frac{u^s-u}{s}, \frac{u^s-u}{s}\right)\right| \geq c \left\| \frac{u^s-u}{s}(.,t) \right\|^2_{V(\Omega)/\mathcal{R}}, \tag{4.18}$$

where c is a strictly positive constant independent of s and t. Using the definitions of $a_i^s(.,.)$ and the estimate (4.1), we obviously obtain

$$\left| a_i^s\left(u^s, \frac{u^s-u}{s}\right)\right| \leq c_{a_i} \left|\left| \frac{u^s-u}{s}(.,t)\right|\right|_{V(\Omega)/\mathcal{R}}, \tag{4.19}$$

where c_{a_i}, for $i = 1,2,3,4$, are strictly positive constants independent of s and t. Considering now the definitions of the forms F_i^s, G_i^s, and H_i^s, associated with the applied forces, we easily check that, for each t,

$$F_i^s\left(\frac{u^s-u}{s}\right) + G_i^s\left(\frac{u^s-u}{s}\right) + H_i^s\left(\frac{u^s-u}{s}\right) \leq c_i \left|\left| \frac{u^s-u}{s}(.,t)\right|\right|_{V(\Omega)/\mathcal{R}}, \tag{4.20}$$

where c_i are strictly positive constants independent of s and t, for $i = 1,2,3$. In addition we also have, for each t, $G_0^s(\frac{u^s-u}{s}) = G_0(\frac{u^s-u}{s})$ and $H_0^s(\frac{u^s-u}{s}) = H_0(\frac{u^s-u}{s})$. Using the mean value theorem for the operator P_η, we deduce

$$\left[\begin{array}{l} \left|\frac{1}{s}\Big[F_0^s(\frac{u^s-u}{s}) - F_0(\frac{u^s-u}{s})\Big]\right| \\ \quad \leq \int_\Omega \gamma\, |\frac{P_\eta(d^s)-P_\eta(d)}{d^s-d}|\, |\frac{d^s-d}{s}|\, |f_\alpha\, \frac{u_\alpha^s-u_\alpha}{s} + f_3\, \frac{u_3^s-u_3}{s}|\, d\Omega \\ \quad\quad \leq c_0\, \|\frac{d^s-d}{s}(.,t)\|_{L^2(\Omega))}\, \|\frac{u^s-u}{s}(.,t)\|_{V(\Omega)/\mathcal{R}}, \end{array}\right. \tag{4.21}$$

where c_0 is a strictly positive constant independent of s and t. Finally, using the mean value theorem for the material coefficient $b_{3333}(.)$, the estimate (4.12) for $s = 0$, and condition (2.7), we get

$$\left[\begin{array}{l} \Big|\int_\Omega \frac{b_{3333}(d)-b_{3333}(d^s)}{s}\big(b_{3333}(d^s)\, b_{3333}(d)\big)^{-1} e_{33}(u) e_{33}(\frac{u^s-u}{s})\, d\Omega\Big| \\ = \Big|\int_\Omega \frac{b_{3333}(d)-b_{3333}(d^s)}{d^s-d}\, \frac{d^s-d}{s}\big(b_{3333}(d^s)\, b_{3333}(d)\big)^{-1} e_{33}(u) e_{33}(\frac{u^s-u}{s})\, d\Omega\Big| \\ \quad\quad \leq c_b\, \|\frac{d^s-d}{s}(.,t)\|_{L^2(\Omega)}\, \|\frac{u^s-u}{s}(.,t)\|_{V(\Omega)/\mathcal{R}}, \end{array}\right. \tag{4.22}$$

where c_b is a strictly positive constant independent of s and t. Therefore using (4.17) and the estimates (4.18)–(4.22), we have, for each t,

$$\left\{\begin{array}{l} c\, \|\frac{u^s-u}{s}(.,t)\|^2_{V(\Omega)/\mathcal{R}} \\ \leq c_1\, \|\frac{d^s-d}{s}(.,t)\|_{L^2(\Omega)}\, \|\frac{u^s-u}{s}(.,t)\|_{V(\Omega)/\mathcal{R}} + c_2\, \|\frac{u^s-u}{s}(.,t)\|_{V(\Omega)/\mathcal{R}}, \end{array}\right. \tag{4.23}$$

where c, c_1, and c_2 are strictly positive constants independent of s and t. The proof is finished, dividing (4.23) by $c\, \|\frac{u^s-u}{s}(.,t)\|_{V(\Omega)/\mathcal{R}}$. ∎

THEOREM 4.4 (estimate for the sequence $\frac{d^s-d}{s}$). *Let (u^s, d^s) and (u, d) be the solutions of problems* (3.11) *and* (3.12)*, respectively. We assume that the hypotheses of Theorem* 4.3 *are satisfied. Then*

$$\left|\left| \frac{d^s-d}{s}\right|\right|_{C^0([0,T];L^2(\Omega))} \leq c, \tag{4.24}$$

where c is a strictly positive constant independent of s and t.

Proof. Subtracting the remodeling rate equations of problems (P^s) and (P^0) (cf. (3.11) and (3.12)), taking the integral in time between 0 and t and then the

$L^2(\Omega)$ norm, using the conditions (2.7)–(2.8) and the mean value theorem for the terms $c(d^s) - c(d)$ and $a(d^s) - a(d)$, we obtain, for each t, the estimate

$$\left\{\begin{array}{l} \|(d^s - d)(.,t)\|_{L^2(\Omega)} \\ \leq \int_0^t \Big[c_1 \, \|e_{33}(u^s - u)(.,r)\|_{L^2(\Omega)} + s\, c_4 \, \|\partial_{33} u^s_\alpha(.,r)\|_{L^2(\Omega)} \\ \qquad + \big(c_2 \, \|e_{33}(u)(.,r)\|_{C^0(\overline{\Omega})} + c_3\big) \, \|(d^s - d)(.,r)\|_{L^2(\Omega)}\Big] \, dr, \end{array}\right. \tag{4.25}$$

where c_1, c_2, c_3, and c_4 are strictly positive constants independent of s and t. However,

$$\begin{aligned} &\|e_{33}(u^s - u)(.,t)\|_{L^2(\Omega)} = \|(u^s - u)(.,t)\|_{V(\Omega)/\mathcal{R}}, \\ &\|e_{33}(u)(.,t)\|_{C^0(\overline{\Omega})} \leq c_2, \\ &\|\partial_{33} u^s_\alpha(.,t)\|_{L^2(\Omega)} \leq c_0 \|e_{33} u^s(.,t)\|_{L^2(\Omega)} = c_0 \|u^s(.,t)\|_{V(\Omega)/\mathcal{R}} \leq c, \end{aligned} \tag{4.26}$$

where c_2 is the constant defined in (4.12) for the case $s = 0$, c_0 is defined in (4.5), and c is a constant depending on the constant defined in (4.1); these three constants are independent of s and t. So, dividing (4.25) by s, we have from (4.25)–(4.26) and Theorem 4.3

$$\left|\left|\frac{d^s - d}{s}(.,t)\right|\right|_{L^2(\Omega)} \leq c_5 + \int_0^t c_6 \, \left|\left|\frac{d^s - d}{s}(.,r)\right|\right|_{L^2(\Omega)} dr, \tag{4.27}$$

where c_5 and c_6 are strictly positive constants independent of s and t. Then, applying the integral Gronwall inequality (cf. Evans [6, p. 625]),

$$\left|\left|\frac{d^s - d}{s}(.,t)\right|\right|_{L^2(\Omega)} \leq c_5 \, (1 + t \, c_6 \, e^{c_6 \, t}) \qquad \forall t \in [0,T], \tag{4.28}$$

which implies (4.24). □

COROLLARY 4.5. *With the hypotheses of Theorem* 4.3*,*

$$\exists c > 0 : \qquad \left|\left|\frac{u^s - u}{s}\right|\right|_{C^0([0,T];V(\Omega)/\mathcal{R})} \leq c, \tag{4.29}$$

where c is independent of s and t.

Thus we conclude that the solutions (u^s, d^s) and (u, d), of problems (P^s) and (P^0), verify for all $s \in [0, \delta]$

$$\left|\left|\frac{u^s - u}{s}\right|\right|_{C^0([0,T];V(\Omega)/\mathcal{R})} \leq c_1 \quad \text{and} \quad \left|\left|\frac{d^s - d}{s}\right|\right|_{C^0([0,T];L^2(\Omega))} \leq c_2, \tag{4.30}$$

where c_1 and c_2 are strictly positive constants independent of s. As a consequence of this property we state the following theorem.

THEOREM 4.6 (weak limits of the quotient sequences). *Let (u^s, d^s) and (u, d) be the solutions of problems (P^s) and (P^0), and assume that the hypotheses of Theorem* 4.3 *are verified. Then, for each t, there exists a subsequence of $\{(u^s, d^s)(.,t)\}$, also denoted by $\{(u^s, d^s)(.,t)\}$, and elements $\bar{u}(.,t) \in V(\Omega)/\mathcal{R}$ and $\bar{d}(.,t) \in L^2(\Omega)$ such*

that, when the parameter $s \to 0^+$,

$$\frac{u^s - u}{s}(.,t) \rightharpoonup \bar{u}(.,t) \quad \textit{weakly in} \quad V(\Omega)/\mathcal{R}, \tag{4.31}$$

$$e_{33}\left(\frac{u^s - u}{s}\right)(.,t) \rightharpoonup e_{33}(\bar{u})(.,t) \quad \textit{weakly in} \quad L^2(\Omega), \tag{4.32}$$

$$\frac{d^s - d}{s}(.,t) \rightharpoonup \bar{d}(.,t) \quad \textit{weakly in} \quad L^2(\Omega). \tag{4.33}$$

Therefore, when $s \to 0^+$, $(u^s - u)(.,t)$ *converges strongly to* 0 *in* $V(\Omega)/\mathcal{R}$, *and the sequences* $e_{33}(u^s - u)(.,t)$ *and* $(d^s - d)(.,t)$ *converge strongly to* 0 *in* $L^2(\Omega)$.

We conclude this section with a convergence result concerning the sequences $\{u^s\}$ and $e_{33}(u^s)$, and a corollary about the convergence of $\{d^s\}$, which will be useful in subsection 4.2.

THEOREM 4.7 (strong limit of $e_{33}(u^s)$). *Let* (u^s, d^s) *and* (u, d) *be the solutions of problems* (P^s) *and* (P^0) *and assume that the hypotheses of Theorems* 2.1, 4.2, *and* 4.3 *are verified. Then there exists a subsequence of* $\{u^s\}$, *also denoted by* $\{u^s\}$, *that verifies the following convergence, when the parameter* $s \to 0^+$:

$$e_{33}(u^s - u) \longrightarrow 0 \quad \textit{strongly in} \quad C^0([0,T]; C^0(\overline{\Omega})). \tag{4.34}$$

Proof. Recalling the definition of $u^s - u$ and its regularity (cf. Theorem 2.1), we have, for each t, $(u^s_\alpha - u_\alpha)(.,t) \in W^{3,2}(0,L)$ for $\alpha = 1, 2$ and $(\underline{u}^s_3 - \underline{u}_3)(.,t) \in W^{2,2}(0,L)$. The calculus of $e_{33}(u^s - u)$ gives

$$e_{33}(u^s - u) = \partial_3(u^s_3 - u_3) = \partial_3(\underline{u}^s_3 - \underline{u}_3) - x_\alpha \partial_{33}(u^s_\alpha - u_\alpha), \tag{4.35}$$

where $[\partial_3(\underline{u}^s_3 - \underline{u}_3)](.,t) \in W^{1,2}(0,L)$ and $[\partial_{33}(u^s_\alpha - u_\alpha)](.,t) \in W^{1,2}(0,L)$, for $\alpha = 1, 2$. But, because of the strong convergence of $e_{33}(u^s - u)(.,t)$ to 0 in $L^2(\Omega)$ (cf. Theorem 4.6), and applying the definition of $\|e_{33}(u^s - u)(.,t)\|_{L^2(\Omega)}$ (cf. (2.15)), we conclude immediately that, for each t, $\partial_3(\underline{u}^s_3 - \underline{u}_3)(.,t)$ and $\partial_{33}(u^s_\alpha - u_\alpha)(.,t)$ converge strongly to 0 in $L^2(0,L)$. On the other hand, we get directly from (4.11)

$$\begin{cases} \|e_{33}(u^s - u)\|_{C^0([0,T];W^{1,2}(\Omega))} \le \|u^s - u\|_{C^0([0,T];W^{2,2}(\Omega))} \\ \qquad \le \|u^s\|_{C^0([0,T];W^{2,2}(\Omega))} + \|u\|_{C^0([0,T];W^{2,2}(\Omega))} \le c, \end{cases} \tag{4.36}$$

where c is a constant independent of s. Therefore the sequences $\partial_3(\underline{u}^s_3 - \underline{u}_3)$ and $\partial_{33}(u^s_\alpha - u_\alpha)$ are bounded in $C^0([0,T]; W^{1,2}(0,L))$, and consequently, we obtain that $\partial_3(\underline{u}^s_3 - \underline{u}_3)(.,t)$ and $\partial_{33}(u^s_\alpha - u_\alpha)(.,t)$ weakly converge to 0, in $W^{1,2}(0,L)$ when $s \to 0^+$. But as the space $W^{1,2}(0,L)$ is compactly embedded in $C^0([0,L])$, we have that $\partial_3(\underline{u}^s_3 - \underline{u}_3)(.,t)$ and $\partial_{33}(u^s_\alpha - u_\alpha)(.,t)$ strongly converge to 0 in $C^0([0,L])$ when $s \to 0^+$. This implies the strong convergence of $e_{33}(u^s - u)$ to 0 in $C^0([0,T]; C^0(\overline{\Omega}))$. ∎

COROLLARY 4.8 (strong limit of d^s). *Let* (u^s, d^s) *and* (u, d) *be the solutions of problems* (P^s) *and* (P^0), *and assume the same hypotheses of Theorem* 4.7. *Then there exists a constant* $c > 0$ *independent of* s *such that when* $s \to 0^+$

$$d^s - d \longrightarrow 0 \quad \textit{strongly in} \quad C^0([0,T]; C^0(\overline{\Omega})). \tag{4.37}$$

Proof. Using exactly the same arguments as in the beginning of the proof of Theorem 4.4,

$$|(d^s-d)(x,t)| \le \int_0^t \Big[c_1|e_{33}(u^s-u)(x,r)| + c_2|(d^s-d)(x,r)| + sc_3\Big]dr, \tag{4.38}$$

where the constants c_i, for $i=1,2,3$, are strictly positive constants independent of s and t, and consequently

$$\begin{cases} |(d^s-d)(x,t)| \le \int_0^t c_2\,|(d^s-d)(x,r)|\,dr \\ \qquad + \underbrace{T\Big[c_1\,\|e_{33}(u^s-u)\|_{C^0([0,T];C^0(\overline{\Omega}))} + s\,c_3\Big]}_{\varphi^s}. \end{cases} \tag{4.39}$$

Because of the strong convergence (4.34), the scalar $\varphi^s \to 0$ when $s \to 0^+$. Then we obtain the convergence (4.37), applying to (4.39) the integral Gronwall inequality (cf. Evans [6, p. 625]). ∎

4.2. Shape semiderivatives. The objective of this section is to identify, for each t, the weak limits $\bar{u}(.,t)$ and $\bar{d}(.,t)$ of the sequences $\{\frac{u^s-u}{s}(.,t)\}$ and $\{\frac{d^s-d}{s}(.,t)\}$, defined in (4.31) and (4.33). The procedure is the following: we subtract and divide by s the equilibrium variational equations and the remodeling rate equations in problems (P^s) and (P^0), and then we take the limit, when the parameter $s \to 0^+$. We conclude that, for each t, the pair $(\bar{u}(.,t), \bar{d}(.,t))$ is the solution of another nonlinear problem. Finally we end up proving that $(\bar{u},\bar{d})$ is the unique solution of this latter problem in the space $C^1([0,T];V(\Omega)/\mathcal{R}) \times C^1([0,T];C^0(\overline{\Omega}))$.

4.2.1. Weak limit $\bar{\mathbf{u}}(.,\mathbf{t})$. Subtracting and dividing by s the two equilibrium variational equations of problems (P^s) and (P^0), we obtain, for each $t \in [0,T]$ (cf. (4.16)–(4.17)),

$$\begin{cases} \int_\Omega \frac{1}{b_{3333}(d^s)} e_{33}(\frac{u^s-u}{s}) e_{33}(v)d\Omega \\ + \int_\Omega \frac{b_{3333}(d)-b_{3333}(d^s)}{s}\big(b_{3333}(d^s)\,b_{3333}(d)\big)^{-1} e_{33}(u)e_{33}(v)\,d\Omega \\ + a_1^s(u^s,v) - F_1^s(v) - G_1^s(v) - H_1^s(v) \\ = \frac{1}{s}\big[F_0^s(v) - F_0(v) + G_0^s(v) - G_0(v) + H_0^s(v) - H_0(v)\big] \\ + s\,\big[-a_2^s(u^s,v) + F_2^s(v) + G_2^s(v) + H_2^s(v)\big] \\ + s^2\,\big[-a_3^s(v) + F_3^s(v) + G_3^s(v) + H_3^s(v)\big] - s^3\,a_4^s(u^s,v). \end{cases} \tag{4.40}$$

Computing now for each t the limit of each term of (4.40), we obtain the following theorem.

THEOREM 4.9 (identification of $\bar{u}(.,t)$). *We assume the hypotheses of Theorems* 2.1, 4.2, *and* 4.3. *Then the weak limit* $\bar{u}(.,t)$ *of the sequence* $\{\frac{u^s-u}{s}\}(.,t)$ *verifies the following variational equation for each* $t \in [0,T]$:

$$B(\bar{u},v) = S(v) \qquad \forall v \in V(\Omega)/\mathcal{R}. \tag{4.41}$$

The linear form $S(.)$ is defined in $V(\Omega)/\mathcal{R}$ by

$$
\left\{
\begin{array}{l}
S(v) = \int_\Omega b'_{3333}(d)\, b_{3333}(d)^{-2}\, \bar{d}\, e_{33}(u)\, e_{33}(v)\, d\Omega \\
- \int_\Omega \frac{1}{b_{3333}(d)} \big[- \theta_\alpha \big(e_{33}(u) \partial_{33} v_\alpha + e_{33}(v) \partial_{33} u_\alpha \big) \\
\qquad\qquad + e_{33}(u) e_{33}(v) \mathrm{div}\, \theta \big] d\Omega \\
+ \int_\Omega \gamma\, \bar{d}\, P'_\eta(d)\, (f_\alpha\, v_\alpha + f_3\, \underline{v}_3)\, d\Omega \\
+ \int_\Omega \gamma (\xi_0 + P_\eta(d)) \left[(f_\alpha\, v_\alpha + f_3\, \underline{v}_3)\, \mathrm{div}\, \theta - f_3\, \theta_\alpha\, \partial_3 v_\alpha \right] d\Omega \\
+ \int_\Gamma \left[(g_\alpha\, v_\alpha + g_3\, \underline{v}_3)\, G_1(\theta, n) - g_3\, \theta_\alpha\, \partial_3 v_\alpha \right] d\Gamma \\
+ \int_{\Gamma_0 \cup \Gamma_L} \left[(h_\alpha\, v_\alpha + h_3\, \underline{v}_3)\, H_1(\theta) - h_3\, \theta_\alpha\, \partial_3 v_\alpha \right] d\Gamma_0 \cup \Gamma_L,
\end{array}
\right. \tag{4.42}
$$

with $b'_{3333}(.)$ and $P'_\eta(.)$ the first derivatives of the scalar functions $b_{3333}(.)$ and $P_\eta(.)$, respectively. The bilinear form $B(.,.)$ is defined by

$$
B(z, v) = \int_\Omega \frac{1}{b_{3333}(d)}\, e_{33}(z)\, e_{33}(v)\, d\Omega \qquad \forall z,\ v \in V(\Omega)/\mathcal{R}. \tag{4.43}
$$

Proof. We give a sketch of the computation of the limits in (4.40), for each t.

The first term in (4.40) verifies, when $s \to 0^+$,

$$
\int_\Omega \frac{1}{b_{3333}(d^s)} e_{33} \left(\frac{u^s - u}{s} \right) e_{33}(v) d\Omega \ \longrightarrow\ \int_\Omega \frac{1}{b_{3333}(d)} e_{33}(\bar{u}) e_{33}(v) d\Omega, \tag{4.44}
$$

because the sequence $e_{33}(\frac{u^s - u}{s})(.,t)$ weakly converges to $e_{33}(\bar{u})(.,t)$ in $L^2(\Omega)$ (cf. (4.32)), and $\frac{e_{33}(v)}{b_{3333}(d^s)}$ converges strongly to $\frac{e_{33}(v)}{b_{3333}(d)}$ in $L^2(\Omega)$, due to the condition (2.7), the mean value theorem for the function $\frac{1}{b_{3333}(.)}$, and the strong convergence of d^s to d in $C^0([0,T]; C^0(\overline{\Omega}))$ (cf. (4.37)).

The second term in (4.40) converges to

$$
- \int_\Omega b'_{3333}(d)\, b_{3333}(d)^{-2}\, \bar{d}\, e_{33}(u)\, e_{33}(v)\, d\Omega, \tag{4.45}
$$

when $s \to 0^+$, because of condition (2.7), the mean value theorem for the function $b_{3333}(.)$, the condition (4.12) for the case $s = 0$, the strong convergence of d_s to d in $C^0([0,T]; C^0(\overline{\Omega}))$ (cf. Corollary 4.8, formula (4.37)), and the weak convergence of $\frac{d^s - d}{s}(.,t)$ to $\bar{d}(.,t)$ in $L^2(\Omega)$.

For the third term in (4.40), we have that $a_1^s(u^s, v)$ converges to

$$
\int_\Omega \frac{1}{b_{3333}(d)} \Big[- \theta_\alpha \big(e_{33}(u) \partial_{33} v_\alpha + e_{33}(v) \partial_{33} u_\alpha \big) + e_{33}(u) e_{33}(v) \mathrm{div}\, \theta \Big] d\Omega \tag{4.46}
$$

when $s \to 0^+$, and $F_1^s(v) + G_1^s(v) + H_1^s(v)$ converges to

$$
\left\{
\begin{array}{l}
\int_\Omega \gamma (\xi_0 + P_\eta(d)) \Big[(f_\alpha\, v_\alpha + f_3\, \underline{v}_3)\, \mathrm{div}\, \theta - f_3\, \theta_\alpha\, \partial_3 v_\alpha \Big] d\Omega \\
+ \int_\Gamma \Big[(g_\alpha\, v_\alpha + g_3\, \underline{v}_3)\, G_1(\theta, n) - g_3\, \theta_\alpha\, \partial_3 v_\alpha \Big] d\Gamma \\
+ \int_{\Gamma_0 \cup \Gamma_L} \Big[(h_\alpha\, v_\alpha + h_3\, \underline{v}_3)\, H_1(\theta) - h_3\, \theta_\alpha\, \partial_3 v_\alpha \Big] d\Gamma_0 \cup \Gamma_L.
\end{array}
\right. \tag{4.47}
$$

To prove (4.46) we remark that, when $s \to 0^+$, $\|\big(e_{33}(u^s) - e_{33}(u)\big)(.,t)\|_{L^2(\Omega)}$ and $\|(d^s - d)(.,t)\|_{L^2(\Omega)}$ converge to 0 (cf. Theorem 4.6). To obtain (4.47) we apply the definitions of the forms F_1^s, G_1^s, H_1^s and the strong convergence of $P_\eta(d^s)$ to $P_\eta(d)$, when $s \to 0^+$, in the space $C^0([0,T]; C^0(\overline{\Omega}))$.

The fourth term in (4.40) converges to

$$\int_\Omega \gamma\, \bar{d}\, P'_\eta(d)\, (f_\alpha\, v_\alpha + f_3\, \underline{v}_3)\, d\Omega \tag{4.48}$$

when $s \to 0^+$. In fact, we have $G_0^s(v) = G_0(v)$ and $H_0^s(v) = H_0(v)$, and

$$\left\{ \begin{array}{l} \frac{1}{s}\Big[F_0^s(v) - F_0(v)\Big] - \int_\Omega \gamma\, \bar{d}\, P'_\eta(d)\, (f_\alpha\, v_\alpha + f_3\, \underline{v}_3)\, d\Omega \\ = \int_\Omega \gamma \Big[\frac{P_\eta(d^s) - P_\eta(d)}{d^s - d} \Big(\frac{d^s - d}{s} - \bar{d} \Big) \\ \qquad\qquad + \Big(\frac{P_\eta(d^s) - P_\eta(d)}{d^s - d} - P'_\eta(d) \Big) \bar{d} \Big] (f_\alpha\, v_\alpha + f_3\, \underline{v}_3)\, d\Omega. \end{array} \right. \tag{4.49}$$

When $s \to 0^+$, $\gamma\, \frac{P_\eta(d^s) - P_\eta(d)}{d^s - d}$ converges strongly to $\gamma\, P'_\eta(d)$ in $C^0([0,T]; C^0(\overline{\Omega}))$, and, for each t, $\frac{d^s - d}{s}$ converges weakly to $\bar{d}$ in $L^2(\Omega)$ (cf. Theorem 4.6, formula (4.33)), and $\bar{d}\,(f_\alpha\, v_\alpha + f_3\, \underline{u}_3)$ belongs to $L^1(\Omega)$. Thus (4.49) converges to 0, for each t, when $s \to 0^+$.

Finally the last two terms in (4.40) converge to 0 when $s \to 0^+$, because those are composed of bounded terms multiplied by a positive power of s. ∎

4.2.2. Weak limit $\bar{\mathbf{d}}(.,\mathbf{t})$. By subtracting and dividing by s the remodeling rate equations of problems (P^s) and (P^0), and integrating in time between 0 and t, we obtain the following theorem.

THEOREM 4.10 (identification of $\bar{d}(.,t)$). *We assume that the hypotheses of Theorems* 2.1, 4.2, *and* 4.3 *are verified. For each t, the weak limit $\bar{d}(.,t)$ of the sequence $\{\frac{d^s - d}{s}(.,t)\}$ is the solution of the following ordinary differential equation with respect to time:*

$$\left\{ \begin{array}{l} \dot{\bar{d}} = c(d) e_{33}(\bar{u}) + \bar{d}\,[c'(d)\, e_{33}(u) + a'(d)] - c(d)\, \theta_\alpha \partial_{33} u_\alpha, \\ \bar{d}(x,0) = 0. \end{array} \right. \tag{4.50}$$

Proof. For any $v \in L^2(\Omega)$ and for each t, we have

$$\left\{ \begin{array}{l} \int_\Omega \frac{d^s - d}{s}\, v\, d\Omega = \int_0^t \Big(\int_\Omega \Big[c(d^s)\, e_{33}(\frac{u^s - u}{s}) + \frac{c(d^s) - c(d)}{s}\, e_{33}(u) \\ \qquad\qquad + \frac{a(d^s) - a(d)}{s} - c(d^s) \theta_\alpha \partial_{33} u_\alpha^s \Big]\, v\, d\Omega \Big)\, dr. \end{array} \right. \tag{4.51}$$

On the other hand, for each t, and when $s \to 0^+$, $c(d^s)$ converges strongly to $c(d)$ in $C^0(\overline{\Omega})$, $e_{33}(\frac{u^s - u}{s})$ converges weakly to $e_{33}(\bar{u})$ in $L^2(\Omega)$, $\frac{c(d^s) - c(d)}{s}\, e_{33}(u)$ converges weakly to $\bar{d}\, c'(d)\, e_{33}(u)$ in $L^2(\Omega)$, $\frac{a(d^s) - a(d)}{s}$ converges weakly to $\bar{d}\, a'(d)$ in $L^2(\Omega)$, and $\partial_{33} u_\alpha^s$ converges strongly to $\partial_{33} u_\alpha$ in $L^2(\Omega)$. Hence, we have that, for each t and when $s \to 0^+$, the sequence $\int_\Omega \frac{d^s - d}{s}\, v\, d\Omega$ converges to

$$\int_\Omega \left(\int_0^t \Big[c(d) e_{33}(\bar{u}) + \bar{d}\, c'(d)\, e_{33}(u) + \bar{d}\, a'(d) - c(d)\, \theta_\alpha \partial_{33} u_\alpha \Big]\, dr \right) v\, d\Omega. \tag{4.52}$$

But by (4.33), $\frac{d^s - d}{s}(.,t)$ converges weakly to $\bar{d}(.,t)$ in $L^2(\Omega)$ when $s \to 0^+$. Therefore $\bar{d}(.,t)$ must verify (4.50), since the weak limit is unique. ∎

4.2.3. Final identification result. Assembling the results of Theorems 4.9 and 4.10, we have the next theorem, which completely identifies, for each t, the (weak) shape semiderivatives $\bar{u}(.,t)$ and $\bar{d}(.,t)$.

THEOREM 4.11. *We assume that the hypotheses of Theorems* 2.1, 4.2, *and* 4.3 *are verified. For each* $t \in [0,T]$, *the weak limit* $(\bar{u},\bar{d})(.,t)$ *is an element of the space* $(V(\Omega)/\mathcal{R}) \times L^2(\Omega)$ *and is the solution of the following nonlinear problem* $(\bar{\mathrm{P}})$:

$$
\left[\begin{array}{l}
\text{Find } \bar{u} : \overline{\Omega} \times [0,T] \to \mathbb{R}^3 \quad \text{and} \quad \bar{d} : \overline{\Omega} \times [0,T] \to \mathbb{R} \quad \text{such that} \\
\bar{u}(.,t) \in V(\Omega)/\mathcal{R}, \\
B(\bar{u},v) = S(v) \qquad \forall v \in V(\Omega)/\mathcal{R}, \\
\dot{\bar{d}} = c(d) e_{33}(\bar{u}) + \bar{d}\,[c'(d)\, e_{33}(u) + a'(d)] - c(d)\,\theta_\alpha \partial_{33} u_\alpha \quad \text{in } \ \Omega \times (0,T), \\
\bar{d}(x,0) = 0 \quad \text{in} \quad \overline{\Omega},
\end{array}\right. \tag{4.53}
$$

where the linear form $S(.)$ *and the bilinear form* $B(.,.)$ *are defined by* (4.42) *and* (4.43), *respectively. We observe that* $S(.)$ *depends on* (u,d), *which is the solution of problem* (P^0), *and also on* $\bar{d}$. *The bilinear form* $B(.,.)$ *depends on* d, *that is, the measure of change in volume fraction of the elastic material of problem* (P^0). *Moreover, there exists a unique solution* $(\bar{u},\bar{d})$ *of problem* $(\bar{\mathrm{P}})$ *such that* $\bar{u} \in C^1([0,T]; V(\Omega)/\mathcal{R})$ *and* $\bar{d} \in C^1([0,T]; C^0(\overline{\Omega}))$. *Consequently, for each* t, *the entire sequence* $\{(\frac{u^s-u}{s}, \frac{d^s-d}{s})(.,t)\}$ *weakly converges to* $(\bar{u},\bar{d})(.,t)$ *in the space* $V(\Omega)/\mathcal{R} \times L^2(\Omega)$ *when* $s \to 0^+$. *Thus, there exists the weak shape semiderivative of the shape map* $J(\Omega_s) = (u^s, d^s)$ *at* $s = 0$ *in the direction of the vector field* θ (*cf.* (2.24)), *and it is perfectly defined, for each* t, *by* $dJ(\Omega;\theta)(.,t) = (\bar{u},\bar{d})(.,t)$, *where* $(\bar{u},\bar{d}) \in C^1([0,T]; V(\Omega)/\mathcal{R}) \times C^1([0,T]; C^0(\overline{\Omega}))$ *is the unique solution of problem* $(\bar{\mathrm{P}})$.

Proof. The arguments used to prove the existence and uniqueness of the solution $(\bar{u},\bar{d})$ to problem $(\bar{\mathrm{P}})$, in the space $C^1([0,T]; V(\Omega)/\mathcal{R}) \times C^1([0,T]; C^0(\overline{\Omega}))$, are analogous to those utilized in the proof of existence and uniqueness of the solution to problem $(\mathrm{P_s})$ (cf. Figueiredo and Trabucho [7]) and rely on the Schauder fixed point theorem. □

5. Conclusions and future work. In this paper we have considered the family $\overline{\Omega}_s$ of perturbed thin rods, for $s \in [0,\delta]$, and the corresponding family of solutions (u^s, d^s) of the nonlinear asymptotic adaptive elastic model, derived in Figueiredo and Trabucho [7]. We have proved that, for each t, the sequence $(\frac{u^s-u}{s}, \frac{d^s-d}{s})(.,t)$ converges weakly to $(\bar{u},\bar{d})(.,t)$ in the space $(V(\Omega)/\mathcal{R}) \times L^2(\Omega)$ when $s \to 0^+$. Consequently, for each t, $(\bar{u},\bar{d})(.,t)$ is the weak shape semiderivative of the function $J(\Omega_s) = (u^s, d^s)$ at $s = 0$ in the direction of the vector field θ. Moreover, we have showed that the pair $(\bar{u},\bar{d})$ is the unique solution of another nonlinear problem that couples a variational equation, depending on (u,d) and $\bar{d}$, and an ordinary differential equation with respect to time, depending on (u,d) and $\bar{u}$. We intend to apply this methodology to analyze the weak shape semiderivative of the solution to the nonlinear adaptive elastic asymptotic model (2.1)–(2.4), but for the case where the remodeling rate equation (2.4) depends nonlinearly on $e_{33}(u_s)$ (cf. Figueiredo and Trabucho [7]). We think that this nonlinear term may generate some difficulties in proving that the sequence $\{\frac{d^s-d}{s}\}$ is bounded, independently of s, and subsequently in the identification of the shape semiderivative.

REFERENCES

[1] P. G. Ciarlet, *Mathematical Elasticity* 1: *Three-Dimensional Elasticity*, Stud. Math. Appl. 20, North–Holland, Amsterdam, 1988.

[2] P. G. Ciarlet, *Introduction to Linear Shell Theory*, Gauthier-Villars, Paris, 1998.

[3] S. C. Cowin and D. H. Hegedus, *Bone remodeling* I: *Theory of adaptive elasticity*, J. Elasticity, 6 (1976), pp. 313–326.

[4] S. C. Cowin and R. R. Nachlinger, *Bone remodeling* III: *Uniqueness and stability in adaptive elasticity theory*, J. Elasticity, 8 (1978), pp. 285–295.

[5] M. C. Delfour and J. P. Zolésio, *Shapes and Geometries, Analysis, Differential Calculus, and Optimization*, Adv. Des. Control 4, SIAM, Philadelphia, 2001.

[6] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.

[7] I. N. Figueiredo and L. Trabucho, *Asymptotic model of a nonlinear adaptive elastic rod*, Math. Mech. Solids, 9 (2004), pp. 331–354.

[8] J. Haslinger and R. A. E. Mäkinen, *Introduction to Shape Optimization, Theory, Approximation, and Computation*, Adv. Des. Control 7, SIAM, Philadelphia, 2003.

[9] D. H. Hegedus and S. C. Cowin, *Bone remodeling* II: *Small strain adaptive elasticity*, J. Elasticity, 6 (1976), pp. 337–352.

[10] J. Monnier and L. Trabucho, *An existence and uniqueness result in bone remodeling theory*, Comput. Methods Appl. Mech. Engrg., 151 (1998), pp. 539–544.

[11] L. Trabucho and J. M. Viaño, *Mathematical modelling of rods*, in Handb. Numer. Anal. 4, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1996, pp. 487–974.

[12] T. Valent, *Boundary Value Problems of Finite Elasticity*, Springer Tracts Nat. Philos. 31, Springer-Verlag, New York, 1988.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 174–186

MAGNETIC SHAPING OF A LIQUID METAL COLUMN AND DEFORMATION OF A BUBBLE IN A VORTEX FLOW*

M. G. BLYTH† AND J.-M. VANDEN-BROECK†

Abstract. Two different physical problems are considered: the magnetic shaping of a liquid metal column and the distortion of a bubble in a corner vortex flow. It is shown that the two problems can be modeled with a virtually identical set of equations. These equations are solved numerically using a conformal mapping and a series truncation method, which permits fast and efficient computation of the bubble or column shapes. It is found that the two problems exhibit different limiting configurations. For the bubble problem, the deformation becomes more severe as the vortex moves further into the corner until eventually the free surface makes contact with the walls. For the magnetic shaping problem, columns approach a limiting configuration featuring either a finite number of cusps or a fixed number of trapped bubbles along the perimeter. The division between these two different behaviors is explained by means of an exact solution for zero surface tension.

Key words. bubble, magnetic field, conformal mapping, series truncation

AMS subject classifications. 76B07, 76D45, 76W05

DOI. 10.1137/050622353

1. Introduction.

This paper is concerned with two apparently unrelated problems. The first problem is the magnetic shaping of a liquid metal column. It is motivated by the process of continuous casting in the metallurgical industry in which a free-falling vertical column of liquid metal is shaped by an electromagnetic field. This avoids any surface imperfections which might be introduced by a mold. Experimental work on magnetic shaping has been performed by Etay [7] and Etay, Gagnoud, and Garnier [8]. It appears to have been first analyzed mathematically by Shercliff [14]. The second problem is the deformation of a two-dimensional bubble in a vortex flow. It is relevant to applications where mixing of fluids or cleaning of unwanted bubbles from an apparatus is important. We show that these two problems can be modeled by almost the same equations. The only difference is a sign in the dynamic boundary condition. We solve these equations numerically, recover previous solutions, and compute new ones. Our method is relatively simple and, for the magnetic problem, for example, can be extended to any number of arbitrarily placed conductors.

The problem of computing the shape and evolution of a bubble under a prescribed set of flow conditions has occupied many workers over the years. In this article, we concentrate on the computation of bubble shapes in steady flow. Unsteady calculations have been performed, for example, by Baker and Moore [1] for a gas bubble rising in an inviscid liquid.

Inviscid models have been applied extensively in the literature to computing bubble shapes. McLeod [10] obtained an exact solution describing uniform flow past a two-dimensional bubble in a special case. Further numerical results were obtained by Vanden-Broeck and Keller [17]. Asymptotic results for the same problem were

*Received by the editors January 10, 2005; accepted for publication (in revised form) May 10, 2005; published electronically October 17, 2005. The research of the first author was supported by the Nuffield Foundation under grant NUF-NAL-O4. The research of the second author was supported by the EPSRC under grant GR/S76847.

http://www.siam.org/journals/siap/66-1/62235.html

†School of Mathematics, University of East Anglia, Norwich, NR4 7TJ, UK (M.Blyth@uea.ac.uk, J.Vanden-Broeck@uea.ac.uk).

presented by Shankar [13], and some analytical results were given by Tanveer [15]. Miksis, Vanden-Broeck, and Keller [11] found numerical solutions for uniform flow past axisymmetric bubbles.

Vanden-Broeck and Keller [18] examined straining flow past a bubble in a right-angled corner using boundary integral methods. Ozugurlu and Vanden-Broeck [12] used a series truncation method to extend these results to flow in a corner of arbitrary angle. In both of these papers, bubble shapes were computed numerically for a range of parameter values, and it was shown that in extreme cases, a configuration is reached featuring smaller, trapped bubbles along the free surface. In the present article, we extend these results to the case when the flow is driven by a line vortex positioned at a finite distance from the corner. Our aim is to compute the prevailing bubble shape, which is unknown in advance. Gravity is neglected and the bubble shape is determined by the ambient flow and by surface tension effects. We use a conformal mapping to transform the flow domain into the unit circle, and then use a series truncation method to compute the bubble shape numerically. The series truncation method has been applied previously to calculate free surface flows by Vanden-Broeck [16], Dias and Vanden-Broeck [6], Vanden-Broeck and Miloh [19], Blyth and Vanden-Broeck [2], and others. When the vortex is moved to infinity, our results reduce exactly to those obtained by Ozugurlu and Vanden-Broeck [12].

As mentioned earlier, the vortex problem and the magnetic shaping problem are intimately related mathematically. Therefore we are able to use the same numerical technique to compute liquid column shapes with little amendment to the analysis. In this way, we are able to recompute and extend the results of Shercliff using a somewhat simplified and more convenient approach.

The layout of the paper is as follows. In section 2, we formulate the two physical problems. In section 3 we describe the conformal mapping and present the numerical method. In section 4 we present our numerical results. Our conclusions are summarized in section 5.

2. Problem formulation. In this section, we present the mathematical formulations for corner vortex flow past a bubble and magnetic shaping of a liquid metal column by an arrangement of electrical conductors. The mathematical details are almost the same.

2.1. Vortex flow past a bubble. First, we consider inviscid, incompressible flow into a corner of general angle α with a trapped air bubble at the apex, as is sketched in Figure 1. Our interest lies in computing the shape of the bubble, which is unknown in advance. The flow in the corner is driven by a line vortex whose intensity and position control the shape of the bubble. We define the vortex circulation by Γ. The vortex lies at some point along the bisector, $y = \tan(\alpha/2)\, x$, at a distance d from the origin. Consequently, the flow is symmetric about this line. The contact angle, β, is a free parameter.

At the surface of the bubble, the fluid pressure undergoes a jump whose magnitude is dictated by the Laplace–Young equation,

$$p_S - p_B = \kappa T. \tag{2.1}$$

Here, p_S is the fluid pressure at the surface of the bubble, p_B is the constant pressure in the bubble, κ is the surface curvature reckoned to be negative when the bubble encloses a convex region (see Figure 1), and T is the constant surface tension. Applying

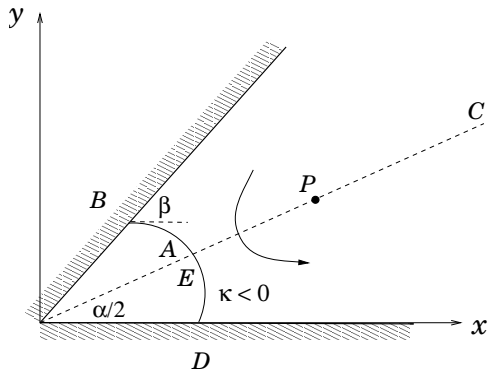


FIG. 1. *Sketch of the corner flow with a trapped air bubble at the vertex. The contact angle between the bubble and the wall is β. The flow is driven by a vortex with circulation Γ at point P which lies along the bisector, shown as a broken line. The point C lies at infinity. The arrow indicates the direction of flow.*

Bernoulli's equation on the bubble surface and invoking condition (2.1), we have

$$\frac{1}{2}q^2 + \frac{T\kappa}{\rho} = \mathcal{B}, \tag{2.2}$$

where q is the fluid speed, ρ is the fluid density, and $\mathcal{B}$ is a constant. If L is a suitable lengthscale, so that Γ/L represents a characteristic velocity scale, then a routine analysis of (2.2) reveals the importance of two dimensionless groups, one involving the surface tension, and the other involving the distance of the vortex from the origin. For example, these might be defined as

$$\frac{Td}{\rho\Gamma^2}, \qquad \frac{\mathcal{B}d^2}{\Gamma^2}. \tag{2.3}$$

The flow outside of the bubble satisfies Laplace's equation, subject to the no-normal flow condition at the walls,

$$v = 0 \quad \text{on} \quad y = 0, \qquad v/u = \tan\alpha \quad \text{on} \quad y = x\tan\alpha, \tag{2.4}$$

where u and v are the x and y components of the velocity, respectively. The dynamic condition (2.2) applies at the bubble surface.

2.2. Magnetic shaping of a liquid metal column. Next, we consider a two-dimensional column of liquid metal surrounded by an even number of wire conductors carrying high-frequency electrical current, as discussed by Shercliff [14]. The conductors are positioned at equally spaced intervals around a circle of radius d centered at the origin. Gravity is neglected. Any stirring effects in the liquid metal are ignored and the problem is treated as quasi-steady. Under this assumption, the magnetic field generated by the current-carrying wires competes with surface tension to shape the molten column. Our goal is to compute the shape of the free surface, which is unknown in advance. The column shape is assumed to be rotationally symmetric according to the number of conductors present. For this reason, the contact angle, β, is taken to be equal to $\pi/2$.

The magnetic field may be described in terms of a potential function satisfying Laplace's equation in the region exterior to the metal column. At the free surface,

the appropriate boundary condition is (see, e.g., Shercliff [14])

$$\frac{B^2}{2\mu_0} - T\,\kappa = p, \tag{2.5}$$

where B is the magnitude of the magnetic field at the surface, μ_0 is the permeability of free space, T is the surface tension, κ is the surface curvature defined as in section 2.1, and p is the uniform pressure in the liquid metal. For simplicity, we recast (2.5) in the notation of section 2.1, writing it instead as

$$\frac{1}{2}q^2 - \frac{T\kappa}{\rho} = \mathcal{E}, \tag{2.6}$$

where $q = B$, $\rho = 1/\mu_0$, and $\mathcal{E} = \mu_0 p$ is a constant. Clearly, the only difference between (2.6) and the free surface condition for the vortex flow (2.2) is the sign on the second term. Together with the fact that both problems are governed by Laplace's equation within the domain of interest, we see that, despite describing wholly different physical phenomena, the two problems are virtually identical mathematically. The solution to both of these problems is considered in the next section.

We note that there are other problems which are closely related to the current work. These include the equilibrium configuration of a charged surface of liquid metal (see Zubarev [20] and Zubarev et al. [21]) and the circulation-induced shape deformation of drops and bubbles (see, e.g., Crowdy [5], [4] and Blyth and Vanden-Broeck [3]). The boundary conditions derived by Zubarev [20] imply that our vortex flow also models the deformation of the surface of a liquid metal due to an arrangement of charges symmetrically distributed around it.

3. Conformal mapping and numerical solution. In the first problem, discussed in section 2.1, the aim is to compute the shape of the bubble for a vortex of given strength and position. In the second problem, discussed in section 2.2, the aim is to compute the shape of the liquid metal column when surrounded by a prescribed number of conductors carrying a fixed current. We have shown that these two problems are virtually mathematically equivalent, save for a difference in sign in the free surface boundary conditions (2.2) and (2.6). Both problems can be solved numerically using a series truncation method to be described in this section.

To prepare the ground, we first map the flow domain onto the unit circle in a transformed plane. To do this, we introduce the complex potential $f(z) = \phi + \mathrm{i}\psi$, where $\phi(x, y)$, $\psi(x, y)$ are the velocity potential and streamfunction, respectively, and $z = x + \mathrm{i}y$. Treating ϕ and ψ as independent coordinates, the domain of interest is illustrated in Figure 2. Note that ϕ varies between $-\Gamma/2$ and $\Gamma/2$ and takes the a priori unknown value of γ at the two contact points B and D. The parameter Γ is either a measure of the vortex strength or the current in the conductors. The mapping from the complex f plane into the unit circle in the target t plane is achieved by the transformation

$$-\frac{2\pi\mathrm{i}f}{\Gamma} = \log\left[\frac{\mathrm{i}(t^2+1) + 2\lambda t}{\mathrm{i}(t^2+1) - 2\lambda t}\right], \tag{3.1}$$

where λ is a positive constant given by

$$\lambda = \tan\left(\frac{\pi\gamma}{\Gamma}\right). \tag{3.2}$$

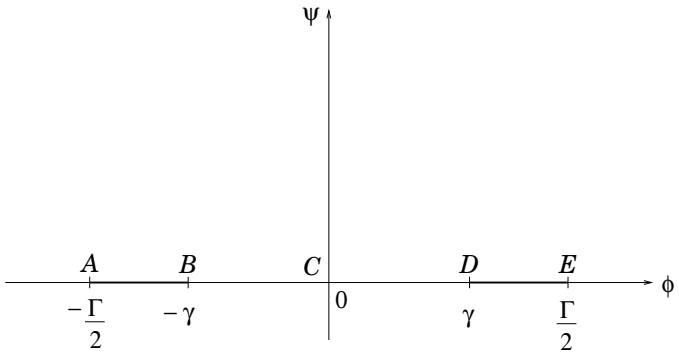


FIG. 2. *The flow domain in the complex f plane. The bubble surface is highlighted by the thick lines.*

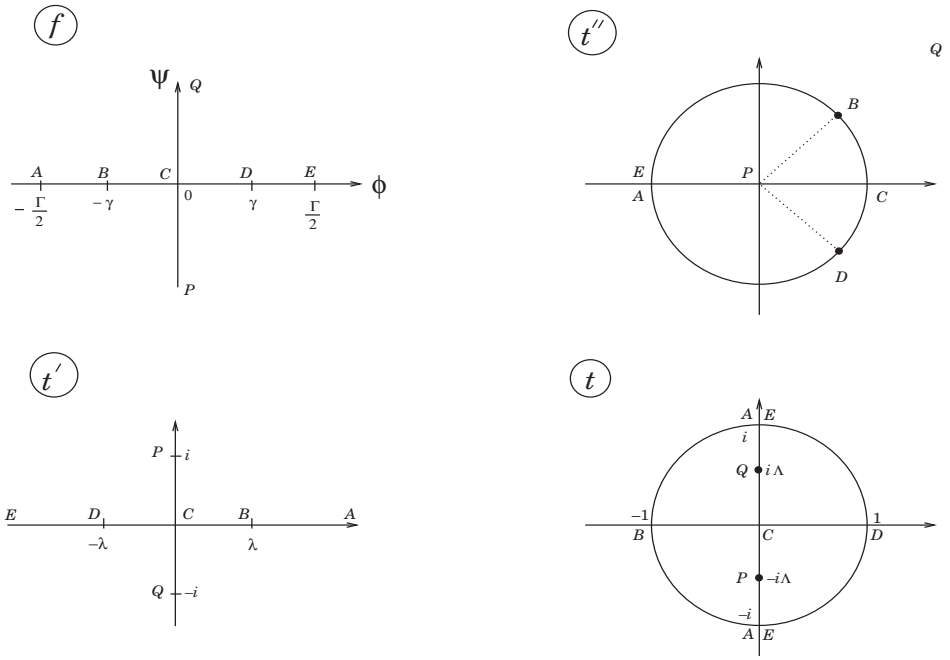


FIG. 3. *The conformal mapping from the f to the t plane via the intermediate t'' and t' planes. The succession of mappings is given by $t'' = \exp(-2\mathrm{i}\pi f/\Gamma)$, $t'' = (\mathrm{i}-t')/(\mathrm{i}+t')$, and $-2\lambda/t' = t+1/t$.*

Moving the corner vortex to infinity corresponds to the double limit $\lambda \to \infty$ and $\Gamma \to \infty$. In this case, $\Lambda \to 0$ and point P collapses onto point C. Moreover, $\gamma \to \Gamma/2$. Writing $b = \Gamma/2 - \gamma$, for constant b, and introducing the shifted complex potential $g = f - \Gamma/2$, we obtain the limiting behavior of the mapping (3.1),

$$\frac{g}{b} \sim -\frac{1+t^2}{2t}, \tag{3.3}$$

which is precisely the mapping used by Ozugurlu and Vanden-Broeck [12] when considering corner flow driven by a vortex at infinity.

The mapping from the f to the t plane is more clearly understood by introducing two intermediate planes, as is shown in Figure 3. For the magnetic shaping problem,

there are $2n$ conductors positioned symmetrically on a circle of radius d at $z = z_k$ and $z = z_k^*$, where

$$z_k = d \exp\left\{ \mathrm{i} \left(k - \frac{1}{2} \right) \frac{\pi}{n} \right\}$$

for $k = 1, \ldots, n$. The current carried by the conductors alternates in sign from one to the other so that as $z \to z_k$, the stream function $\psi \to (-1)^k \infty$. These limits are represented by the points P and Q in Figure 3.

The complex velocity is defined by

$$\zeta \equiv \frac{\mathrm{d}f}{\mathrm{d}z} = u - \mathrm{i}v. \tag{3.4}$$

Taking the limit as $z \to \infty$, we note the behavior of the complex velocity near the origin in the transformed plane,

$$\zeta \sim O\Big(t^{\frac{n+1}{n}} \Big) \quad \text{as} \quad t \to 0. \tag{3.5}$$

Consistent with the local flow at the contact points B and D shown in Figure 1, the complex velocity has the following singular behavior in the t plane:

$$\zeta \sim O\Big((1 - t^2)^{2(1-\beta/\pi)} \Big) \quad \text{as} \quad t \to \pm 1. \tag{3.6}$$

As shown in Figure 3, the points P and Q are mapped onto the points $t = -\mathrm{i}\Lambda$ and $t = \mathrm{i}\Lambda$, respectively, where

$$\Lambda = -\lambda + \sqrt{\lambda^2 + 1}. \tag{3.7}$$

Accordingly, the complex velocity has the local structure,

$$\zeta \sim O\Big((t^2 + \Lambda)^{-1} \Big) \quad \text{as} \quad t \to \pm \mathrm{i}\Lambda. \tag{3.8}$$

Our aim is to construct a function inside the unit circle in the transformed plane, which is analytic except at the singularities just discussed. To this end, we expand the complex velocity in an infinite series, writing

$$\zeta = t^{1/n} \, \frac{(1 - t^2)^{2(1-\beta/\pi)}}{(t^2 + \Lambda)} \, S(t), \qquad S(t) = \sum_{k=1}^{\infty} a_k t^k, \tag{3.9}$$

where the a_k are unknown, real coefficients. It will be noted that, while the magnetic shaping problem requires n to be an integer, no such restriction is necessary for the vortex problem. In the latter case, where the corner angle $\alpha = \pi/n$, n may be taken to be any positive real number. In the limit when the vortex tends to infinity studied above, the complex velocity given in (3.9) has the leading order behavior, $\zeta \sim a_1 \, t^{(1/n)-1} \, (1 - t^2)^{2(1-\beta/\pi)}$, in agreement with the analysis of Ozugurlu and Vanden-Broeck [12].

Since the coefficients, a_k, in (3.9) are real, the boundary conditions (2.4) are satisfied automatically and it only remains to enforce the dynamic condition (2.2). In practice, the infinite series is terminated after M terms, leaving M unknown coefficients $a_1, \ldots, a_M$. These are determined numerically by placing M collocation

points along AB and DE which comprise the free surface. Writing $t = \mathrm{e}^{\mathrm{i}\,\sigma}$, where $0 < \sigma < \pi$, we select the M points $\sigma_k = (k - 1/2)h$ for $k = 1, \ldots, M$, where the step length $h = \pi/M$. Noting the identity,

$$\frac{\partial x}{\partial \phi} + \mathrm{i}\,\frac{\partial y}{\partial \phi} = \frac{1}{u - \mathrm{i}v} = \frac{u + \mathrm{i}v}{u^2 + v^2}, \tag{3.10}$$

we may write the curvature in the convenient form

$$\kappa = \frac{vu' - uv'}{\sqrt{u^2 + v^2}}\,\frac{\partial \sigma}{\partial \phi}, \tag{3.11}$$

where a prime denotes partial differentiation with respect to σ. According to the mapping (3.1),

$$\frac{\partial \sigma}{\partial \phi} = \frac{\pi}{\Gamma}\,\frac{\lambda^2 + \cos^2\sigma}{\lambda \sin\sigma} \tag{3.12}$$

on the unit circle. Substituting (3.11) and (3.12) into the dynamic condition (2.2) and applying the result at each of the collocation points, we derive M nonlinear algebraic equations for the M unknowns $a_1, \ldots, a_M$. These equations are solved numerically using Newton's method. Once the coefficients are known, the shapes are constructed by integrating the identity (3.10). The position of the vortex is given by $z_v = z_A + Z$, where z_A is the position of the midpoint, A, on the bubble surface, and Z is given by

$$Z = \frac{2}{\pi}\lambda\Gamma\,\mathrm{i}\int_1^{\Lambda} \frac{(1 + s^2)^{(\frac{2\beta}{\pi} - 1)}}{(-\mathrm{i}s)^{\frac{1+n}{n}}(s^2 - \hat{\Lambda}^2)S(-is)}\,\mathrm{d}s, \tag{3.13}$$

where $\hat{\Lambda} = \lambda + \sqrt{\lambda^2 + 1}$.

Solutions with a cusp at some point along their perimeter can be described by a simple, exact solution for a free surface with zero surface tension and $\beta = \pi$. According to (2.2), such solutions have constant velocity on the bubble surface. They can be constructed by taking $a_1 = U$, $a_2 = 0$, $a_3 = U\Lambda^2$, and $a_k = 0$ for all $k \geq 4$ in (3.9). This guarantees that $|\zeta| = U$, a constant, on the bubble surface $|t| = 1$. Successful comparison is made below between this exact solution and numerically computed cusp solutions for the magnetic shaping problem.

4. Results. We present results for each of the two different physical problems in turn. For the vortex problem, the contact angle, β, is a free parameter. For the metal column shaping problem, β is equal to $\pi/2$. All solutions were computed using the method described in section 3. In most cases, we took $M = 150$ collocation points along the free surface. This was found to be sufficient to accurately resolve the bubble and liquid metal column shapes. In exceptional cases, we needed to take $M = 200$ to obtain an accurate solution. An example set of coefficients for a sample calculation is presented below.

We begin by presenting results for corner flow past a bubble. To define dimensionless variables, we set $T = \rho = \Gamma = 1$. This leaves a two-parameter family of solutions obtained by varying γ and $\mathcal{B}$. In Figure 4(a), we show a number of bubble shapes with contact angle $\beta = \pi/2$ for the case $\mathcal{B} = 1$ and for various values of γ. The vortex is indicated by a solid disk for each bubble shape. As γ is reduced, the vortex moves towards the origin, forcing the bubble towards the walls until it eventually makes contact at $\gamma = 0.149$. Beyond this point, the free surface intersects the

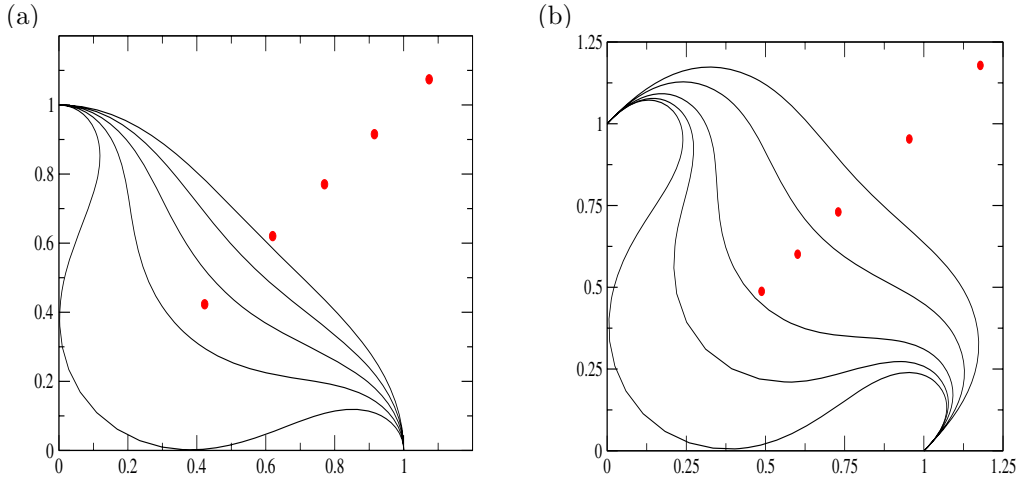


FIG. 4. *Flow past a bubble in a right-angled corner for* $\mathcal{B} = 1$, $T = \rho = \Gamma = 1$*:* (a) $\beta = \pi/2$ *and* 0.55 *(outermost),* 0.45, 0.35, 0.25, *and* 0.149 *(innermost);* (b) $\beta = \pi/4$ *and* $\gamma = 0.4$ *(outermost),* 0.3, 0.2, 0.15, 0.113 *(innermost). The position of the vortex for each case is shown as a solid disk. The vortex moves closer to the origin as* γ *decreases. Each picture is scaled so that each bubble makes contact with the walls a unit distance from the origin.*

walls and any physical meaning is lost. For the last case, $\gamma = 0.149$, the coefficients a_k in (3.9) are given by $a_1 = -2.163$, $a_3 = -1.038$, $a_5 = 0.3103$, ..., $a_{11} = -0.01918$, ..., $a_{21} = -6.375 \times 10^{-4}$, ..., $a_{41} = -3.748 \times 10^{-5}$, ..., $a_{61} = -1.2561 \times 10^{-6}$, ..., $a_{81} = -5.75 \times 10^{-8}$. All of the even coefficients are zero. Bubble shapes under the same conditions but with a contact angle of $\beta = \pi/4$ for various values of γ are shown in Figure 4(b). We note that fixing γ and varying the Bernoulli constant $\mathcal{B}$ produces little qualitative difference in the bubble shapes.

Free surface shapes inside a corner of angle $\alpha = \pi/3$ are displayed in Figure 5(a) for the contact angle $\beta = \pi/2$. As γ increases, the vortex moves further into the corner until contact is made between the free surface and the walls. For the contact angle $\beta = \pi$, considered in Figure 5(b), intersection between the bubble and the walls tends to occur much closer to the contact points. As γ increases and the vortex moves a long distance from the corner, the free surface tends to a circular arc, except in small regions close to the walls, where the bubble meets the wall tangentially. The circular arc is the expected shape for a bubble in a stagnant fluid. It is shown as a broken line in Figure 5(b).

Results for flow past a bubble adhering to a plane wall are displayed in Figure 6 for $\mathcal{B} = 0.5$. As γ decreases and the vortex moves towards the wall, the bubble is squashed downwards until its midpoint touches the wall at $\gamma = 0.247$. Beyond this value, the free surface intersects the wall, preventing physical interpretation. As γ increases and the vortex moves away from the wall, the bubble expands and eventually acquires the semicircular shape expected for an attached bubble residing in a quiescent fluid. If, simultaneously, the vortex strength tends to infinity, we obtain uniform flow past a bubble on a flat wall, as discussed by Vanden-Broeck and Keller [17].

We now turn to the liquid metal column shaping problem. We start by recomputing some of Shercliff's results using our simplified conformal mapping. First, we nondimensionalize (2.6), writing $q = B_s q^*$, where B_s is a reference magnetic field strength, and $\kappa = \kappa^*/L$, where L is a reference length. Dropping the asterisks for

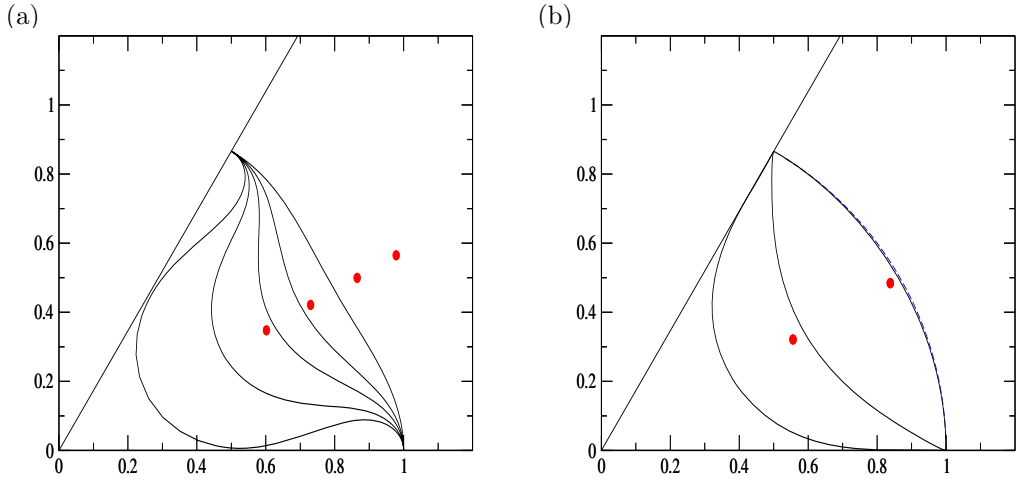


FIG. 5. *Flow past a bubble in a corner with* $\alpha = \pi/3$ *for* $\mathcal{B} = 1$, $T = \rho = 1$*:* (a) $\beta = \pi/2$ *and* $\gamma = 0.6$ *(outermost),* 0.4, 0.3, 0.2, *and* 0.135 *(innermost);* (b) $\beta = \pi$ *for* $\gamma = 0.27$ *(innermost),* 0.5, 0.99 *(outermost) together with the limiting circular arc, shown as a broken line. The position of the vortex for each case (except for* (b) $\gamma = 0.99$*) is shown as a solid disk. The vortex moves closer to the origin as* γ *decreases. Each picture is scaled so that each bubble makes contact with the walls a unit distance from the origin.*

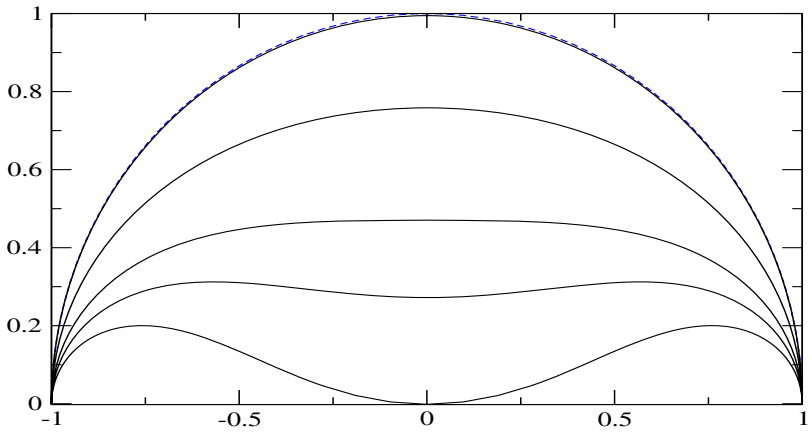


FIG. 6. $T = \rho = 1$ *and* $\beta = \pi/2$*: Flow past a bubble on a flat wall with* $\mathcal{B} = 0.5$ *for* $\gamma =$ 0.247 *(innermost),* 0.35, 0.45, 0.65, 0.98 *(outermost), shown with the limiting circle (broken line).*

convenience, (2.6) becomes

$$\frac{1}{2}q^2 - \frac{\kappa}{2ak} = \frac{1}{2a}, \tag{4.1}$$

where

$$a = \frac{B_s^2}{2\mathcal{E}}, \qquad k = \frac{\rho\mathcal{E}L}{T} \tag{4.2}$$

are the same as Shercliff's dimensionless parameters a and k. Shercliff's parameter, α, to be written here as α_S, is related to our parameter γ by $\alpha_S = \Gamma/2 - \gamma$, where Γ

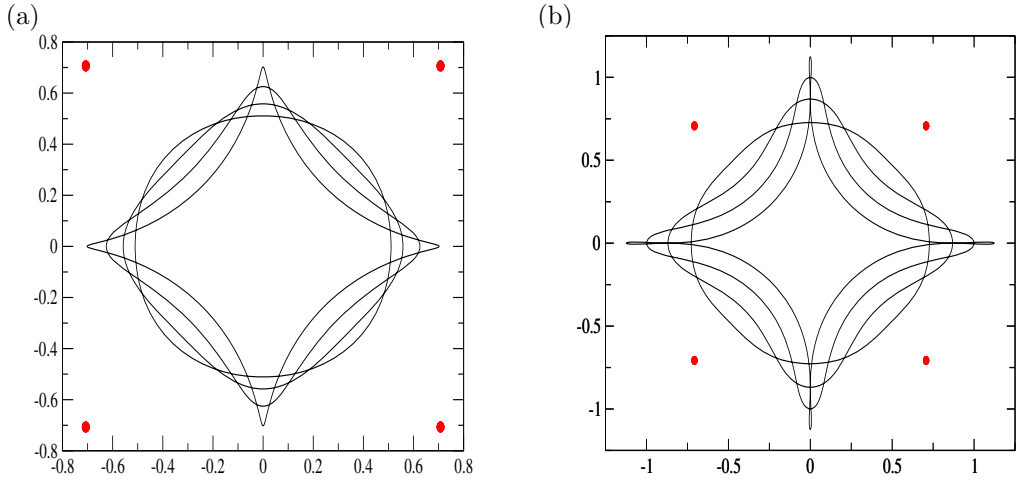


FIG. 7. *Column shapes for four conductors for* (a) $\sin\alpha_S = 0.5$ *and* $a = 1$, 10, 100, 10000, *and* (b) $\sin\alpha_S = 0.8$ *and* $a = 2.5$, 25, 250, 25000. *The largest value of* a *corresponds to the limiting shape exhibiting a cusp and a trapped bubble, respectively. In both figures, the shapes are scaled so that the conductors, shown as four solid dots, are always a unit distance from the origin.*

now represents the dimensionless circulation. To conform with Shercliff's results, we take $k = 0.5$ and $\Gamma = \pi$ throughout.

In Figure 7, we plot column shapes for a number of values of a for the two cases $\sin(\alpha_S) = 0.5$ and $\sin(\alpha_S) = 0.8$, corresponding to Shercliff's Figures 12 and 13. As a increases, the shapes in Figure 7(a) tend to a limiting cusped configuration, while those in Figure 7(b) approach a shape with trapped bubbles along the perimeter. The critical value dividing these two characteristic limits is provided by Shercliff as $\sin\alpha_S = 0.66$. In either the subcritical or supercritical case, the limiting configuration is easily obtained as the exact solution for zero surface tension discussed at the end of section 3. To confirm this statement, we temporarily drop the current nondimensionalization and, in Figure 8(a), present the numerical solution for the case $\mathcal{E} = \rho = 1$ and $T = 10^{-6}$ with contact angle $\beta = \pi/2$. According to (2.6), the surface velocity is approximately $\sqrt{2}$ except near the cusps, where the curvature is large. The exact cusp solution with $U = \sqrt{2}$ is shown as the broken line in Figure 8(a). The agreement between the two curves is excellent. A close-up view in Figure 8(b) illustrates the true cusplike nature of the exact solution in contrast to the sharp turning of the numerical solution.

The effect on the cusp shape of increasing α_S is shown in Figure 9. As α_S is raised, the cusps eventually intersect themselves. The first intersection occurs at the critical value $\sin\alpha_{S_c} = 0.66$, which is in excellent agreement with Shercliff's prediction. Typical subcritical and supercritical profiles are shown in the figure. The behavior of the exact cusp solutions suggests that, as a increases, metal columns for which $\alpha_S < \alpha_{S_c}$ will approach a limiting physical shape exhibiting trapped bubbles, and those for which $\alpha_S > \alpha_{S_c}$ will approach a cusped configuration. These predictions are confirmed by the shapes shown in Figure 7. A magnified view of one of the trapped bubbles seen in Figure 7(a) is shown in Figure 10(a). The shape eventually intersects itself when a is increased further, as can be seen in Figure 10(b).

Shercliff presented results for only four conductors. Our numerical method allows us to compute shapes for any even number of conductors. By way of illustration, we

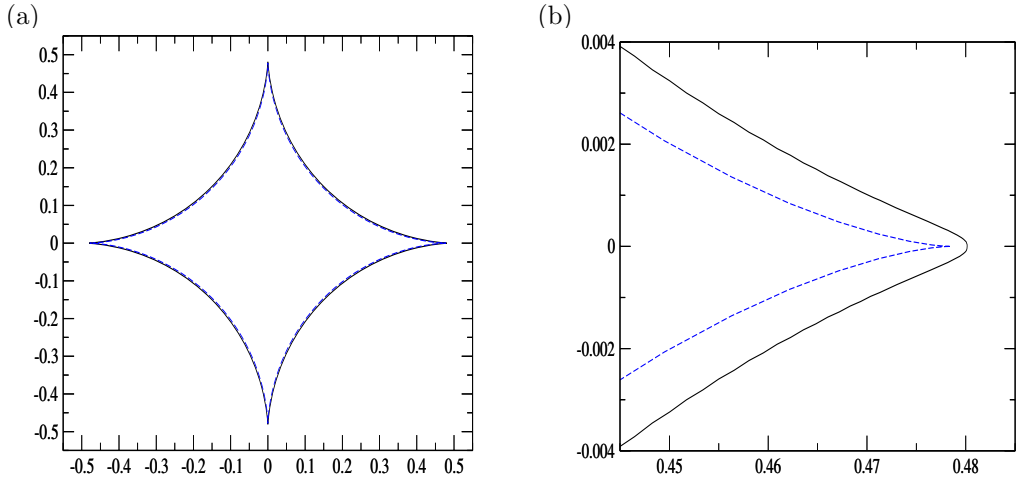


FIG. 8. $\sin\alpha_S = 0.5$: (a) *Numerical solution with* $\beta = \pi/2$ *and* $\mathcal{E} = \rho = 1$, $T = 10^{-6}$, *shown as a solid line, compared with the exact cusp solution with* $\beta = \pi$ *and* $U = \sqrt{2}$, *shown as a broken line.* (b) *Close up of* (a) *near the cusp.*

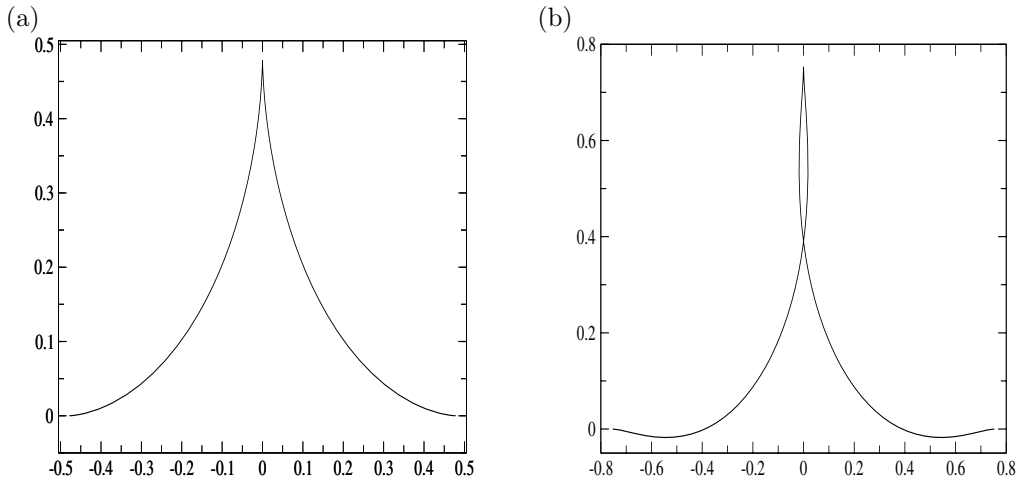


FIG. 9. *Exact solutions with zero surface tension and* $\beta = \pi$, $U = \sqrt{2}$: (a) *Subcritical cusp shape for* $\sin\alpha_S = 0.5$, *and* (b) *supercritical cusp shape for* $\sin\alpha_S = 0.8$.

show in Figure 11 the various column shapes produced by six regularly placed conductors. Computing the exact cusp solution for zero surface tension as described above, we find the critical value $\sin\alpha_S = 0.75$. Subcritical values correspond to limiting cusp solutions appropriate to Figure 11(a), and supercritical values correspond to shapes with trapped bubbles appropriate to Figure 11(b).

5. Summary. We have considered the free surface shapes adopted by a bubble in a vortex corner flow or by a column of liquid metal surrounded by an even number of conductors, and have noted the near mathematical equivalence between these two problems. In both cases, our approach has been to use a conformal map to transform the domain of interest into the unit circle. In the process, the kinematic condition on the walls in the case of the vortex flow, or the symmetry conditions in the case of the

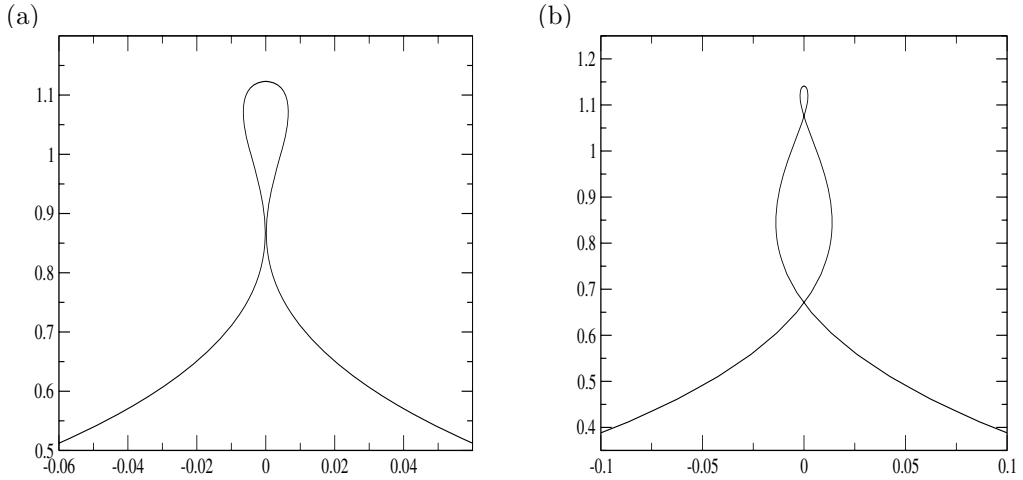


FIG. 10. *The case* $\sin\alpha_S = 0.8$*:* (a) *A close-up view of the trapped bubble shown in Figure* 7(b) *when* $a = 25000$*, and* (b) *a self-intersecting shape when* $a = 2 \times 10^5$.

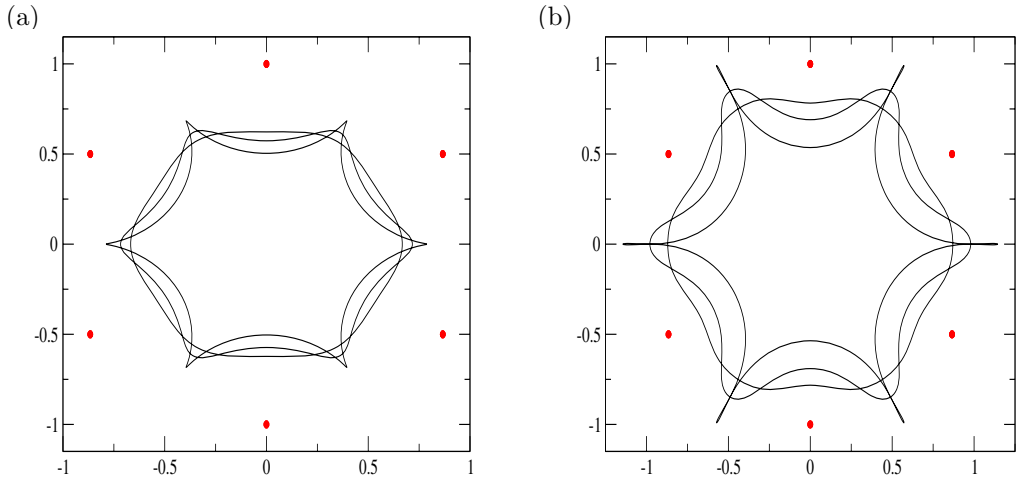


FIG. 11. *Column shapes for six conductors for* (a) $\sin\alpha_S = 0.5$ *and* $a = 5,\ 50,\ 5 \times 10^4$*, and* (b) $\sin\alpha_S = 0.85$ *and* $a = 3.8,\ 38,\ 3.8 \times 10^4$. *The largest value of* a *corresponds to the limiting shape exhibiting a cusp and a trapped bubble, respectively. In both figures, the shapes are scaled so that the conductors, shown as six solid dots, are always a unit distance from the origin.*

metal column, were satisfied automatically, and the free surface condition was satisfied by expanding in a convergent power series, which was truncated to permit numerical computation. For the magnetic shaping problem, our approach has considerably simplified the analysis used by Shercliff to compute the various shapes.

For both problems, we have presented a variety of shapes over a range of conditions and have demonstrated the limiting configurations featuring either cusps or trapped bubbles along the free surface perimeter. Moreover, we have constructed exact zero surface tension cusp solutions and have shown that as one parameter is varied, these solutions begin to intersect themselves. The critical value at which this self-intersection occurs corresponds to the division between limiting cusp and trapped

bubble solutions of the free surfaces with surface tension present. Below the critical value, the bubble shapes approach a cusped configuration; above it, the free surfaces tend to a shape with trapped bubbles along its perimeter.

The issue of the stability of the configurations computed in this paper is left as a topic for future work. Stability calculations for related problems involving liquid metals have been performed by Felici [9].

REFERENCES

[1] G. R. Baker and D. W. Moore, *The rise and distortion of a two-dimensional bubble in an inviscid liquid*, Phys. Fluids, A1 (1989), pp. 1451–1459.

[2] M. G. Blyth and J.-M. Vanden-Broeck, *New solutions for capillary waves on fluid sheets*, J. Fluid Mech., 507 (2004), pp. 255–264.

[3] M. G. Blyth and J.-M. Vanden-Broeck, *New solutions for capillary waves on curved sheets of fluid*, IMA J. Appl. Math., 70 (2005), pp. 588–601.

[4] D. G. Crowdy, *Steady nonlinear capillary waves on curved sheets*, European J. Appl. Math., 12 (2001), pp. 689–708.

[5] D. G. Crowdy, *Circulation-induced shape deformations of drops and bubbles: Exact two-dimensional models*, Phys. Fluids, 11 (1999), pp. 2836–2845.

[6] F. Dias and J.-M. Vanden-Broeck, *Flows emerging from a nozzle and falling under gravity*, J. Fluid Mech., 123 (1990), pp. 465–477.

[7] J. Etay, *Formage et guidage des métaux liquides sous l'action de champs magnétiques alternatifs*, Rep. de DEA de Méchanique des Fluides., Inst. Nat. Polytechnique, Grenoble, 1980.

[8] J. Etay, A. Gagnoud, and M. Garnier, *Le problème de frontièr libre en lévitation électromagnétique*, J. Mec. Theor. Appl., 5 (1986), pp. 911–934.

[9] T. P. Felici, *On the surface stability of liquid conductors in electromagnetic shaping*, J. Fluid Mech. (1995), pp. 1–28.

[10] E. B. McLeod, *The explicit solution of a free boundary problem involving surface tension*, J. Rational Mech. Anal., 4 (1955), pp. 557–567.

[11] M. Miksis, J.-M. Vanden-Broeck, and J. B. Keller, *Axisymmetric bubble or drop in a uniform flow*, J. Fluid Mech., 108 (1981), pp. 89–100.

[12] E. Ozugurlu and J.-M. Vanden-Broeck, *The distortion of a bubble in a corner flow*, European J. Appl. Math., 11 (2000), pp. 171–179.

[13] P. N. Shankar, *On the shape of a two-dimensional bubble in uniform motion*, J. Fluid Mech., 244 (1992), pp. 187–200.

[14] J. A. Shercliff, *Magnetic shaping of molten metal columns*, Proc. Roy. Soc. London Ser. A, 375 (1981), pp. 455–473.

[15] S. Tanveer, *Some analytical properties of solutions to a two-dimensional steadily translating inviscid bubble*, Proc. Roy. Soc. London Ser. A, 452 (1996), pp. 1397–1410.

[16] J.-M. Vanden-Broeck, *Rising bubbles in a two-dimensional tube with surface tension*, Phys. Fluids, 27 (1984), pp. 2604–2607.

[17] J.-M. Vanden-Broeck and J. B. Keller, *Deformation of a bubble or drop in a uniform flow*, J. Fluid Mech., 101 (1980), pp. 673–686.

[18] J.-M. Vanden-Broeck and J. B. Keller, *Bubble or drop distortion in a straining flow in two dimensions*, Phys. Fluids, 23 (1980), pp. 1491–1495.

[19] J.-M. Vanden-Broeck and T. Miloh, *Computations of steep gravity waves by a refinement of Davies–Tulin's approximation*, SIAM J. Appl. Math., 55 (1995), pp. 892–903.

[20] N. M. Zubarev, *Exact solution of the problem of the equilibrium configuration of the charged surface of a liquid metal*, J. Exp. Theor. Phys., 89 (1999), pp. 1078–1085.

[21] N. M. Zubarev and O. V. Zubarev, *Exact solutions for equilibium configurations of charged conducting liquid jets*, Phys. Rev. E, 71 (2005), article 016307.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 187–205

AN APPROXIMATE METHOD FOR SCATTERING BY THIN STRUCTURES*

S. MOSKOW†, F. SANTOSA‡, AND J. ZHANG‡

Abstract. Scattering of waves by a thin structure is considered in this work. The Helmholtz equation with variable coefficient models the wave phenomena. The scatterer is assumed to have a high index of refraction while at the same time it is very small in one of the dimensions. We show that if the index scales as $O(1/h)$, where h is the thickness of the scatterer, then an approximate solution, based on perturbation analysis, can be obtained. The approximate solution consists of a leading order term plus a corrector, each of which solves an integral equation in two dimensions for a three-dimensional problem. We provide error analysis on the approximation. The approximate method can be viewed as an efficient computational approach since it can potentially greatly simplify scattering calculations. Numerical results provide an assessment of the accuracy of the approximate solution.

Key words. scattering, Helmholtz equation, approximate solution, asymptotics, error estimates

AMS subject classifications. 65R20, 78A45, 45E99, 34E10, 78M99

DOI. 10.1137/040617388

1. Introduction.

The problem under investigation arises in the study of photonic band gap (PBG) structures. Optical devices that exploit photonic band gap phenomena to guide and manipulate light are expected to play an important role in optical communication networks and optical computing. Thin-film or membrane devices are particularly attractive because of the relative ease with which they can be made.

A typical thin-film device is made of a material with a high index of refraction. The high index is needed to confine light within the structure. To manipulate light within the structure, holes are drilled into film. Typical structures under study can be found in several recent papers [3, 5, 6].

In order to simulate how light behaves in such a structure, it is necessary to solve the wave equation. In most studies, the structure is surrounded by air. Thus, the domain in which the wave equation must be solved will be all of $\mathbb{R}^3$. The thin film structure can be modeled by prescribing index of refraction to a subdomain of $\mathbb{R}^3$.

The classical approach to performing the required simulation of wave propagation in such a complicated structure is the finite-difference time-domain (FDTD) method [2], with absorbing boundary conditions. While the computation proceeds in a straightforward manner, it is very time consuming.

In this paper, we propose an approximate method to solve the scattering problem. The method starts with the time-harmonic wave equations and applies a perturbation approach based on an identified small parameter. The advantage of our method is that it reduces the complexity of the computation by one dimension. The Lippmann–Schwinger formulation of the scattering problem will involve a three-dimensional

*Received by the editors October 20, 2004; accepted for publication (in revised form) May 16, 2005; published electronically October 17, 2005.
http://www.siam.org/journals/siap/66-1/61738.html
†Department of Mathematics, University of Florida, Gainesville, FL 32611 (moskow@math.ufl.edu).
‡School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (santosa@math.umn.edu, jzhong@math.umn.edu). The research of the second author was supported in part by the National Science Foundation.

(3-D) (volume) integral equation. Our method simplifies the calculation to solving a sequence of two-dimensional (2-D) integral equations.

The present work addresses only the case of the scalar wave equation. Maxwell's equation, which is the correct model for the propagation phenomena, will be treated in a separate, future work.

This paper is organized as follows. We give a description of the problem we wish to solve in the next section. The perturbation approach is presented in section 3. Justification of the approximate method follows in sections 4 and 5. Section 6 contains numerical examples in two dimensions. The paper closes with a discussion.

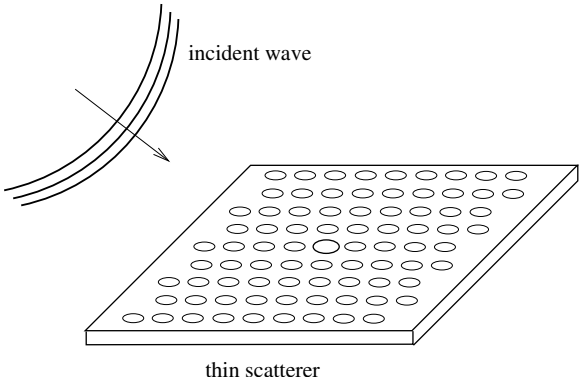


FIG. 2.1. *Scattering by a thin structure.*

2. Problem statement. The situation we are attempting to model is depicted in Figure 2.1. The propagation of waves is modeled by the Helmholtz equation

$$\Delta u + k^2 \epsilon(x,z) u = 0, \tag{2.1}$$

where $x = (x_1, x_2) \in \mathbb{R}^2$ and $z \in \mathbb{R}$. In (2.1) we call $\epsilon(x,z)$ the dielectric constant (an abuse in terminology), which will be set to unity in air and to some value in the structure. The field $u(x,z)$ is of normalized real frequency k and comprises two components:

$$u(x,z) = u_{\mathrm{i}}(x,z) + u_{\mathrm{s}}(x,z),$$

where the incident wave u_{i} is given, and the scattered field u_{s} satisfies the Sommerfeld radiation conditions

$$\frac{\partial u_{\mathrm{s}}}{\partial r} - iku_{\mathrm{s}} = O\left(\frac{1}{r^2}\right) \quad \text{for } r = \sqrt{|x|^2 + z^2} \to \infty.$$

The thin structure is incorporated into the definition of $\epsilon(x,z)$. Let $\Omega \times (-h/2, h/2)$ be the region occupied by the thin structure, where Ω is a bounded domain in $\mathbb{R}^2$. Then the dielectric constant is defined by

$$\epsilon(x,z) = \begin{cases} 1 & \text{for} \quad |z| > h/2, \\ \epsilon_0(x)/h & \text{for} \quad |z| < h/2, x \in \Omega, \\ 1 & \text{for} \quad |z| < h/2, x \notin \Omega. \end{cases} \tag{2.2}$$

Thus, we have assumed that $\epsilon_0(x)$ is supported in Ω.

The problem we wish to solve is to find the scattered field $u_{\rm s}(x,z)$ given a thin structure (2.2) and an incident field $u_{\rm i}(x,z)$. For the purpose of simulating waves in PBG structures, the total field $u(x,z)$ for (x,z) in the thin structure, $\Omega\times(-h/2,h/2)$, is of great interest.

3. Perturbation approach. One can rewrite the problem for $u(x,z)$ in (2.1) as an integral equation [1] by observing that since

$$\Delta u_{\rm i}+k^2u_{\rm i}=0,$$

we have that

$$\Delta u_{\rm s}+k^2u_{\rm s}=k^2(1-\epsilon)u.$$

If $G(x,z,x',z')$ is the free space Green's function for the Helmholtz equation in $\mathbb{R}^3$, i.e., G satisfies

$$(\Delta_{(x',z')}+k^2)G=\delta_{(x,z)}$$

in $\mathbb{R}^3$ with the Summerfeld radiation condition, then by using integration by parts and the decay at infinity, we get

$$u_{\rm s}(x,z)=k^2\int_\Omega\int_{-h/2}^{h/2}\left(1-\frac{\epsilon_0(x')}{h}\right)G(x,z,x',z')u(x',z')dz'dx'.$$

Hence we have that the field satisfies the well-known Lippmann–Schwinger integral equation

$$u(x,z)=u_{\rm i}(x,z)+k^2\int_\Omega\int_{-h/2}^{h/2}\left(1-\frac{\epsilon_0(x')}{h}\right)G(x,z,x',z')u(x',z')dz'dx'. \tag{3.1}$$

We note that $G(x,z,x',z')$ is given by

$$G(x,z,x',z')=\frac{1}{4\pi}\frac{e^{ik\sqrt{|x-x'|^2+|z-z'|^2}}}{\sqrt{|x-x'|^2+|z-z'|^2}}.$$

See [1] for the full proof of equivalence. To solve for the field, we need to view (3.1) as an integral equation satisfied by $u(x,z)$ for $(x,z)\in\Omega\times(-h/2,h/2)$. Once we have found $u(x,z)$ for (x,z) in the thin domain, we can then use (3.1) as a way to compute the field outside the thin domain. Therefore, our first step will be to find an asymptotic approximation for u inside the thin region.

To find a first order approximation, we will scale the variable in the z direction, $z=h\zeta$, so that the integral equation is now

$$u(x,\zeta)=u_{\rm i}(x,h\zeta)+k^2\int_\Omega\int_{-1/2}^{1/2}\left(1-\frac{\epsilon_0(x')}{h}\right)G(x,h\zeta,x',h\zeta')u(x',\zeta')hd\zeta'dx', \tag{3.2}$$
$$(x,\zeta)\in\Omega\times(-1/2,1/2).$$

Formally, we assume a perturbation series ansatz

$$u(x,\zeta)=u_0(x)+hu_1(x,\zeta)+\cdots,\quad (x,\zeta)\in\Omega\times[-1/2,1/2]. \tag{3.3}$$

The goal is now to obtain equations by which $u_0(x,z)$ and $u_1(x,z)$ can be found.

Substituting (3.3) into (3.2), we see that

$$u_0(x) + hu_1(x,\zeta) + \cdots = u_{\mathrm{i}}(x,0) + h\frac{\partial u_{\mathrm{i}}}{\partial z}(x,0) + O(h^2) + k^2 \int_\Omega \int_{-1/2}^{1/2} (h - \epsilon_0(x')) G(x, h\zeta, x', h\zeta') \left[u_0(x') + hu_1(x',\zeta') + \cdots\right] d\zeta' dx'. \tag{3.4}$$

The classical way to find u_0 and u_1 is to equate like powers of h on both sides of (3.4). However, the situation is complicated by the fact that $G(x,z,x',z')$ is singular.

We make the observation that if the integral

$$\int_{-1/2}^{1/2} G(x, h\zeta, x', h\zeta') d\zeta'$$

converges as $h \to 0$ to

$$G(x,0,x',0),$$

then, setting equal like powers of h in (3.4), we see that $u_0(x)$ satisfies the integral equation

$$u_0(x) = u_{\mathrm{i}}(x,0) - k^2 \int_\Omega \epsilon_0(x') G(x,0,x',0) u_0(x') dx'. \tag{3.5}$$

We will justify this step in the next section and show that $u_1(x,z)$ can be calculated, as well as justified, in section 5.

Once we have solved for $u_0(x)$ and $u_1(x,z)$, we can insert them in the right-hand side of (3.1) to obtain an approximation of the field for all points outside the thin domain. Note that the equation for u_0 and, as we shall see, that for u_1 are 2-D integral equations. Therefore, in terms of computational cost we have reduced the dimension of the problem by one.

4. Justification for the first term. Now we provide a rigorous error estimate for the approximation (3.5) derived above. The solution $u(x,z)$ satisfies

$$u(x,z) = u_{\mathrm{i}}(x,z) + k^2 \int_\Omega \int_{-h/2}^{h/2} \left(1 - \frac{\epsilon_0(x')}{h}\right) G(x,z,x',z') u(x',z') dz' dx'. \tag{4.1}$$

The candidate for an approximation to u on the thin strip to substitute into (3.1) is $u_0(x)$, the solution to the lower dimensional problem:

$$u_0(x) = u_{\mathrm{i}}(x,0) - k^2 \int_\Omega \epsilon_0(x') G(x,0,x',0) u_0(x') dx'. \tag{4.2}$$

To show this is a good approximation, let

$$\zeta = z/h \quad \text{and} \quad \tilde{u}(x,\zeta) = u(x,z),$$

so that

$$\tilde{u}(x,\zeta) = u_{\mathrm{i}}(x,h\zeta) + k^2 \int_\Omega \int_{-1/2}^{1/2} (h - \epsilon_0(x')) G(x, h\zeta, x', h\zeta') \tilde{u}(x',\zeta') d\zeta' dx'. \tag{4.3}$$

For convenience define S to be the scaled strip

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

We will show the following uniform norm estimate.

PROPOSITION 1. *There exists a constant C independent of h (but depending on k) such that*

$$\|u_0(x) - \tilde{u}(x,\zeta)\|_{L^\infty(S)} \leq Ch.$$

Using (4.2) and (4.3) and interchanging the order of integration, we can write

$$\tilde{u}(x,\zeta) - u_0(x) = u_{\mathrm{i}}(x,h\zeta) - u_{\mathrm{i}}(x,0) + hk^2 \int_\Omega \int_{-1/2}^{1/2} G(x,h\zeta,x',h\zeta')\tilde{u}(x',\zeta')d\zeta' dx'$$
$$+ k^2 \int_{-1/2}^{1/2} \int_\Omega \epsilon_0(x') \left[G(x,0,x',0)u_0(x') - G(x,h\zeta,x',h\zeta')\tilde{u}(x',\zeta')\right] dx' d\zeta'.$$

We add and subtract appropriate terms to obtain

$$\tilde{u}(x,\zeta) - u_0(x) = u_{\mathrm{i}}(x,h\zeta) - u_{\mathrm{i}}(x,0) + hk^2 \int_\Omega \int_{-1/2}^{1/2} G(x,h\zeta,x',h\zeta')\tilde{u}(x',\zeta')d\zeta' dx'$$
$$+ k^2 \int_{-1/2}^{1/2} \int_\Omega \epsilon_0(x')\tilde{u}(x',\zeta') \left[G(x,0,x',0) - G(x,h\zeta,x',h\zeta')\right] dx' d\zeta'$$
$$+ k^2 \int_{-1/2}^{1/2} \int_\Omega \epsilon_0(x') G(x,0,x',0) \left[u_0(x') - \tilde{u}(x',\zeta')\right] dx' d\zeta'.$$

For a given $\epsilon_0 \in L^\infty(\Omega)$ which is also piecewise continuous, define the integral operator

$$T : L^2(S) \to L^2(S)$$

by

$$T(f) = \int_{-1/2}^{1/2} \int_\Omega \epsilon_0(x') G(x,0,x',0) f(x',\zeta') dx' d\zeta'.$$

By an abuse of notation, we will also use T to denote the same integral operator on the space of continuous functions, $C^0(\overline{S})$, equipped with the L^∞ norm:

$$T : C^0(\overline{S}) \to C^0(\overline{S}).$$

(Note that $T(f)$ will always be independent of ζ, so the range of T is really only functions on Ω.) Then $\tilde{u} - u_0$ satisfies

$$(I + k^2 T)(\tilde{u} - u_0) = u_{\mathrm{i}}(x,h\zeta) - u_{\mathrm{i}}(x,0) + hk^2 \int_\Omega \int_{-1/2}^{1/2} G(x,h\zeta,x',h\zeta')\tilde{u}(x',\zeta')d\zeta' dx'$$
$$+ k^2 \int_{-1/2}^{1/2} \int_\Omega \epsilon_0(x')\tilde{u}(x',\zeta') \left[G(x,0,x',0) - G(x,h\zeta,x',h\zeta')\right] dx' d\zeta'. \tag{4.4}$$

The lemmas that follow are used to bound the right-hand side of (4.4) and to show that we can invert $(I + k^2 T)$.

LEMMA 1. *There exists a constant C independent of h and ζ' but depending on k such that*

$$\sup_{(x,\zeta)\in S}\int_\Omega |G(x,0,x',0)-G(x,h\zeta,x',h\zeta')|dx' \le Ch.$$

Proof. The difference of these Green's functions can be written as

$$\begin{aligned}
&G(x,0,x',0)-G(x,h\zeta,x',h\zeta')\\
&=\frac{1}{4\pi}\frac{e^{ik|x-x'|}}{|x-x'|}-\frac{1}{4\pi}\frac{e^{ik\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}\\
&=\frac{1}{4\pi}e^{ik|x-x'|}\left[\frac{1}{|x-x'|}-\frac{1}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}\right]\\
&\qquad+\frac{1}{4\pi}\frac{1}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}\left[e^{ik|x-x'|}-e^{ik\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}\right].
\end{aligned}\tag{4.5}$$

We first work on the second term on the right-hand side of (4.5). Since we know that for $(x,\zeta)\in S$,

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

there exists a constant C independent of h and ζ' but depending on (real) k such that

$$|e^{ik|x-x'|}-e^{ik\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}|\le Ch.\tag{4.6}$$

Since

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

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

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

is bounded independently of h, ζ', and $(x,z)\in S$. This along with (4.6) gives that we can choose C independent of h, ζ', and $(x,\zeta)\in S$ such that

$$\int_\Omega\frac{1}{4\pi}\frac{1}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}|e^{ik|x-x'|}-e^{ik\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}|dx'\le Ch.$$

The integral of the first term on the right-hand side of (4.5) can be bounded as follows:

$$\begin{aligned}
&\int_\Omega\left|\frac{1}{4\pi}e^{ik|x-x'|}\left[\frac{1}{|x-x'|}-\frac{1}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}\right]\right|dx'\\
&\le\int_\Omega\left|\frac{1}{|x-x'|}-\frac{1}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}\right|dx'\\
&=\int_\Omega\left(\frac{1}{|x-x'|}-\frac{1}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}\right)dx'
\end{aligned}$$

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

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

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

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

Change to polar coordinates centered at x with

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

The integral transforms to

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

by direct calculation. This is then $O(h)$, where the constant is independent of $(x,\zeta) \in S$ and $\zeta' \in (-1/2, 1/2)$. This, combined with the bounds on the first integral, proves the lemma. □

Recall that our scaled domain S is given as

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

LEMMA 2. *Let $\epsilon_0(x)$ be piecewise continuous on Ω. Then the operator $T : L^2(S) \to L^2(S)$ given by*

$$(Tf)(x) = \int_S \epsilon_0(x') G(x,0,x',0) f(x',\zeta') dx' d\zeta',$$

where

$$G(x,0,x',0) = \frac{1}{4\pi} \frac{e^{ik|x-x'|}}{|x-x'|},$$

is compact. Moreover, if we view T on the Banach space of continuous functions,

$$T : C^0(\overline{S}) \to C^0(\overline{S}),$$

it is also a compact operator. Furthermore, $(I + k^2 T)$ is continuously invertible on both $L^2(S)$ and $C^0(\overline{S})$.

Proof. Since ϵ_0 is piecewise continuous, the kernel $\epsilon_0(x')G(x,0,x',0)$ is a finite sum of weakly singular kernels. Hence the fact that T is a compact operator on $C^0(\overline{S})$ follows from Theorem 1.11 of [1]. To show T is compact on $L^2(S)$, we will show that for any sequence $\{f_n\}$ such that $\|f_n\|_{L^2(S)} < M$ and $f_n \rightharpoonup 0$, the sequence $Tf_n \to 0$ in $L^2(S)$. Let

$$D := \{(x-x', \zeta-\zeta') \; ; \; (x,\zeta), (x',\zeta') \in S\}$$

and define

$$g(y,\eta) := e^{ik|y|}/|y| \text{ for } (y,\eta) \in D.$$

Then since $g \in L^1(D)$, there exists

$$g_m \in C^\infty(D) \quad \text{such that} \quad \|g - g_m\|_{L^1(D)} \to 0 \quad \text{as} \quad m \to \infty.$$

In the estimates that follow we use a standard result (see, for example, [7]). If $g \in L^1(D)$ and $f \in L^p(S)$ $(1 < p < \infty)$, then $(g * f) \in L^p(S)$ and

$$\|g * f\|_{L^p(S)} \le \|g\|_{L^1(D)} \|f\|_{L^p(S)}. \tag{4.7}$$

Now consider

$$\begin{aligned}
\|Tf_n\|^2_{L^2(S)} &= \int_S \left[\int_S \epsilon_0(x') G(x,0,x',0) f_n(x',\zeta') dx' d\zeta' \right]^2 dx d\zeta \\
&= \frac{1}{4\pi} \int_S \left[\int_S \epsilon_0(x') g(x-x',\zeta-\zeta') f_n(x',\zeta') dx' d\zeta' \right]^2 dx d\zeta.
\end{aligned}$$

Let $M_\epsilon = \|\epsilon_0\|_\infty$. By adding and subtracting g_m, we can bound the above by

$$\begin{aligned}
&\frac{M_\epsilon^2}{4\pi} \int_S \left[|g_m * f_n| + |(g - g_m) * f_n| \right]^2 dx d\zeta \\
&\le \frac{M_\epsilon^2}{2\pi} (\| g_m * f_n \|^2_{L^2(S)} + \| (g - g_m) * f_n \|^2_{L^2(S)}) \\
&\le \frac{M_\epsilon^2}{2\pi} (\| g_m * f_n \|^2_{L^2(S)} + \|g - g_m\|^2_{L^1(D)} M^2), \qquad (4.8)
\end{aligned}$$

with the last inequality obtained by using (4.7) with $p = 2$. Also, since

$$|g_m * f_n| \le \|g_m\|_{L^\infty(D)} \|f_n\|_{L^1(S)}$$

and $\|f_n\|_{L^1(S)}$ is bounded, by the Lebesgue dominated convergence theorem,

$$\lim_{n\to\infty} \|g_m * f_n\|_{L^2(S)} = \| \lim_{n\to\infty} (g_m * f_n) \|_{L^2(S)} = 0$$

since $\{f_n\}$ goes to zero weakly. For any given $\varepsilon > 0$, choose m large enough so that

$$\|g - g_m\|^2_{L^1(D)} M^2 < \varepsilon.$$

Then for this m we can choose n large enough that

$$\| g_m * f_n \|^2_{L^2(S)} < \varepsilon$$

also. Hence from (4.8)

$$\|Tf_n\|^2_{L^2(S)} \le C\varepsilon.$$

Therefore, $\lim_{n\to\infty} \|Tf_n\|^2_{L^2(S)} = 0$, and hence T is compact on $L^2(S)$.

Now, by the Fredholm theory (Corollary 1.17 of [1]), $I + k^2 T$ is invertible if $(I + k^2 T) f = 0$ has only the zero solution. The following argument holds on the Banach space X for both $X = L^2(S)$ and $X = C^0(\overline{S})$. Let $f \in X$ be a solution to

$(I + k^2T)f = 0$. Since for any $f \in X$, Tf depends only on $x \in \Omega$, the solution f satisfies

$$f = -k^2 Tf$$

and hence also depends only on x. Let

$$w(x,z) = \int_\Omega \epsilon_0(x') G(x,z,x',0) f(x') dx'.$$

By a slight abuse of notation, in what follows we will use Ω to denote $\Omega \times \{0\}$. Using the equation for G, its conditions at infinity, and standard arguments of single layer potential theory [4],

(a) $(\triangle + k^2)w(x,z) = 0$ in $\mathbb{R}^3 \setminus \Omega$,
(b) w satisfies the radiation condition,
(c) $[w]_\Omega = 0$,
(d) $\left[\frac{\partial w}{\partial z}\right]_\Omega = \epsilon_0 f$,

where $[\cdot]$ denotes the jump across Ω, i.e.,

$$[g] = \lim_{z \to 0^+} g(x,z) - \lim_{z \to 0^-} g(x,z).$$

Let B_R be any ball in $\mathbb{R}^3$ containing Ω, multiply through by $\overline{w}$, and integrate by parts to get

$$\int_{B_R \setminus \Omega} |\nabla w|^2 - k^2 \int_{B_R \setminus \Omega} |w|^2 - \int_{\partial B_R} \overline{w} \frac{\partial w}{\partial \nu} - \int_\Omega \overline{w} \left[\frac{\partial w}{\partial z}\right]_\Omega = 0.$$

Since $(I + k^2T)f = 0$, we conclude that $f = -k^2 w$ on Ω. Substituting this and property (d) in the identity above, we get

$$\int_{B_R \setminus \Omega} |\nabla w|^2 - k^2 \int_{B_R \setminus \Omega} |w|^2 - \int_{\partial B_R} \overline{w} \frac{\partial w}{\partial \nu} + \frac{1}{k^2} \int_\Omega \epsilon_0 |f|^2 = 0.$$

Hence

$$\operatorname{Im} \int_{\partial B_R} \overline{w} \frac{\partial w}{\partial \nu} = 0.$$

We can now use the Rellich lemma (see Theorem 3.12 of [1]) on any domain $U \in \mathbb{R}^3$ arbitrarily close to Ω to obtain $w = 0$ in $B_R \setminus U$. Hence $w = 0$ on $\mathbb{R}^3 \setminus \Omega$. Thus the jump $\left[\frac{\partial w}{\partial z}\right]_\Omega = 0$. By property (d), we can conclude that $f = 0$. □

Note an immediate corollary of the above lemma is that u_0 exists, is unique, and is in $C^0(\overline{S})$.

Proof of Proposition 1. Note that for each fixed h, the right-hand side of (4.4) is a continuous function. We take a Taylor expansion of the smooth incident wave $u_{\mathbf{i}}$ for the first term. We then invoke Lemma 1 for the third term to obtain the bound

$$\|(I + k^2T)(\tilde{u} - u_0)\|_{L^\infty(S)} \leq h\|u_{\mathbf{i}}\|_{C^1(S)} + C_1 h \|\tilde{u}\|_{L^\infty(S)} + C_2 h \|\tilde{u}\|_{L^\infty(S)}.$$

By the boundedness of $(I + k^2T)^{-1}$, we have that

$$\begin{aligned} \|\tilde{u} - u_0\|_{L^\infty(S)} &\leq C_3 h \|u_{\mathbf{i}}\|_{C^1(S)} + C_4 h \|\tilde{u}\|_{L^\infty(S)} \\ &\leq C_3 h \|u_{\mathbf{i}}\|_{C^1(S)} + C_4 h \|\tilde{u} - u_0\|_{L^\infty(S)} + C_4 h \|u_0\|_{L^\infty(S)}. \end{aligned}$$

Since we know u_0 is continuous and u_{i} is bounded in C^1, we have that there exists a C_5 independent of h such that

$$(1 - C_4 h)\|\tilde{u} - u_0\|_{L^\infty(S)} \leq C_5 h,$$

from which the result follows for h small enough. □

5. The next term and its justification. First we find formally an equation for u_1, the next term in the expansion. We begin with the ansatz and Taylor expansion

$$\tilde{u}(x,\zeta) = u_0(x) + h u_1(x,\zeta) + O(h^2),$$

$$u_{\mathrm{i}}(x, h\zeta) = u_{\mathrm{i}}(x,0) + h\zeta \frac{\partial u_{\mathrm{i}}}{\partial z}(x,0) + O(h^2),$$

and, in order to obtain some sort of expansion for G, we consider the function v_h,

$$v_h(x,z) = \int_S \epsilon_0(x') G(x,z,x',h\zeta') u_0(x') d\zeta' dx'.$$

From Lemma 1, it seems that as $h \to 0$, $v_h(x,h\zeta)$ should converge $O(h)$ to $v_0(x)$, where

$$v_0(x) = \int_\Omega \epsilon_0(x') G(x,0,x',0) u_0(x') dx'.$$

So it seems reasonable to define

$$v_1(x,\zeta) = \lim_{h\to 0} \frac{v_h(x,h\zeta) - v_0(x)}{h},$$

so that if the limit exists, we have

$$v_h(x,h\zeta) = v_0(x) + h v_1(x,\zeta) + o(h).$$

Insert these expansions into (4.3) and match like powers of h. The $O(1)$ terms give the equation for u_0. The terms of $O(h)$ yield

$$\begin{aligned} u_1(x,\zeta) = \zeta \frac{\partial u_{\mathrm{i}}}{\partial z}(x,0) - k^2 \int_\Omega \int_{-1/2}^{1/2} \epsilon_0(x') G(x,0,x',0) u_1(x,\zeta') d\zeta' dx' \\ + k^2 \int_\Omega \int_{-1/2}^{1/2} G(x,0,x',0) u_0(x') d\zeta' dx' - k^2 v_1(x,\zeta). \end{aligned}$$

We use the definition of the operator T defined in the last section to obtain

$$\begin{aligned} u_1(x,\zeta) = -k^2 T u_1 + \zeta \frac{\partial u_{\mathrm{i}}}{\partial z}(x,0) - k^2 v_1(x,\zeta) \\ + k^2 \int_\Omega G(x,0,x',0) u_0(x') dx'. \end{aligned} \tag{5.1}$$

To complete our definition of u_1 we need to find an expression for $v_1(x,\zeta)$.

LEMMA 3. *Given any fixed $(x,\zeta) \in S$ and any ρ small enough such that the 2-D ball around x of radius ρ, $B_{x,\rho}$, is contained in Ω, then*

$$\lim_{h\to 0} \frac{1}{h} \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{B_{x,\rho}} [G(x,h\zeta,x',h\zeta') - G(x,0,x',0)] dx' d\zeta' = -\frac{1}{2}\left(\zeta^2 + \frac{1}{4}\right).$$

Proof. Before dividing by h, the integral above can be written as

$$\begin{aligned} I &= \frac{2\pi}{4\pi}\int_{-\frac{1}{2}}^{\frac{1}{2}}\int_0^{\rho}\left(\frac{e^{ik\sqrt{r^2+h^2(\zeta-\zeta')^2}}}{\sqrt{r^2+h^2(\zeta-\zeta')^2}}-\frac{e^{ikr}}{r}\right)rdrd\zeta' \\ &= \frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}}\frac{1}{ik}\left(e^{ik\sqrt{\rho^2+h^2(\zeta-\zeta')^2}}-e^{ik\rho}\right)d\zeta'-\frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}}\frac{1}{ik}\left(e^{ik|h(\zeta-\zeta')|}-1\right)d\zeta' \\ &= I_1+I_2. \end{aligned}$$

Since

$$\frac{\partial}{\partial h}\left(e^{ik\sqrt{\rho^2+h^2(\zeta-\zeta')^2}}\right)$$

is integrable, one can compute

$$\lim_{h\to 0^+}\frac{1}{h}I_1=\lim_{h\to 0^+}\frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}}\frac{1}{ik}\frac{\partial}{\partial h}\left(e^{ik\sqrt{\rho^2+h^2(\zeta-\zeta')^2}}\right)d\zeta'=0.$$

The second term,

$$\lim_{h\to 0^+}\frac{1}{h}I_2=-\frac{1}{2}\int_{-\frac{1}{2}}^{\frac{1}{2}}|(\zeta-\zeta')|d\zeta'=-\frac{1}{2}\left(\zeta^2+\frac{1}{4}\right). \qquad \square$$

Note that in the previous lemma, although the limit holds pointwise, it is not uniform as x approaches the boundary of Ω. It is this observation that will lead to a boundary correction which we will examine in a forthcoming paper. For shorthand, in what follows we use the notation

$$G=G(x,h\zeta,x',h\zeta') \quad \text{and} \quad G_0=G(x,0,x',0).$$

PROPOSITION 2. *For any $g\in L^\infty(\Omega)$ such that $g\in C^0(B_{x,\rho})$ for some ρ small enough,*

$$\lim_{h\to 0}\frac{1}{h}\int_S g(x')[G-G_0]dx'd\zeta'=-\frac{1}{2}g(x)\left(\zeta^2+\frac{1}{4}\right).$$

Proof. For any small $\varepsilon>0$, choose ρ small enough so that

$$|g(x')-g(x)|<\varepsilon$$

for any $x',x\in B_{x,\rho}$. Consider

$$\begin{aligned}\int_S g(x')[G-G_0]dx'd\zeta' &= \int_{-\frac{1}{2}}^{\frac{1}{2}}\int_{\Omega\setminus B_{x,\rho}}[G-G_0]g(x')dx'd\zeta' \\ &\quad + \int_{-\frac{1}{2}}^{\frac{1}{2}}\int_{B_{x,\rho}}[G-G_0]g(x')dx'd\zeta' \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}}\int_{\Omega\setminus B_{x,\rho}}[G-G_0]g(x')dx'd\zeta' \\ &\quad + g(x)\int_{-\frac{1}{2}}^{\frac{1}{2}}\int_{B_{x,\rho}}[G-G_0]dx'd\zeta' \\ &\quad + \int_{-\frac{1}{2}}^{\frac{1}{2}}\int_{B_{x,\rho}}[G-G_0][g(x')-g(x)]dx'd\zeta' \\ &= I_1+I_2+I_3.\end{aligned}$$

We first examine I_1. For fixed ρ, we are away from the singularity and can expand the integrand. Define

$$h(y)=\frac{e^{\sqrt{y}}}{\sqrt{y}}.$$

By the mean value theorem, for $y>0$,

$$|h(y+\delta)-h(y)|\le\delta\sup_{[y,y+\delta]}|h'|. \tag{5.2}$$

Note that here,

$$h'(y)=\frac{e^{\sqrt{y}}}{2}\left[\frac{1}{y}-\frac{1}{y^{3/2}}\right].$$

Apply (5.2) with

$$y=|x-x'|^2$$

and

$$\delta=h^2|\zeta-\zeta'|^2.$$

This yields

$$|G-G_0|\le Ch^2\left[\frac{1}{\rho^2}-\frac{1}{\rho^3}\right],$$

where C is independent of h, x', and ζ'. Since $g\in L^\infty$ and ρ is fixed,

$$\lim_{h\to0}\frac{1}{h}I_1=0.$$

By Lemma 3,

$$\lim_{h\to0}\frac{1}{h}I_2=-\frac{1}{2}g(x)\left(\zeta^2+\frac{1}{4}\right).$$

For the last term,

$$\begin{aligned}\frac{1}{h}|I_3| &\le \left\|\frac{G-G_0}{h}\right\|_{L^1(S)} \|g(x')-g(x)\|_{L^\infty(B_{x,\rho})} \\ &\le C\varepsilon\end{aligned}$$

by Lemma 1 and our choice of ρ. From the limits above, we can choose h small enough that

$$\left|\frac{1}{h}\int_S g(x')[G-G_0]dx'd\zeta' + \frac{1}{2}g(x)\left(\zeta^2+\frac{1}{4}\right)\right| \le C\varepsilon+\varepsilon$$

from which the result follows. □

Note that u_0 is continuous from Lemma 2. By setting $g(x) = \epsilon_0(x)u_0(x)$ in Proposition 2, we have that

$$v_1(x) = -\frac{1}{2}\epsilon_0(x)u_0(x)\left(\zeta^2+\frac{1}{4}\right)$$

pointwise almost everywhere in Ω, assuming $\epsilon_0(x)$ is piecewise continuous. Using (5.1), this means that u_1 satisfies

$$\begin{aligned}u_1(x,\zeta) = \zeta\frac{\partial u_{\mathrm{i}}}{\partial z}(x,0) - k^2\int_S \epsilon_0(x')u_1(x',\zeta')G_0 d\zeta' dx' \\ + k^2\int_\Omega u_0(x')G_0dx' + \frac{1}{2}k^2\left(\zeta^2+\frac{1}{4}\right)u_0(x)\epsilon_0(x).\end{aligned} \tag{5.3}$$

To compute a simpler expression for u_1, we first note that from the symmetry of the integral with respect to ζ' we have

$$T\left(\zeta\frac{\partial u_{\mathrm{i}}}{\partial z}(x,0)\right) \equiv 0.$$

Hence $u_1(x,\zeta)$ has the form

$$u_1(x,\zeta) = \hat{u}_1(x) + \zeta\frac{\partial u_{\mathrm{i}}}{\partial z}(x,0) + \frac{1}{2}\zeta^2k^2u_0(x)\epsilon_0(x), \tag{5.4}$$

where $\hat{u}_1(x)$ is the solution to the lower dimensional integral equation

$$\begin{aligned}(I+k^2T)\hat{u}_1(x) = k^2\int_\Omega G(x,0,x',0)u_0(x')dx' \\ -\frac{k^4}{24}\int_\Omega G(x,0,x',0)\epsilon_0^2(x')u_0(x')dx' + \frac{1}{8}k^2u_0(x)\epsilon_0(x).\end{aligned} \tag{5.5}$$

One can verify this by taking $(I+k^2T)$ of both sides of (5.4) and using (5.3) to eliminate u_1. We show the following convergence estimate.

PROPOSITION 3. *Suppose that $\epsilon_0(x)$ is piecewise continuous. Let $\tilde{u}(x,\zeta), u_0(x)$, and $u_1(x,\zeta)$ be given by* (4.3), (4.2), (5.4), *and* (5.5), *respectively. Then as $h \to 0$,*

$$\|\tilde{u} - (u_0 + hu_1)\|_{L^2(S)} = o(h).$$

Proof. Define the error by

$$e = \tilde{u} - (u_0 + hu_1).$$

Then, by using the integral equations for each term we obtain

$$\begin{aligned} e &= u_{\mathrm{i}}(x, h\zeta) + k^2 h \int_S G(x, h\zeta, x', h\zeta')\tilde{u}(x', \zeta')d\zeta' dx' \\ &- k^2 \int_S \epsilon_0(x')G(x, h\zeta, x', h\zeta')\tilde{u}(x', \zeta')d\zeta' dx' - u_{\mathrm{i}}(x, 0) + k^2 T u_0 \\ &- h\zeta \frac{\partial u_{\mathrm{i}}}{\partial z}(x, 0) + hk^2 T u_1 - hk^2 \int_\Omega u_0(x')G(x, 0, x', 0)dx' - \frac{h}{2}k^2 \left(\zeta^2 + \frac{1}{4}\right) u_0(x)\epsilon_0(x). \end{aligned}$$

By adding $k^2 Te$ to both sides and rearranging terms,

$$\begin{aligned} (I + k^2 T)e = u_{\mathrm{i}} &- \left[u_{\mathrm{i}}(x, 0) + h\zeta \frac{\partial u_{\mathrm{i}}}{\partial z}(x, 0)\right] \\ &+ k^2 h \int_S G(x, h\zeta, x', h\zeta')\tilde{u}(x', \zeta')d\zeta' dx' - hk^2 \int_\Omega u_0(x')G(x, 0, x', 0)dx' \\ &+ k^2 \int_S \epsilon_0(x')u_0(x')[G_0 - G]dx' d\zeta' - h\frac{k^2}{2}\left(\zeta^2 + \frac{1}{4}\right) u_0(x)\epsilon_0(x) \\ &+ k^2 \int_S \epsilon_0(x')[G_0 - G](\tilde{u}(x', \zeta') - u_0(x'))d\zeta' dx'. \end{aligned}$$

We will refer to each set of expressions on each line on the right-hand side of the above as $term_1$, $term_2$, $term_3$, and $term_4$. Now, $term_1$ is clearly $O(h^2)$ in L^2 by a Taylor expansion of u_{i}. In $term_2$, we can use Lemma 1 to approximate G by G_0 and commit an error of $O(h)$. Hence $term_2$ becomes

$$hk^2 T\left(\frac{\tilde{u} - u_0}{\epsilon_0}\right) + o(h),$$

which from Proposition 1 and the boundedness of T is $o(h)$ in $L^2(S)$. Consider $term_3/h$. By Proposition 2 and the fact that $g = \epsilon_0 u_0$ is piecewise continuous, this ratio approaches zero pointwise almost everywhere. So,

$$\left\{\left(\frac{term_3}{h}\right)^2\right\}$$

is a sequence of functions converging pointwise almost everywhere to zero, and by Lemma 1, they are uniformly bounded on a bounded domain (and hence in L^1). The Lebesgue dominated convergence theorem therefore yields $term_3/h \to 0$ in $L^2(S)$. Finally, for $term_4$, we can again use Proposition 2 with $g = \epsilon_0(\tilde{u} - u_0)$ to obtain

$$term_4 = h\frac{k^2}{2}\left(\zeta^2 + \frac{1}{4}\right)\epsilon_0(x)(\tilde{u}(x) - u_0(x)) + o(h),$$

which is $o(h)$ in $L^2(S)$ by Proposition 1. We have now shown that

$$\|(I + k^2 T)e\|_{L^2(S)} = o(h).$$

The result follows from Lemma 2. ∎

6. Numerical results. In this section we will show some numerical results. Our goal is to demonstrate the properties of the approximation method, and in order to reduce the computational complexity, we consider 2-D scattering. We will compare results obtained using the approximate method with those obtained by solving the full Lippmann–Schwinger equation numerically.

In two dimensions, we reduce the region Ω to a line segment $\Omega = [-L, L]$. Of course, more general regions consisting of multiple line segments can be considered. The fundamental solution in two dimensions is

$$G(x, \zeta, x', \zeta') = \frac{i}{4} H_0^{(1)} \left(k|(x, \zeta) - (x', \zeta')| \right),$$

where $H_0^{(1)}$ is a Hankel function of the first kind. One can justify that the formula for u_0 is the same as in the 3-D case. The equation satisfied by u_1 hinges on Lemma 3. The result of Lemma 3 applies to the 2-D case without modification. This can be shown by direct calculation. Therefore, the equations for u_0 and u_1 are again (3.5) and (5.4)–(5.5), with the Green's function replaced by the above 2-D version.

In order to obtain solutions to which we compare our approximate solutions, we will solve a 2-D scattering problem. The equation we need to solve is the 2-D version of (3.1). We use piecewise bilinear functions to approximate the exact solution u and discretize the integral equation (3.1) to solve for u. Of particular interest is the solution $u(x, z)$ in the scatterer $S = \Omega \times [-h/2, h/2]$.

We will solve for the approximate solutions $u_0(x)$ using (3.5), and $u_1(x, z)$ using (5.4)–(5.5). To accomplish this, we discretize the integral equations using piecewise linear representations of $u_0(x)$ and $\hat{u}_1$.

The length of the scatterer is $L = 5$, and the thickness is $h = 0.1$. In solving the 2-D problem, we choose a mesh size of 0.02 in the direction of the membrane and 0.025 across its thickness. When solving for u_0 and $\hat{u}_1$, we discretize the interval with mesh size 0.02.

We will solve the scattering problem for three wave numbers, $k = 4$, 8, and 12. We choose the incident wave to be a plane wave of the form

$$u_{\mathrm{i}} = \exp ik(x \cos\theta + z \sin\theta).$$

The wavelengths in the scatterer under these conditions are computed and summarized in Table 6.1. Therefore, one can compare the wavelength with the scatterer thickness $h = 0.1$. For example, at $k = 8$ the thickness is approximately 1/4 the wavelength when $\epsilon = 3$ and approximately 1/3 the wavelength when $\epsilon = 9$.

TABLE 6.1
Wavelength in the scatterer as a function of wave number k and dielectric constant ϵ.

k	$\epsilon = 3$	$\epsilon = 9$
4	0.91	0.52
8	0.45	0.26
12	0.30	0.17

We are particularly interested in the accuracy of the solution on the scatterer itself. In the first experiment, we solve the scattering problem with $k = 8$ with incident wave hitting a uniform scatterer at -45° degrees. For dielectric constant $\epsilon = 3$, the results are shown in Figure 6.1. This is a good situation as the wavelength in the scatterer is more than 4 times the thickness. The error, as can be seen in the figure, is quite small, with the largest disagreement occurring at the bottom of the

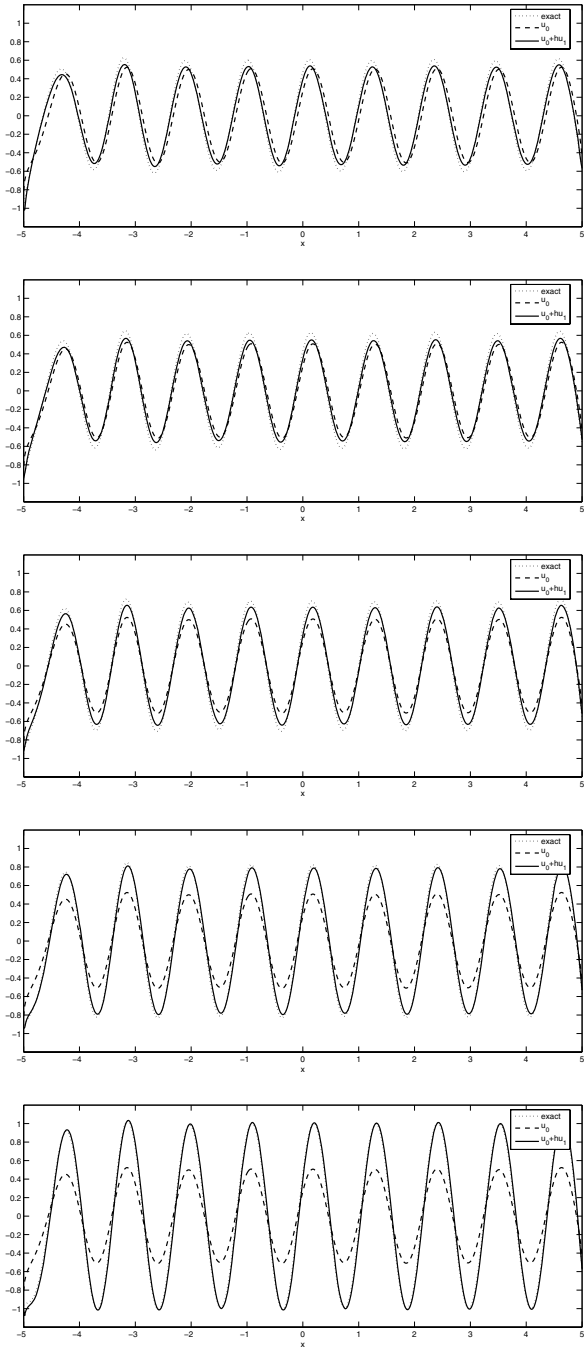


FIG. 6.1. *Top to bottom: real part of* $u(x,z)$ *for* $\epsilon = 3$ *at* $z = -0.05, -0.025, 0, 0.025, 0.05$. *Shown in dots are the exact solutions, and in solid, the approximation* $u_0 + hu_1$. *Also, shown in dashes is the leading order approximation* u_0.

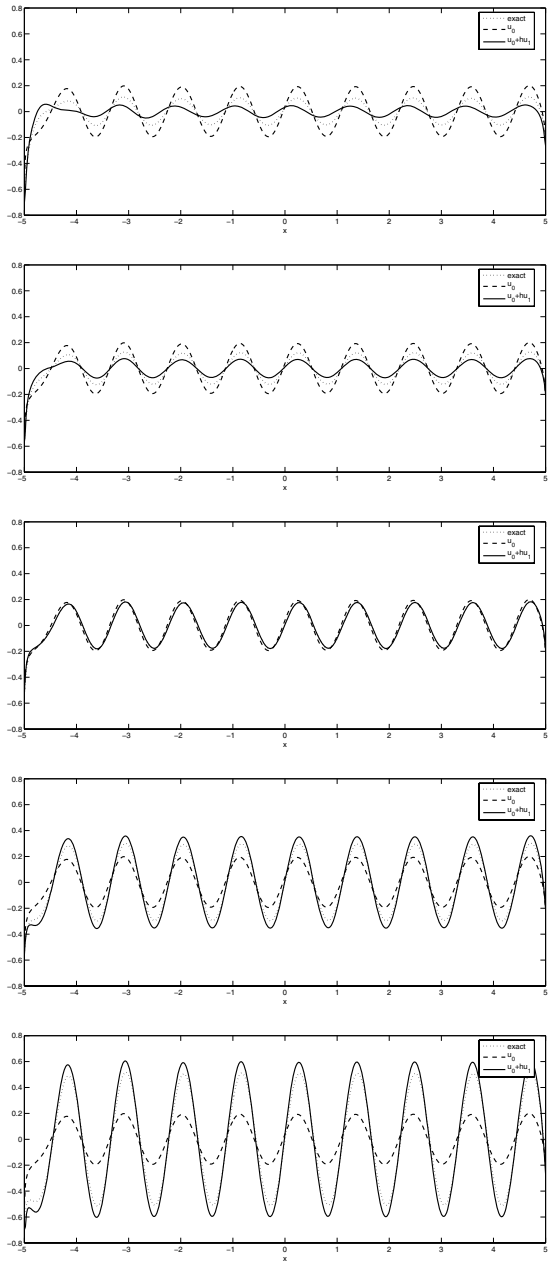


FIG. 6.2. *Top to bottom: real part of* $u(x,z)$ *when* $\epsilon = 9$ *at* $z = -0.05, -0.025, 0, 0.025, 0.05$.

scatterer. There is also some discrepancy at the leading and trailing edges of the scatterer since we have not accounted for the boundary layer.

When the dielectric constant ϵ is 9, the approximation deteriorates. Under this condition, the wavelength is only 2.6 times bigger than the thickness of the scatterer. While the corrector u_1 does improve over the leading order approximation u_0, the error is still quite noticeable. The results are shown in Figure 6.2.

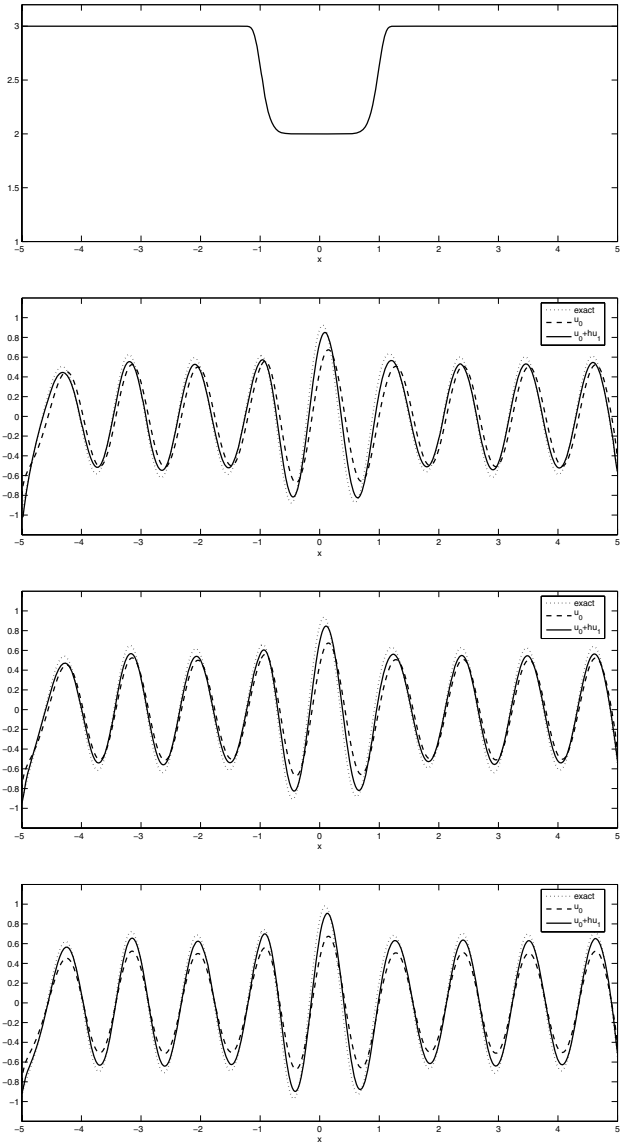


FIG. 6.3. *Top to bottom: the function $\epsilon(x)$, followed by the real part of $u(x,z)$ at $z = -0.05, -0.025, 0$.*

A final example is given when ϵ is x dependent. The distribution of $\epsilon(x)$ is shown in Figure 6.3, together with the solution at various z-values.

Next we calculate the relative L^2 error of the approximation. The L^2 norm is calculated by taking the values of a function at the node points of the regular mesh and interpolating with bilinear splines to obtain an estimate of the function. The interpolated function is then squared and integrated over the domain. The results of the calculations are displayed in Table 6.2. When $\epsilon = 3$, the error increases as a function of k. This is to be expected as the wavelength in the scatterer becomes smaller in comparison to the thickness. When $\epsilon = 9$, the errors follow the same trend

TABLE 6.2
Relative L^2 error of the approximation $u_0 + hu_1$.

	k	Angle of incidence 45°	Normal incidence
$\epsilon = 3$	4	0.0614	0.0529
	8	0.1065	0.1129
	12	0.1413	0.1614
$\epsilon = 9$	4	0.0226	0.0248
	8	0.2058	0.2146
	12	0.6332	0.6567

as k is increased. However, notice that the error is actually smaller for $k = 4$ when $\epsilon = 9$ than when $\epsilon = 3$.

According to our estimates, for a fixed k, as we decrease h and scale ϵ as $1/h$, the error should decrease according to $o(h)$. The numerical examples presented here are meant to give an indication of the accuracy of our approximation.

7. Discussion. In this work, we developed an approximate method for solving a scattering problem where the scatterer is thin. We assume that the dielectric constant of the scatterer scales as $1/h$, where h is the thickness of the scatterer. We formulate the scattering problem using the Lippmann–Schwinger equation. Solution to this equation is approximated by a series in h. Both the leading order solution and the first order corrector can be found by solving an integral equation involving one fewer spatial variable than the original problem. This could lead to substantial savings in realistic computations.

We show that the leading order approximation is $O(h)$ accurate, while the approximation including the first order corrector is $o(h)$ accurate. Boundary layer correctors will be needed to improve the approximation. Finally, we present numerical examples that provide some quantitative assessment of the accuracy of the approximation.

Acknowledgments. The authors are grateful to Habib Ammari, Jay Gopalakrishnan, and Murali Rao for helpful discussions on this work. In particular, we thank Habib Ammari for his clever suggestion which led to the simpler uniqueness proof in Lemma 2.

REFERENCES

[1] D. COLTON AND R. KRESS, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer-Verlag, Berlin, 1998.
[2] G. GOHEN, *Higher-Order Numerical Methods for Transient Wave Equations*, Springer-Verlag, Berlin, 2002.
[3] S. FAN, J. WINN, A. DEVENYI, J. CHEN, R. MEADE, AND J. JOANNOPOULOS, *Guided and defect modes in periodic waveguides*, J. Opt. Soc. Amer. B Opt. Phys., 12 (1995), pp. 1267–1283.
[4] G. FOLLAND, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1976.
[5] P. VILLENEUVE, S. FAN, S. JOHNSON, AND J. JOANNOPOULOS, *Three-dimensional photon confinement in photonic crystals of low-dimensional periodicity*, IEE Proc. Optoelectron., 145 (1998), pp. 384–390.
[6] J. VUCKOVIC, M. LONCAR, H. MABUCHI, AND A. SCHERER, *Optimization of the Q factor in photonic crystal microcavities*, IEEE J. Quantum Elec., 38 (2002), pp. 850–856.
[7] R. WHEEDEN AND A. ZYGMUND, *Measure and Integral: An Introduction to Real Analysis*, Marcel Dekker, New York, 1977.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 206–227

AMERICAN OPTIONS WITH LOOKBACK PAYOFF*

MIN DAI† AND YUE KUEN KWOK‡

Abstract. We examine the early exercise policies and pricing behaviors of one-asset American options with lookback payoff structures. The classes of option models considered include floating strike lookback options, Russian options, fixed strike lookback options, and the pricing model of the dynamic protection fund. For each class of the American lookback options, we analyze the optimal stopping region, in particular the asymptotic behavior at times close to expiration and at infinite time to expiration. The interrelations between the price functions of these American lookback options are explored. The mathematical technique of analyzing the exercise boundary curves of lookback options at infinitesimally small asset values is also applied to the American two-asset minimum put option model.

Key words. lookback options, American feature, free boundary problems, two-asset minimum put option

AMS subject classifications. 90A09, 91B28, 93E20

DOI. 10.1137/S0036139903437345

1. Introduction.

In this paper, we consider the theoretical analysis of the optimal exercise policies of an American option with lookback payoff. An American lookback option involves the combination of two exotic features: the early exercise feature and the lookback feature. Like other American option models, the analysis of an American lookback option requires the solution of a free boundary value problem. The solution procedure involves the determination of the free exercise boundary that separates the stopping region and the continuation region. The analysis is further complicated by the presence of the path-dependent lookback state variable. For floating strike lookback options, the analysis is easier since the dimensionality of the pricing model can be reduced through homogeneity of the price function. This is achieved by taking the asset price as the numeraire. However, for American fixed strike lookback options, the exercise boundary is a two-dimensional curve in the state space described by the asset price and the lookback state variable.

Several earlier papers on American lookback options concentrated on the analysis of the Russian option [7, 17, 18], which is essentially a perpetual zero-strike fixed strike lookback call option. There have been only a few papers which have analyzed the optimal exercise behaviors of finite time American lookback options. Yu, Kwok, and Wu [22] developed finite difference algorithms to compute the exercise boundaries of both American fixed strike and floating strike lookback options. In their two papers [15, 16], Lai and Lim proposed the Bernoulli walk approach to compute the price functions and optimal exercise boundaries of American fixed strike and floating strike lookback options. They also obtained analytic price formulas for American lookback options using a decomposition, which expresses the price as the sum of the corresponding European value and an early exercise premium. Dai, Wong, and

*Received by the editors November 5, 2003; accepted for publication (in revised form) May 16, 2005; published electronically October 17, 2005.
http://www.siam.org/journals/siap/66-1/43734.html
†Department of Mathematics, National Singapore University, Singapore 117543 (matdm@nus.edu.sg).
‡Corresponding author. Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China (maykwok@ust.hk).

Kwok [4] analyzed the exercise policies of American floating strike lookback options with quanto payoff. These quanto options involve an underlying foreign currency asset, but the payoffs are denominated in domestic currency.

We would like to provide a more comprehensive and thorough analysis of the exercise behaviors of the commonly traded American lookback options. Our analysis framework relies more on the partial differential equation approach, as opposed to the usual stochastic approach in most earlier works (say, [14, 15]). For the sake of completeness, we attempt to provide a comprehensive list of analytic properties of the exercise boundaries and stopping regions of the lookback option models. The classes of American lookback option models considered in this paper include the floating strike and fixed strike lookback call and put options, Russian options, and the pricing model of the dynamic protection fund. We analyze the exercise boundary of each class of lookback options, in particular the asymptotic behavior at times close to expiration and at infinite time to expiration. The interrelations between the price functions of these American lookback options are explored. We observe that our mathematical technique developed for analyzing the exercise boundary at infinitesimally small asset values for lookback options can be extended to the American two-asset minimum put option model. For all types of American lookback options considered in this paper, we performed numerical calculations to compute the corresponding exercise boundaries. These plots of exercise boundaries serve as the verification for all results derived from the theoretical studies of the optimal exercise policies.

2. Floating strike lookback options. In this section, we explore some analytic properties of the price functions and optimal exercise policies of the American floating strike lookback options. The usual assumptions of the Black–Scholes option pricing framework are adopted in this paper. Let S denote the price of the underlying asset of the lookback option, whose stochastic dynamics under the risk neutral measure is governed by

$$\frac{dS}{S} = (r - q)dt + \sigma\, dZ, \tag{2.1}$$

where t is the calendar time, r is the riskless interest rate, σ and q are the volatility and dividend yield of S, respectively, and Z is the standard Wiener process. We write τ as the time to expiry, $0 \le \tau < \infty$. Let m and M denote the realized minimum value and realized maximum value, respectively, of the asset price over the lookback monitoring period (continuous monitoring is assumed) up to the current time. The payoff functions of the American floating strike lookback call and lookback put are taken to be

$$(\alpha S - m)^+ \quad \text{and} \quad (M - \alpha S)^+,$$

respectively, where α is a positive parameter value, $0 < \alpha < \infty$, and $x^+ = \max(x, 0)$. When $\alpha = 1$, we recover the usual lookback payoffs. While lookback options are less attractive to investors due to their high option premium, the parameter α allows flexible adjustment of the resulting option premium. For example, we may take α to be less (greater) than one in the floating strike call (put) payoff so as to achieve option premium reduction. Furthermore, the addition of the parameter α in the pricing model facilitates our asymptotic analysis of the exercise boundary curves at the limit of infinitesimally small asset values.

2.1. American floating strike lookback call. Let $C_{f\ell}(S,m,\tau)$ denote the price function of an American floating strike lookback call with payoff $(\alpha S - m)^+$. The linear complementarity formulation that governs $C_{f\ell}(S,m,\tau)$ is given by (see [12] and [21])

$$\begin{aligned} &\frac{\partial C_{f\ell}}{\partial \tau} - \mathcal{L} C_{f\ell} \geq 0, \quad C_{f\ell} \geq \alpha S - m, \\ &\left(\frac{\partial C_{f\ell}}{\partial \tau} - \mathcal{L} C_{f\ell}\right)[C_{f\ell} - (\alpha S - m)] = 0, \quad S > m > 0, \quad \tau > 0, \end{aligned} \tag{2.2}$$

with auxiliary conditions

$$\begin{aligned} &\left.\frac{\partial C_{f\ell}}{\partial m}\right|_{S=m} = 0, \\ &C_{f\ell}(S,m,0) = (\alpha S - m)^+. \end{aligned} \tag{2.3}$$

The operator $\mathcal{L}$ is defined by

$$\mathcal{L} = \frac{\sigma^2}{2} S^2 \frac{\partial^2}{\partial S^2} + (r-q) S \frac{\partial}{\partial S} - r.$$

Note that the payoff upon early exercise is guaranteed to be positive so that we can replace the payoff function $(\alpha S - m)^+$ by $\alpha S - m$. However, we cannot do so for the terminal payoff at $\tau = 0$. The dimension of the above formulation can be reduced by one if we define the following transformation of variables:

$$\eta = \frac{m}{S} \quad \text{and} \quad \widetilde{C_{f\ell}}(\eta,\tau) = \frac{C_{f\ell}(S,m,\tau)}{S}. \tag{2.4}$$

This is equivalent to taking S as the numeraire. The new linear complementarity formulation for $\widetilde{C_{f\ell}}(\eta,\tau)$ is given by

$$\begin{aligned} &\frac{\partial \widetilde{C_{f\ell}}}{\partial \tau} - \widetilde{\mathcal{L}}\, \widetilde{C_{f\ell}} \geq 0, \quad \widetilde{C_{f\ell}} \geq \alpha - \eta, \\ &\left(\frac{\partial \widetilde{C_{f\ell}}}{\partial \tau} - \widetilde{\mathcal{L}}\, \widetilde{C_{f\ell}}\right)[\widetilde{C_{f\ell}} - (\alpha - \eta)] = 0, \quad 0 < \eta < 1, \quad \tau > 0, \end{aligned} \tag{2.5}$$

with auxiliary conditions

$$\begin{aligned} &\left.\frac{\partial \widetilde{C_{f\ell}}}{\partial \eta}\right|_{\eta=1} = 0, \\ &\widetilde{C_{f\ell}}(\eta,0) = (\alpha - \eta)^+, \end{aligned} \tag{2.6}$$

where the operator $\widetilde{\mathcal{L}}$ is given by

$$\widetilde{\mathcal{L}} = \frac{\sigma^2}{2} \eta^2 \frac{\partial^2}{\partial \eta^2} + (q-r)\eta \frac{\partial}{\partial \eta} - q.$$

Remark. The normal reflection condition in (2.6) plays a crucial role in distinguishing the optimal exercise policies of American lookback options from the usual American options. The auxiliary condition is derived from the observation that the

lookback option value is insensitive to the running extremum value when the current asset value equals the extremum value. This is because the probability that the current extremum value remains to be the realized extremum value at maturity is essentially zero when the current asset value and running extremum value are equal (see [10]). In a more recent work, Peskir [17] presented a proof on the normal reflection condition for the finite time Russian option. A similar proof can be mimicked for an American lookback option with a more general lookback payoff.

The holder optimally exercises the lookback call whenever S reaches a sufficiently high level. In terms of η, the holder chooses to exercise when $\eta \leq \eta^*$, where the threshold η^* has dependence on τ. The domain of the pricing model can be divided into two regions: the stopping region $\mathcal{S} = \{(\eta, \tau) : 0 < \eta \leq \eta^*(\tau),\ 0 < \tau < \infty\}$, inside which it is optimal to exercise the option, and the continuation region $\mathcal{S}^C = \{(\eta, \tau) : \eta^*(\tau) < \eta \leq 1,\ 0 \leq \tau < \infty\}$, inside which it is optimal to continue to hold the option. Upon exercise, we have $\widetilde{C}_{f\ell} = \alpha - \eta$ so that the stopping region is defined by

$$\mathcal{S} = \{(\eta, \tau) : 0 < \eta \leq 1,\ 0 \leq \tau < \infty, \text{ and } \widetilde{C}_{f\ell}(\eta, \tau) = \alpha - \eta\}.$$

The analysis of the optimal exercise policies amounts to the analysis of the analytic properties of $\eta^*(\tau)$ that separate the continuation and stopping regions. Some of the analytic properties of $\eta^*(\tau)$ are summarized in Proposition 2.1.

PROPOSITION 2.1. *The exercise boundary $\eta^*(\tau; \alpha)$ of the American floating strike lookback call option observes the following properties:*

(i) *Suppose $(\eta, \tau) \in \mathcal{S}^C$; then $(\lambda_1 \eta, \lambda_2 \tau) \in \mathcal{S}^C$ for all $\lambda_1 \geq 1$, $\lambda_2 \geq 1$.*

(ii) *The line $\eta = 1$ always lies inside $\mathcal{S}^C$ for a finite value of α.*

(iii) *The behavior of $\eta^*(\tau; \alpha)$ near expiry, $\tau \to 0^+$, is given by*

$$\eta^*(0^+; \alpha) = \min\left(1, \alpha, \frac{q}{r}\alpha\right).$$

When $q > 0$, $\eta^*(0^+; \alpha)$ is guaranteed to be positive so that there exists at least a line segment, $\tau = 0$, where $0 < \eta < \eta^*(0^+; \alpha)$, in the stopping region. Property (ii) reveals that the line $\eta = 1$ lies in the continuation region. Hence, we can conclude that both the continuation and stopping regions exist in the η-τ plane. Further, by virtue of (i), the free boundary $\eta^*(\tau; \alpha)$ that separates the stopping and continuation regions can be deduced to be monotonically decreasing with respect to τ. In conclusion, for $q > 0$, there exists the monotonic free boundary $\eta^*(\tau; \alpha)$ such that $\widetilde{C}_{f\ell} = \alpha - \eta$ for $\eta \leq \eta^*(\tau; \alpha)$, $\tau > 0$. The details of the proof of Proposition 2.1 are presented in Appendix A. Further asymptotic properties of $\eta^*(\tau; \alpha)$ with respect to $\tau \to \infty$ and $\alpha \to \infty$ are stated in Proposition 2.2.

PROPOSITION 2.2. *When $q > 0$, the asymptotic behaviors at $\tau \to \infty$ and $\alpha \to \infty$ of the exercise boundary $\eta^*(\tau; \alpha)$ of the American floating strike lookback call option are summarized as follows.*

(i) *Write $\eta^*_\infty(\alpha)$ as $\lim_{\tau \to \infty} \eta^*(\tau; \alpha)$; $\eta^*_\infty(\alpha)$ is given by the solution of the root inside the interval $(0, 1)$ of the following algebraic equation:*

$$(\eta^*_\infty)^{\lambda_+ - \lambda_-} = \frac{\lambda_+}{\lambda_-} \frac{(1 - \lambda_-)\eta^*_\infty + \lambda_- \alpha}{(1 - \lambda_+)\eta^*_\infty + \lambda_+ \alpha},$$

where

$$\lambda_\pm = \frac{r - q}{\sigma^2} + \frac{1}{2} \pm \sqrt{\left(\frac{r - q}{\sigma^2} + \frac{1}{2}\right)^2 + \frac{2q}{\sigma^2}}.$$

(ii) $\lim_{\alpha \to \infty} \eta^*(\tau; \alpha) = 1$ *for all τ.*

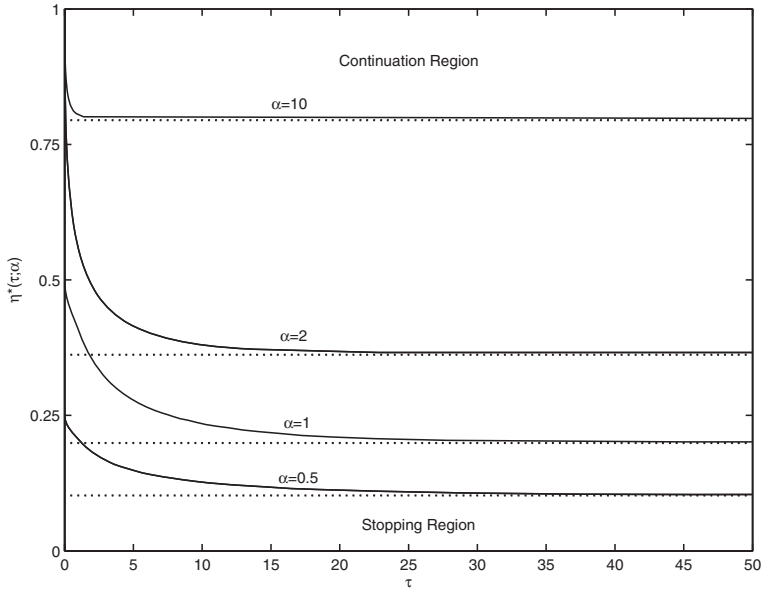


FIG. 1. *The critical threshold $\eta^*(\tau;\alpha)$ of the American floating strike lookback call option is plotted against τ for different values of α. The parameter values of the pricing model are $r = 0.04$, $q = 0.02$, and $\sigma = 0.3$.*

The proof of Proposition 2.2 is presented in Appendix B. From the monotonic decreasing property of $\eta^*(\tau;\alpha)$ with respect to τ and the finiteness property of $\eta^*_\infty(\alpha)$ for $q > 0$, we infer that $\eta^*(\tau;\alpha) > 0$ exists for all τ when $q > 0$. When $\alpha \to \infty$, the continuation region vanishes.

When the underlying asset is non–dividend-paying, $q = 0$, we have $\eta^*(0^+;\alpha) = 0$. Furthermore, since $\eta^*(\tau;\alpha)$ is monotonically decreasing with respect to τ, we deduce that $\eta^*(\tau;\alpha) = 0$ for $\tau > 0$. That is, the stopping region does not exist when $q = 0$. Interpreted in a financial sense, it is never optimal to exercise the American floating strike lookback call at any asset price level if the underlying asset is non–dividend-paying. Such a result agrees intuitively with a similar result of the usual American call.

Figure 1 shows the plot of $\eta^*(\tau;\alpha)$ against τ at varying values of α. The parameter values used in the calculations are $r = 0.04$, $q = 0.02$, and $\sigma = 0.3$. The monotonicity properties of $\eta^*(\tau;\alpha)$ with respect to τ and α and the asymptotic behaviors at $\tau \to 0^+$ and $\tau \to \infty$ as shown in the plots do agree with the results stated in Propositions 2.1 and 2.2. Our calculations give the following asymptotic values for $\eta^*(\tau;\alpha)$:

$$\begin{aligned} &\eta^*(0^+;0.5) = 0.25, \qquad &&\eta^*(\infty;0.5) = 0.1023,\\ &\eta^*(0^+;1) = 0.5, &&\eta^*(\infty;1) = 0.1988,\\ &\eta^*(0^+;2) = 1, &&\eta^*(\infty;2) = 0.3617,\\ &\eta^*(0^+;10) = 1, &&\eta^*(\infty;10) = 0.7947. \end{aligned}$$

2.2. American floating strike lookback put. Let $P_{f\ell}(S, M, \tau)$ denote the price function of an American floating strike lookback put with payoff $(M - \alpha S)^+$. The Russian option is the perpetual version of the American floating strike lookback

put with $\alpha = 0$. In a similar manner, we use S as the numeraire and define

$$\xi = \frac{M}{S} \quad \text{and} \quad \widetilde{P_{f\ell}}(\xi, \tau) = \frac{P_{f\ell}(S, M, \tau)}{S}. \tag{2.7}$$

The linear complementarity formulation for $\widetilde{P_{f\ell}}(\xi, \tau)$ is given by

$$\begin{aligned} &\frac{\partial \widetilde{P_{f\ell}}}{\partial \tau} - \widetilde{\mathcal{L}}\, \widetilde{P_{f\ell}} \geq 0, \quad \widetilde{P_{f\ell}} \geq \xi - \alpha, \\ &\left(\frac{\partial \widetilde{P_{f\ell}}}{\partial \tau} - \widetilde{\mathcal{L}}\, \widetilde{P_{f\ell}} \right) [\widetilde{P_{f\ell}} - (\xi - \alpha)] = 0, \quad 1 < \xi < \infty, \quad \tau > 0, \end{aligned} \tag{2.8}$$

with auxiliary conditions

$$\begin{aligned} &\left. \frac{\partial \widetilde{P_{f\ell}}}{\partial \xi} \right|_{\xi=1} = 0, \\ &\widetilde{P_{f\ell}}(\xi, 0) = (\xi - \alpha)^+. \end{aligned} \tag{2.9}$$

Similarly, we have the free boundary $\xi^*(\tau)$ that separates the stopping region $\{(\xi, \tau) : \xi \geq \xi^*(\tau),\ 0 \leq \tau < \infty\}$ and the continuation region $\{(\xi, \tau) : 1 \leq \xi < \xi^*(\tau),\ 0 \leq \tau < \infty\}$. The analytic properties of $\xi^*(\tau)$ are summarized in Proposition 2.3.

PROPOSITION 2.3. *The free boundary $\xi^*(\tau; \alpha)$ observes the following properties:*

(i) *$\xi^*(\tau; \alpha)$ is monotonically increasing with respect to τ and α.*

(ii) *The behavior of $\xi^*(\tau; \alpha)$ near expiry, $\tau \to 0^+$, is given by*

$$\xi^*(0^+; \alpha) = \max\left(1, \alpha, \frac{q}{r}\alpha\right).$$

(iii) *Write $\xi^*_\infty(\alpha)$ as $\lim_{\tau \to \infty} \xi^*(\tau; \alpha)$; $\xi^*_\infty(\alpha)$ is given by the solution of the root inside the interval $(1, \infty)$ of the following algebraic equation:*

$$(\xi^*_\infty)^{\lambda_+ - \lambda_-} = \frac{\lambda_+}{\lambda_-} \frac{(1 - \lambda_-)\xi^*_\infty + \lambda_- \alpha}{(1 - \lambda_+)\xi^*_\infty + \lambda_+ \alpha}.$$

In particular, when $q = 0$, we have

$$\xi^*_\infty(\alpha) = \infty.$$

As a remark, it is well known that it is never optimal to exercise a Russian option when the underlying asset is non–dividend-paying [18]. The above result shows that such an optimal exercise policy holds even for a nonzero value of α (a Russian option is the special case of $\alpha = 0$).

The ideas behind the proof of Proposition 2.3 are similar to those used in proving Propositions 2.1 and 2.2. In Figure 2, we show the plot of $\xi^*(\tau; \alpha)$ against τ with different values of α. The parameter values used in the calculations are $r = 0.02$, $q = 0.04$, and $\sigma = 0.3$. We obtained the following asymptotic values for $\xi^*(\tau; \alpha)$:

$$\begin{aligned} &\xi^*(0^+; 0) = 1, && \xi^*(\infty; 0) = 3.4939, \\ &\xi^*(0^+; 0.5) = 1, && \xi^*(\infty; 0.5) = 4.8536, \\ &\xi^*(0^+; 1) = 2, && \xi^*(\infty; 1) = 6.6068, \\ &\xi^*(0^+; 2) = 4, && \xi^*(\infty; 2) = 10.7613. \end{aligned}$$

The monotonic behaviors of $\xi^*(\tau; \alpha)$ as exhibited by the plots in Figure 2 are consistent with the results stated in Proposition 2.3.

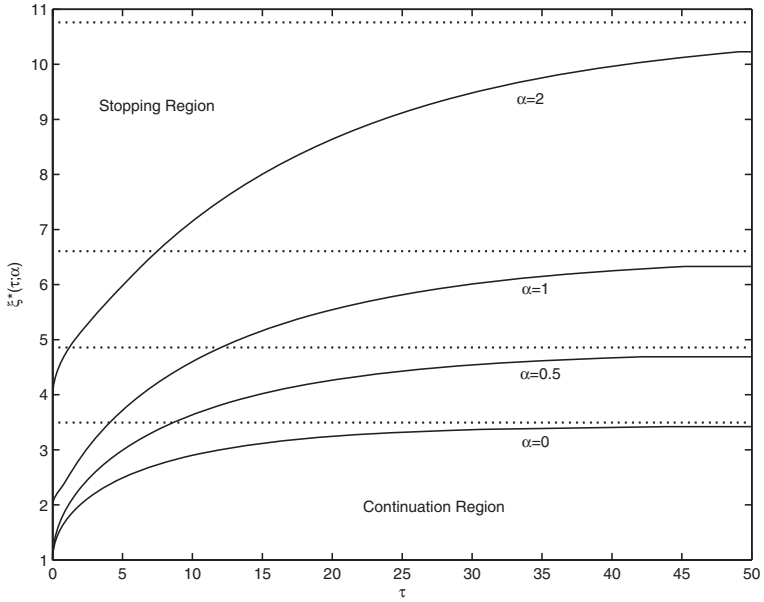


FIG. 2. *The critical threshold* $\xi^*(\tau;\alpha)$ *of the American floating strike lookback put option is plotted against* τ *for different values of* α. *The parameter values of the pricing model are* $r = 0.02$, $q = 0.04$, *and* $\sigma = 0.3$.

3. Fixed strike lookback options. We now consider the pricing behaviors and optimal exercise policies of American fixed strike lookback options, where the payoff involves the strike price K and either realized maximum value M or realized minimum value m. The payoff functions of the American fixed strike lookback call and lookback put are given by

$$(M-K)^+ \quad \text{and} \quad (K-m)^+,$$

respectively. We also consider the American option model with lookback payoff of the form

$$\max(M, K),$$

which is related to the pricing model of the dynamic protection fund with early withdrawal right [8, 11]. According to the guarantee clause, the fund holder acquires more units of the fund from the fund sponsor whenever the fund value falls below the guaranteed protection floor. The early withdrawal right embedded in the protection fund resembles the early exercise right of an American option. When we set $K = 0$ in the payoff $\max(M, K)$, the option model becomes the finite-time Russian option.

It is tempting to seek possible fixed-floating symmetry relations between American lookback call and put options that are similar to those obtained by Detemple [6] for usual American options. While it is possible to obtain symmetry relations between the grant-date price functions of European lookback options (with no dependence on the running extremum value), such relations do not hold for the in-progress counterparts. We do not expect to have nice fixed-floating symmetry relations between the price functions of in-progress American lookback options.

3.1. American fixed strike lookback call. Let $C_{fix}(S, M, \tau; K)$ denote the price function of an American fixed strike lookback call with payoff $(M-K)^+$. The linear complementarity formulation that governs $C_{fix}(S, M, \tau; K)$ is given by

$$\begin{aligned} &\frac{\partial C_{fix}}{\partial \tau} - \mathcal{L}C_{fix} \geq 0, \quad C_{fix} \geq (M-K), \\ &\left(\frac{\partial C_{fix}}{\partial \tau} - \mathcal{L}C_{fix}\right)[C_{fix} - (M-K)] = 0, \quad 0 < S < M, \quad \tau > 0, \end{aligned} \tag{3.1}$$

with auxiliary conditions

$$\begin{aligned} &\left.\frac{\partial C_{fix}}{\partial M}\right|_{S=M} = 0, \\ &C_{fix}(S, M, 0) = (M-K)^+. \end{aligned} \tag{3.2}$$

Let $\mathcal{S}(K)$ denote the stopping region of the American fixed strike lookback call with strike price K. Inside $\mathcal{S}(K)$, the price function equals the exercise payoff; that is,

$$\mathcal{S}(K) = \{(S, M, \tau) \in \{0 < S \leq M\} \times (0, \infty) : C_{fix}(S, M, \tau) = (M-K)^+\}.$$

Propositions 3.1–3.2 summarize the characterization of the optimal exercise policy of the American fixed strike lookback call and the analytic properties of the stopping region.

PROPOSITION 3.1. *The stopping region $\mathcal{S}(K)$ and the price function $C_{fix}(S, M, \tau; K)$ of the American fixed strike lookback call observe the following properties:*

(i) $C_{fix}(S, M, \tau; K_2) - C_{fix}(S, M, \tau; K_1) \leq K_1 - K_2$ *if* $K_1 > K_2$.
(ii) $\mathcal{S}(K_1) \subset \mathcal{S}(K_2)$ *if* $K_1 > K_2$.
(iii) *Suppose* $(S, M, \tau) \in \mathcal{S}(K)$ *and* $0 < \lambda_1 \leq 1$, $\lambda_2 \geq 1$, $0 < \lambda_3 \leq 1$; *we have*

$$(\lambda_1 S, \lambda_2 M, \lambda_3 \tau) \in \mathcal{S}(K).$$

The proof of Proposition 3.1 is presented in Appendix C. In Figure 3, we plot the exercise boundary that separates the stopping region and the continuation region in the S-M plane and use $M^*(S, \tau; K)$ to denote the exercise boundary. Such representation reveals the dependence of the critical realized maximum value M^* on S, τ, and K. By virtue of (iii) in Proposition 3.1, we deduce that the stopping region lies to the upper left-hand side of the exercise boundary in the S-M plane. Hence, we may rewrite $\mathcal{S}(K)$ in the following alternative form:

$$\mathcal{S}(K) = \{(S, M, \tau) \in \{0 < S \leq M\} \times (0, \infty) : \quad M > M^*(S, \tau)\}.$$

Further properties of $M^*(S, \tau; K)$ are summarized in Proposition 3.2.

PROPOSITION 3.2. *Let $M^*(S, \tau; K)$ denote the exercise boundary of the American fixed strike lookback call in the S-M plane; then $M^*(S, \tau; K)$ observes the following properties:*

(i) $\lim_{\tau \to 0^+} M^*(S, \tau; K) = K$ *for all* S.
(ii) $M^*(S, \tau; K)$ *is monotonically increasing with respect to* S *and* τ.
(iii) $\lim_{S \to 0^+} M^*(S, \tau; K) = K$ *for all* τ.
(iv) *When* $K = 0$, $M^*(S, \tau; 0)$ *is a linear function of* S. *Furthermore,* $\frac{M^*(S,\tau;0)}{S}$ *is a monotonically increasing function of* τ *and*

$$\lim_{S \to \infty} \frac{M^*(S, \tau; K)}{S} = \frac{M^*(S, \tau; 0)}{S} \quad \text{for} \quad K > 0. \tag{3.3}$$

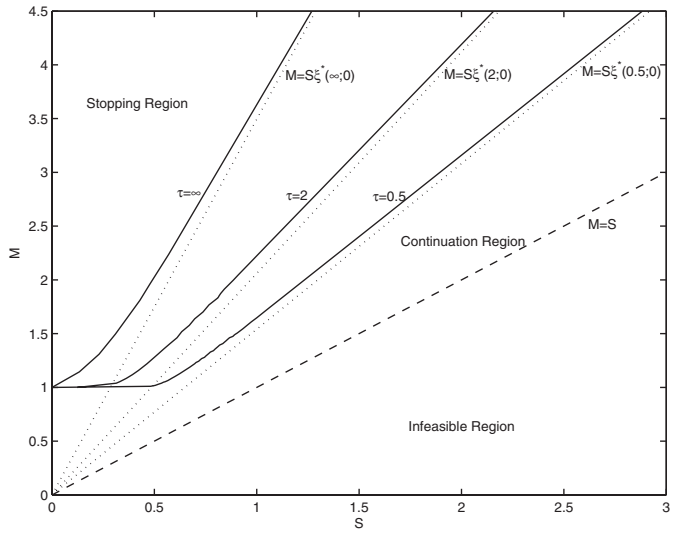


FIG. 3. *The exercise boundaries (solid curves) of the American fixed strike lookback call option with varying values of maturity τ are plotted in the S-M plane. At a given τ, the stopping region is lying to the left of and above the corresponding exercise boundary. The dotted lines are asymptotic lines of the exercise boundaries, corresponding to the exercise boundaries of the zero-strike counterparts. The stopping region of the Russian option lies to the left of the dotted line $M = S\xi^*(\infty; 0)$. The parameter values used in the calculations are $K = 1$, $r = 0.02$, $q = 0.04$, and $\sigma = 0.3$.*

Part (i) gives the zeroth order asymptotic expansion of $M^*(S, \tau; K)$ as $\tau \to 0^+$ (see [15] for a higher order asymptotic expansion of $M^*(S, \tau; K)$ as $\tau \to 0^+$). One can prove part (i) by following an approach similar to that of (iii) in Proposition 2.1. Part (ii) is a corollary of part (iii) in Proposition 3.1. The proofs of parts (iii) and (iv) in Proposition 3.2 are presented in Appendix D.

In Figure 3, we show the plot of the exercise boundaries of the American fixed strike lookback call option with varying values of maturity τ in the S-M plane. The parameter values used in the calculations are $K = 1$, $r = 0.02$, $q = 0.04$, and $\sigma = 0.3$. The exercise boundary corresponding to the zero-strike lookback call is a straight line, the slope of which depends on τ. By virtue of (3.3), the exercise boundaries for the nonzero-strike lookback call options tend to those of their zero-strike counterparts as $S \to \infty$. Note that $M^*(S, \tau; 0)/S = \xi^*(\tau; 0)$, where $\xi^*(\tau; \alpha)$ denotes the exercise boundary in the pricing model for $\widetilde{P_{f\ell}}(\xi, \tau)$ (see (2.8), (2.9)). Our calculations give the following numerical values for $\xi^*(\tau; 0)$:

$$\begin{aligned} \xi^*(\infty; 0) &= 3.4939, \\ \xi^*(2; 0) &= 2.0300, \\ \xi^*(0.5; 0) &= 1.5450. \end{aligned}$$

The finite-time Russian option is seen to be identical to the zero-strike American fixed strike lookback call. Let $V_{Rus}(S, M, \tau)$ denote the price function of the finite-time Russian option so that

$$V_{Rus}(S, M, \tau) = C_{fix}(S, M, \tau; 0). \tag{3.4}$$

Since K does not appear in the price function $V_{Rus}(S, M, \tau)$, the asset value S can

be used as a numeraire. We may write

$$\widetilde{V}_{Rus}(\xi,\tau) = \frac{V_{Rus}(S,M,\tau)}{S}, \quad \text{where} \quad \xi = \frac{M}{S}. \tag{3.5}$$

This explains why $M^*(S,\tau;0)/S$ becomes independent of S. More detailed theoretical analysis of the price function $V_{Rus}(S,M,\tau)$ can be found in Peskir's paper [17].

The exercise boundaries plotted in Figure 3 do agree with our financial intuition about the optimal early exercise policies of the American fixed strike lookback call options. If $S \to 0^+$ or $\tau \to 0^+$, the chance of achieving a higher realized maximum value M becomes vanishingly small, so it becomes optimal to exercise even when M reaches the level K. When the asset price is very high, $M^*(S,\tau;K)$ becomes almost insensitive to the strike price K, since the value K has only a small effect on the exercise payoff. Hence, when $S \to \infty$, the asymptotic behavior of $M^*(S,\tau;K)$ as stated in (3.3) is observed.

3.2. American fixed strike lookback put. Consider an American fixed strike lookback put with payoff $(K-m)^+$; the linear complementarity formulation that governs its price function $P_{fix}(S,m,\tau)$ is given by

$$\begin{aligned} &\frac{\partial P_{fix}}{\partial \tau} - \mathcal{L}P_{fix} \geq 0, \qquad P_{fix} \geq (K-m), \\ &\left(\frac{\partial P_{fix}}{\partial \tau} - \mathcal{L}P_{fix}\right)[P_{fix} - (K-m)] = 0, \qquad 0 < m < S, \quad \tau > 0, \end{aligned} \tag{3.6}$$

with auxiliary conditions

$$\begin{aligned} &\left.\frac{\partial P_{fix}}{\partial m}\right|_{S=m} = 0, \\ &P_{fix}(S,m,0) = (K-m)^+. \end{aligned} \tag{3.7}$$

In a similar manner, we let $m^*(S,\tau;K)$ denote the exercise boundary that separates the stopping region and the continuation region in the S-m plane. The analytic properties of $m^*(S,\tau;K)$ are summarized in Proposition 3.3.

PROPOSITION 3.3. *The exercise boundary $m^*(S,\tau;K)$ of the American fixed strike lookback put satisfies the following properties:*

(i) $\lim_{\tau\to 0^+} m^*(S,\tau;K) = K$ *for all* S.
(ii) $m^*(S,\tau;K)$ *is monotonically increasing with respect to* S.
(iii) $\lim_{S\to\infty} m^*(0,\tau;K) = K$ *for all* τ.
(iv) $\lim_{S\to 0^+} \frac{m^*(S,\tau;K)}{S} = 1$ *for all* τ.

Parts (i)–(iii) in Proposition 3.3 can be proved by using arguments similar to those used in proving parts (i)–(iii) in Proposition 3.2. The proof of (iv) in Proposition 3.3 is interesting and challenging. It relies on the asymptotic result on $\eta^*(\tau;\alpha)$ as stated in (ii) in Proposition 2.2 (see Appendix E for details).

Figure 4 shows the plot of the exercise boundaries $m^*(S,\tau;K)$ of the American fixed strike lookback put with varying values of maturity τ in the S-m plane. The parameter values used in the calculations are $K = 1$, $r = 0.04$, $q = 0.02$, and $\sigma = 0.3$. According to (iii) and (iv) in Proposition 3.3, the exercise boundaries are seen to tend asymptotically to $m = K$ as $S \to \infty$ and to $m = S$ as $S \to 0^+$.

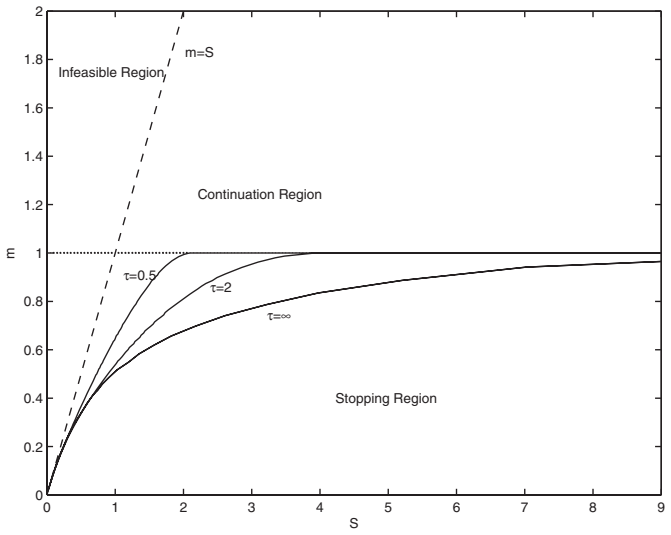


FIG. 4. *The exercise boundaries of the American fixed strike lookback put option with varying values of maturity τ are plotted in the S-m plane. All exercise boundaries tend to the oblique asymptotic line $m = S$ as $S \to 0^+$, and to the horizontal asymptotic line $m = K$ as $S \to \infty$. The parameter values used in the calculations are $K = 1$, $r = 0.04$, $q = 0.02$, and $\sigma = 0.3$.*

3.3. American lookback option with payoff $\max(M, K)$. Let $V_M(S, M, \tau)$ denote the price function of the American option with lookback payoff $\max(M, K)$. First, we argue from financial intuition that $V_M(S, M, \tau)$ should be insensitive to the *current* realized maximum value of asset price M when $M < K$; that is,

$$\frac{\partial V_M}{\partial M} = 0 \quad \text{for} \quad M < K. \tag{3.8}$$

The option payoff is given by K if the *future* realized maximum value of the asset price is less than or equal to K; otherwise, the payoff equals the future realized maximum value. In either case, the current realized maximum value M does not enter into the payoff function. Hence, $V_M(S, M, \tau)$ does have dependence on M when $M < K$. On the other hand, when $M \geq K$, the future realized maximum value is always greater than or equal to K, so the payoff is simply given by M. This is the same payoff as that of the finite-time Russian option. Hence, we have

$$V_M(S, M, \tau) = V_{Rus}(S, M, \tau) \quad \text{for} \quad M \geq K. \tag{3.9}$$

By virtue of the continuity property of the price function $V_M(S, M, \tau)$ with respect to M, we then have

$$V_M(S, M, \tau) = \begin{cases} V_{Rus}(S, M, \tau) & \text{for } M \geq K, \\ V_{Rus}(S, K, \tau) & \text{for } M < K. \end{cases} \tag{3.10}$$

For $M \geq K$, V_M and V_{Rus} should share the same optimal exercise policy. At $M = K$, the exercise boundary of the finite-time Russian option is given by $S = K/\xi^*(\tau; 0)$. Hence, for $M < K$, the American option with payoff $\max(M, K)$ will be exercised optimally when $S \leq K/\xi^*(\tau; 0)$ and unexercised otherwise.

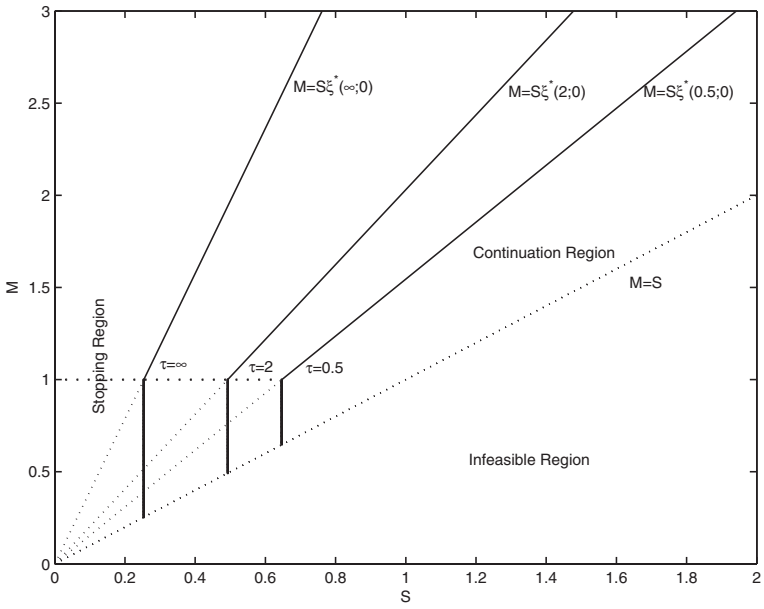


FIG. 5. *The exercise boundaries of the American option with payoff function* $\max(M, K)$ *with varying values of maturity* τ *are plotted in the S-M plane. The parameter values used in the calculations are* $K = 1$, $r = 0.02$, $q = 0.04$, *and* $\sigma = 0.3$.

In Figure 5, we plot the stopping region and the continuation region in the S-M plane of the American option with payoff $\max(M, K)$. The set of parameter values used in the calculations is $K = 1$, $r = 0.02$, $q = 0.04$, and $\sigma = 0.3$. When $M \geq K$, the stopping region and the continuation region for a fixed value of τ are separated by the oblique line $M = S\xi^*(\tau; 0)$. On the other hand, when $M < K$, the exercise boundary becomes the vertical line $S = K/\xi^*(\tau; 0)$.

3.4. A related two-asset American option model. As a slight departure from the option models with lookback payoff structures, we consider the optimal exercise policies of a two-asset American option with a put payoff on the minimum of two asset values. There have been several comprehensive papers that analyze the early exercise policies of two-asset American options [2, 5, 9, 13, 14, 19, 20]. We would like to demonstrate that the mathematical technique of analyzing the exercise boundaries of the American fixed strike lookback put option at $S \to 0^+$ can be adopted to resolve the mystery of the asymptotic behaviors of the exercise boundaries of the two-asset American minimum put option at infinitesimally small asset values.

Let S_1 and S_2 denote the prices of the two underlying assets, whose dynamics under the risk neutral measure is governed by

$$\frac{dS_i}{S_i} = (r - q_i)dt + \sigma_i \, dZ_i, \qquad i = 1, 2, \tag{3.11}$$

where $dZ_1 \, dZ_2 = \rho \, dt$, ρ is the correlation coefficient between the two Wiener processes dZ_1 and dZ_2. The exercise payoff is given by $(K - \min(S_1, S_2))^+$, where K is the strike price. Let $P_{min}(S_1, S_2, \tau; K)$ denote the price function of this two-asset American minimum put option. Let $\mathcal{S}_2(K)$ denote the continuation region in the S_1-S_2 plane, with dependence on K. The linear complementarity formulation for $P_{min}(S_1, S_2, \tau; K)$

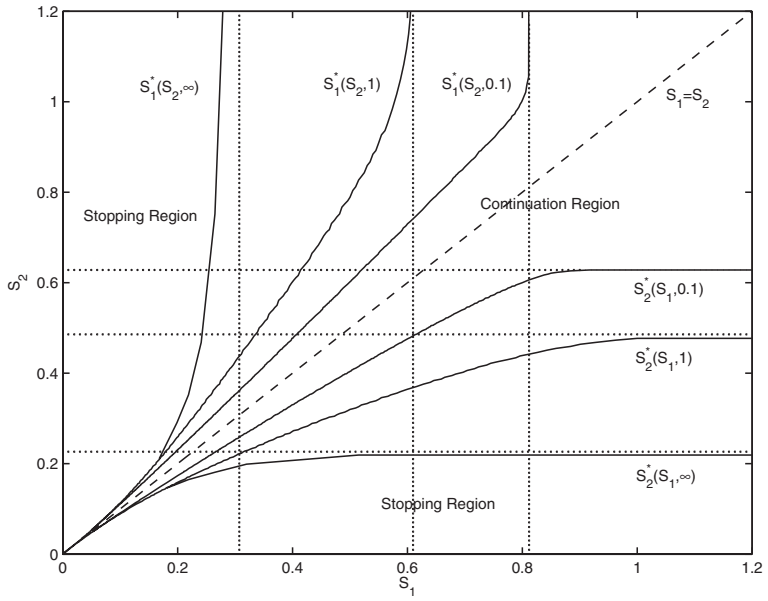


FIG. 6. *The exercise boundaries of the two-asset American minimum put option with varying values of maturity τ are plotted in the S_1-S_2 plane. The continuation region is bounded between the two branches of the exercise boundaries. The parameter values used in the calculations are $K = 1$, $r = 0.02$, $q_1 = 0$, $q_2 = 0.03$, $\sigma_1 = \sigma_2 = 0.3$, and $\rho = 0.5$.*

is given by

$$\begin{aligned} &\frac{\partial P_{min}}{\partial \tau} - \mathcal{L}_2 P_{min} \geq 0, \qquad P_{min} \geq (K - \min(S_1, S_2))^+, \\ &\left[\frac{\partial P_{min}}{\partial \tau} - \mathcal{L}_2 P_{min}\right] [P_{min} - (K - \min(S_1, S_2))^+] = 0, \\ &\qquad\qquad 0 < S_1 < \infty, \quad 0 < S_2 < \infty, \quad \tau > 0. \end{aligned} \tag{3.12}$$

The operator $\mathcal{L}_2$ is defined by

$$\begin{aligned} \mathcal{L}_2 = &\frac{\sigma_1^2}{2} S_1^2 \frac{\partial^2}{\partial S_1^2} + \rho \sigma_1 \sigma_2 S_1 S_2 \frac{\partial^2}{\partial S_1 \partial S_2} + \frac{\sigma_2^2}{2} S_2^2 \frac{\partial^2}{\partial S_2^2} \\ &+ (r - q_1) S_1 \frac{\partial}{\partial S_1} + (r - q_2) S_2 \frac{\partial}{\partial S_2} - r. \end{aligned} \tag{3.13}$$

In Figure 6, we show the plot of the exercise boundaries of the two-asset American minimum put option in the S_1-S_2 plane. The following set of parameter values is used in the calculations: $K = 1$, $r = 0.02$, $q_1 = 0$, $q_2 = 0.03$, $\sigma_1 = \sigma_2 = 0.3$, and $\rho = 0.5$. The whole line $S_1 = S_2$ always lies in the continuation region. The continuation region is bounded by the two branches of the exercise boundaries. In the region $S_1 > S_2$, we let $S_2^*(S_1, \tau)$ denote the exercise boundary at time to expiry τ. We observe that the curve $S_2^*(S_1, \tau)$ tends to the line $S_1 = S_2$ as $S_1 \to 0^+$ and tends to some asymptotic limit as $S_1 \to \infty$. Similar phenomena occur in the region $S_2 > S_1$, where the exercise boundary at time to expiry τ is represented by $S_1^*(S_2, \tau)$. For the

above set of parameter values chosen for the option model, we obtain

$$\lim_{S_1\to\infty} S_2^*(S_1,0.1)=0.6277,\quad \lim_{S_1\to\infty} S_2^*(S_1,1)=0.4855,\quad \lim_{S_1\to\infty} S_2^*(S_1,\infty)=0.2268,$$
$$\lim_{S_2\to\infty} S_1^*(S_2,0.1)=0.8118,\quad \lim_{S_2\to\infty} S_1^*(S_2,1)=0.6100,\quad \lim_{S_2\to\infty} S_1^*(S_2,\infty)=0.3077.$$

Some of the analytic properties of the exercise boundaries $S_1^*(S_2,\tau)$ and $S_2^*(S_1,\tau)$ are summarized in Proposition 3.4.

PROPOSITION 3.4. *Let $S_1^*(S_2,\tau)$ and $S_2^*(S_1,\tau)$ denote the exercise boundaries at time to expiry τ in the two respective regions, $S_2 > S_1$ and $S_1 > S_2$, in the S_1-S_2 plane of the two-asset American minimum put option. The exercise boundaries and the continuation region observe the following properties:*

(i) *Let $S_{1,P}^*(\tau)$ and $S_{2,P}^*(\tau)$ denote the exercise boundary of the one-asset American put option with the underlying asset S_1 and S_2, respectively. We have*

$$\lim_{S_2\to\infty} S_1^*(S_2,\tau)=S_{1,P}^*(\tau)\quad \textit{and}\quad \lim_{S_1\to\infty} S_2^*(S_1,\tau)=S_{2,P}^*(\tau).$$

(ii) *Both $S_1^*(S_2,\tau)$ and $S_2^*(S_1,\tau)$ are monotonically decreasing with respect to time to expiry and monotonically increasing with respect to the asset price level.*

(iii) *The whole line $S_1=S_2$ is contained completely inside the continuation region.*

(iv) *At infinitesimally small asset values, we have*

$$\lim_{S_1\to 0^+}\frac{S_2^*(S_1,\tau)}{S_1}=1\quad \textit{and}\quad \lim_{S_2\to 0^+}\frac{S_1^*(S_2,\tau)}{S_2}=1\quad \textit{for all } \tau. \tag{3.14}$$

All exercise boundaries tend asymptotically to the line $S_1=S_2$ as S_1 and S_2 both tend to zero.

The intuition behind the asymptotic properties stated in part (i) of Proposition 3.4 is quite obvious. When $S_1\to\infty$, $P_{min}(S_1,S_2,\tau;K)\longrightarrow P(S_2,\tau;K)$, where $P(S_2,\tau;K)$ denotes the price function of the one-asset American put option with underlying asset S_2. We would expect that both option models follow the same optimal exercise strategy, thus leading to the asymptotic properties stated in (i). The proof of these asymptotic properties can be pursued by following similar arguments used in the proof of Proposition 4.8 in Villeneuve's paper [20]. Also, the monotonicity properties of $S_1^*(S_2,\tau)$ and $S_2^*(S_1,\tau)$ have been discussed in other papers (say [2] and [20]). Property (iii) states that when $S_1=S_2$, it is never optimal to exercise the two-asset American minimum put option. This optimal exercise policy is similar to that of the two-asset American maximum call option. The proof of (iii) can follow an argument similar to that presented by Detemple, Feng, and Tian [5] on the American maximum call option. The proof of the asymptotic behavior of the exercise boundaries at $S_1\to 0$ and $S_2\to 0$ requires specifically the technique developed in the proof of property (iii) in Proposition 3.3. The proof of part (iv) of Proposition 3.4 is presented in Appendix F.

4. Conclusion. This paper demonstrates the richness of the optimal exercise behaviors adopted by holders of the American options with payoff structures involving lookback state variables. The analysis of the optimal exercise policies of an American lookback option is complicated by the presence of an additional lookback state variable. For fixed strike lookback options, we characterize the exercise behaviors by analyzing the analytic properties of the stopping region and continuation region in

the two-dimensional state space (asset price and lookback state variable). For floating strike lookback options, the dimension of the pricing model can be reduced by one if the asset price is used as the numeraire. We reveal the close relationship between the price functions of the finite-time Russian option and the dynamic protection fund with withdrawal right. For the American put option on the minimum value of two assets, the exercise region consists of two branches of exercise surfaces. Compared to earlier works, our analyses provide a more comprehensive understanding of the optimal exercise policies of commonly traded American lookback options. In particular, we provide a more precise description of the asymptotic behaviors of the exercise boundaries. All the optimal exercise policies of American lookback options derived from our theoretical studies have been verified by plots of the exercise boundaries obtained via numerical calculations.

Appendix A. Proof of Proposition 2.1.

(i) First, we show that if $(\eta, \tau) \in \mathcal{S}^C$, then $(\eta, \lambda_2\tau) \in \mathcal{S}^C$ for $\lambda_2 \geq 1$. By applying the comparison principle, one can show that $\frac{\partial \widetilde{C}_{f\ell}}{\partial \tau} > 0$. This is consistent with the financial intuition that the price function of any American option is an increasing function of τ. Suppose (η, τ) lies in the continuation region; then $\widetilde{C}_{f\ell}(\eta, \tau) > \alpha - \eta$. By virtue of $\frac{\partial \widetilde{C}_{f\ell}}{\partial \tau} > 0$, we deduce that $\widetilde{C}_{f\ell}(\eta, \lambda_2\tau) > \alpha - \eta$ for $\lambda_2 \geq 1$. Hence, $(\eta, \lambda_2\tau)$ also lies in the continuation region.

Next, we show that if $(\eta, \tau) \in \mathcal{S}^C$, then $(\lambda_1\eta, \tau) \in \mathcal{S}^C$ for $\lambda_1 \geq 1$. It suffices to show that

$$\frac{\partial}{\partial \eta}[\widetilde{C}_{f\ell}(\eta, \tau) - (\alpha - \eta)] \geq 0. \tag{A.1}$$

We write $U(\eta, \tau) = \widetilde{C}_{f\ell}(\eta, \tau) - (\alpha - \eta)$; then the linear complementarity formulation for $U(\eta, \tau)$ is given by

$$\frac{\partial U}{\partial \tau} - \widetilde{\mathcal{L}} U \geq r\eta - q\alpha, \quad U \geq 0,$$
$$\left(\frac{\partial U}{\partial \tau} - \widetilde{\mathcal{L}} U\right) U = 0, \quad 0 < \eta < 1, \quad \tau > 0,$$

with auxiliary conditions

$$\left.\frac{\partial U}{\partial \eta}\right|_{\eta=1} = 1 \quad \text{and} \quad U(\eta, 0) = (\eta - \alpha)^+.$$

Both the initial condition $(\eta - \alpha)^+$ and the nonhomogeneous term $r\eta - q\alpha$ are increasing functions of η, and $\frac{\partial U}{\partial \eta}|_{\eta=1} > 0$. By virtue of the comparison principle, we deduce that $\frac{\partial U}{\partial \eta} \geq 0$. □

(ii) We prove by contradiction. Suppose there exists $\tau_0 > 0$ such that $(1, \tau_0) \in S$; by applying (A.1), we can show that $(\eta, \tau_0) \in S$ for $\eta < 1$. We then have

$$\widetilde{C}_{f\ell}(\eta, \tau_0) = \alpha - \eta, \quad \eta < 1.$$

This implies

$$\frac{\partial \widetilde{C}_{f\ell}}{\partial \eta} = -1 \quad \text{at} \quad (1, \tau_0),$$

which contradicts the Neumann boundary condition stated in (2.6). □

(iii) A necessary condition for (η, τ) to lie inside $\mathcal{S}$ is given by

$$\left(\frac{\partial}{\partial \tau} - \widetilde{\mathcal{L}}\right)(\alpha - \eta) = \alpha q - r\eta \geq 0;$$

that is, $\eta \leq \frac{q}{r}\alpha$. Hence, we should have $\eta^*(0^+) \leq \frac{q}{r}\alpha$. Since the exercise payoff must be nonnegative, another necessary condition is given by $\eta \leq \alpha$. Finally, the feasible region for η is $\{\eta : \eta \leq 1\}$. Combining all three necessary conditions, we should have

$$\eta^*(0^+) \leq \min\left(1, \alpha, \frac{q}{r}\alpha\right).$$

Suppose $\eta^*(0^+) < \min(1, \alpha, \frac{q}{r}\alpha)$; then for $\eta \in (\eta^*(0^+), \min(1, \alpha, \frac{q}{r}\alpha))$, we have

$$\left.\frac{\partial \widetilde{C}_{f\ell}}{\partial \tau}\right|_{\tau=0} = \left.\widetilde{\mathcal{L}}\widetilde{C}_{f\ell}\right|_{\tau=0} = \widetilde{\mathcal{L}}(\alpha - \eta) = r\eta - \alpha q < 0.$$

This contradicts $\frac{\partial \widetilde{C}_{f\ell}}{\partial \tau} \geq 0$ for all τ. Hence, we obtain

$$\eta^*(0^+) = \min\left(1, \alpha, \frac{q}{r}\alpha\right). \qquad \square$$

Appendix B. Proof of Proposition 2.2.

(i) Write $\eta^*_\infty(\alpha) = \lim_{\tau\to\infty} \eta^*(\tau; \alpha)$ and $\widetilde{C_{f\ell}}^{\infty}(\eta) = \lim_{\tau\to\infty} \widetilde{C_{f\ell}}(\eta, \tau)$; then $\widetilde{C_{f\ell}}^{\infty}(\eta)$ satisfies the following differential equation:

$$\widetilde{\mathcal{L}}\ \widetilde{C_{f\ell}}^{\infty} = 0, \quad \eta^*_\infty < \eta < 1,$$

subject to the auxiliary conditions

$$\widetilde{C_{f\ell}}^{\infty}(\eta^*_\infty) = \alpha - \eta^*_\infty, \quad \frac{\partial \widetilde{C_{f\ell}}}{\partial \eta}(\eta^*_\infty) = -1, \quad \frac{\partial \widetilde{C_{f\ell}}}{\partial \eta}(1) = 0.$$

The general solution to $\widetilde{C_{f\ell}}^{\infty}(\eta)$ is given by

$$\widetilde{C_{f\ell}}^{\infty}(\eta) = A_1\eta^{\lambda_+} + A_2\eta^{\lambda_-}, \quad \eta^*_\infty < \eta < 1.$$

Applying the auxiliary conditions, we obtain

$$A_1 = \frac{(1-\lambda_-)\eta^*_\infty + \lambda_-\alpha}{(\lambda_- - \lambda_+)(\eta^*_\infty)^{\lambda_+}} \quad \text{and} \quad A_2 = \frac{(1-\lambda_+)\eta^*_\infty + \lambda_+\alpha}{(\lambda_+ - \lambda_-)(\eta^*_\infty)^{\lambda_-}},$$

and η^*_∞ satisfies the nonlinear algebraic equation

$$(\eta^*_\infty)^{\lambda_+ - \lambda_-} = \frac{\lambda_+}{\lambda_-}\frac{(1-\lambda_-)\eta^*_\infty + \lambda_-\alpha}{(1-\lambda_+)\eta^*_\infty + \lambda_+\alpha}. \tag{B.1}$$

The above algebraic equation has two roots; one lies in $(0, 1)$, and the other lies in $(1, \infty)$ (the proof of these properties is found in [3]). Here, η^*_∞ corresponds to the root in $(0, 1)$. Hence, the results in part (i) are established. $\square$

(ii) When $\alpha \to \infty$, the nonlinear algebraic equation (B.1) reduces to

$$(\eta^*_\infty)^{\lambda_+ - \lambda_-} = 1$$

so that the solution for η^*_∞ becomes 1. Also, $\eta^*(0^+) = 1$ when α becomes sufficiently large. Since $\eta^*(\tau)$ is monotonically decreasing with respect to τ, and $\eta^*(0^+) = \eta^*(\infty) = 1$ as $\alpha \to \infty$, we can deduce that

$$\lim_{\alpha \to \infty} \eta^*(\tau; \alpha) = 1 \quad \text{for all } \tau. \qquad \square$$

Appendix C. Proof of Proposition 3.1.

(i) Define the function $V(S, M, \tau; K) = C_{fix}(S, M, \tau; K) + K$. Similar to (3.1)–(3.2), the linear complementarity formulation for $V(S, M, \tau; K)$ is given by

$$\frac{\partial V}{\partial \tau} - \mathcal{L}V \geq rK, \quad V \geq \max(M, K),$$
$$\left[\frac{\partial V}{\partial \tau} - \mathcal{L}V - rK\right][V - \max(M, K)] = 0,$$

with auxiliary conditions

$$\left.\frac{\partial V}{\partial M}\right|_{S=M} = 0 \quad \text{and} \quad V(S, M, 0; K) = \max(M, K).$$

By virtue of the comparison principle, we have

$$V(S, M, \tau; K_1) \geq V(S, M, \tau; K_2) \quad \text{if} \quad K_1 > K_2,$$

and hence the result. $\square$

(ii) From (i), for $K_1 > K_2$ we have

$$C_{fix}(S, M, \tau; K_1) - (M - K_1) \geq C_{fix}(S, M, \tau; K_2) - (M - K_2). \tag{C.1}$$

Suppose $(S, M, \tau) \in \mathcal{S}^C(K_2)$, where $\mathcal{S}^C(K_2)$ denotes the continuation region. In the continuation region, the option value is strictly greater than the exercise payoff so that

$$C_{fix}(S, M, \tau; K_2) > M - K_2.$$

Combining this with inequality (C.1), we can deduce

$$C_{fix}(S, M, \tau; K_1) > M - K_1,$$

so that $(S, M, \tau) \in \mathcal{S}^C(K_1)$. Hence, we establish $\mathcal{S}^C(K_2) \subset \mathcal{S}^C(K_1)$, and thus $\mathcal{S}(K_1) \subset \mathcal{S}(K_2)$. $\square$

(iii) Since $C_{fix}(S, M, \tau)$ is monotonically increasing with respect to both S and τ and the exercise payoff is independent of S and τ, we deduce that if $(S, M, \tau) \in \mathcal{S}(K)$, then

$$(\lambda_1 S, M, \lambda_3 \tau) \in \mathcal{S}(K) \quad \text{for all} \quad 0 < \lambda_1 \leq 1 \text{ and } 0 < \lambda_3 \leq 1.$$

Next, we would like to show that $(S, M, \tau) \in \mathcal{S}(K)$ would imply $(S, \lambda_2 M, \tau) \in \mathcal{S}(K)$ for all $\lambda_2 \geq 1$. Suppose $(S, M, \tau) \in \mathcal{S}(K)$; then $(S/\lambda_2, M, \tau) \in \mathcal{S}(K)$

for $\lambda_2 \geq 1$. Furthermore, by virtue of the linear homogeneity property of the price function and the result in (i), we obtain

$$\begin{aligned} C_{fix}(S, \lambda_2 M, \tau; K) &= \lambda_2 C_{fix}\left(\frac{S}{\lambda_2}, M, \tau; \frac{K}{\lambda_2}\right) \\ &\leq \lambda_2 \left[C_{fix}\left(\frac{S}{\lambda_2}, M, \tau; K\right) + \left(1 - \frac{1}{\lambda_2}\right) K\right] \\ &= \lambda_2 \left[M - K + \left(1 - \frac{1}{\lambda_2}\right) K\right] = \lambda_2 M - K. \end{aligned}$$

On the other hand, the option value $C_{fix}(S, \lambda_2 M, \tau; K)$ cannot fall below the exercise payoff $\lambda_2 M - K$. Combining the results, we then have

$$C_{fix}(S, \lambda_2 M, \tau; K) = \lambda_2 M - K;$$

that is, $(S, \lambda_2 M, \tau) \in \mathcal{S}(K)$. Hence, we obtain the desired result. □

Appendix D. Proof of Proposition 3.2.

(iii) It is clear that $M^*(0^+, \tau; K) \geq K$. From the monotonic increasing property of $M^*(S, \tau; K)$ with respect to S, suppose we can show that the line $M = M_0$ lies in the stopping region in the S-M plane for any $M_0 > K$; then one can deduce that $M^*(S, \tau; K) \to K$ as $S \to 0^+$. This is because the minimum value of $M^*(S, \tau; K)$ is achieved when S is approaching zero from above, and this minimum value is K. We write $U_{fix}(S, \tau) = C_{fix}(S, M_0, \tau) - (M_0 - K)$. The linear complementarity formulation of $U_{fix}(S, \tau)$ is given by

$$\begin{aligned} &\left(\frac{\partial}{\partial \tau} - \mathcal{L}\right) U_{fix} \geq -r(M_0 - K), \quad U_{fix} \geq 0, \\ &\left[\left(\frac{\partial}{\partial \tau} - \mathcal{L}\right) U_{fix}\right] U_{fix} = 0 \end{aligned}$$

with initial condition $U_{fix}(S, 0) = 0$. Since the right-hand term $-r(M_0 - K)$ is always negative and the initial value has compact support, we apply the theorem by Brezis and Friedman [1] that the solution $U_{fix}(S, \tau)$ has compact support, too. The stopping region is nonempty; that is, there exists (S, τ) such that $C_{fix}(S, M_0, \tau) = M_0 - K$ for any $M_0 > K$. Hence, the line $M = M_0 \in \mathcal{S}(K)$ for any $M_0 > K$. □

(iv) When $K = 0$, the American fixed strike lookback call is the same as the American floating strike lookback put (with $\alpha = 0$ in (2.8)). The monotonically increasing property of $\xi^*(\tau) = M^*(S, \tau; 0)/S$ follows directly from Proposition 2.3(i).

For $K > 0$, by virtue of the linear homogeneity property of $M^*(S, \tau; K)$, we obtain

$$\begin{aligned} \lim_{S \to \infty} \frac{M^*(S, \tau; K)}{S} &= \lim_{S \to \infty} \frac{M^*\left(\frac{S}{K}, \tau; 1\right)}{\frac{S}{K}} = \lim_{K \to 0} \frac{M^*\left(\frac{S}{K}, \tau; 1\right)}{\frac{S}{K}} \\ &= \lim_{K \to 0} \frac{M^*(S, \tau; K)}{S} = \frac{M^*(S, \tau; 0)}{S}. \end{aligned}$$

□

Appendix E. Proof of Proposition 3.3.

(iv) First, we consider the proof with $q > 0$, whose arguments rely on the existence of $\eta^*(\tau;\alpha)$. Since $\eta^*(\tau;\alpha)$ does not exist when $q = 0$, we will deal with the special case of zero dividend separately later. For $\alpha \geq 1$, we observe that

$$(K-m)^+ \leq (K-\alpha S)^+ + \alpha S - m$$

so that

$$P_{fix}(S,m,\tau;K) \leq \alpha P\left(S,\tau;\frac{K}{\alpha}\right) + C_{f\ell}(S,m,\tau;\alpha), \tag{E.1}$$

where $P(S,\tau;\frac{K}{\alpha})$ denotes the price function of the American vanilla put option with strike price $\frac{K}{\alpha}$. Let $S_P^*(\tau;\frac{K}{\alpha})$ be the critical asset price of the American vanilla put with payoff $(\frac{K}{\alpha}-S)^+$. Consider the point $(\widehat{S},\widehat{m})$ in the S-m plane which lies inside the region

$$R_\alpha = \left\{(S,m): \quad m \leq S\eta^*(\tau;\alpha) \quad \text{and} \quad S \leq S_P^*\left(\tau;\frac{K}{\alpha}\right)\right\}.$$

$(\widehat{S},\widehat{m})$ lies in the corresponding stopping region of both the American floating strike call and the American vanilla put. We then have

$$P\left(\widehat{S},\tau;\frac{K}{\alpha}\right) = \frac{K}{\alpha} - \widehat{S} \quad \text{and} \quad C_{f\ell}(\widehat{S},\widehat{m},\tau;\alpha) = \alpha\widehat{S} - \widehat{m}. \tag{E.2}$$

Now, we argue that $(\widehat{S},\widehat{m})$ also lies in the stopping region of the American fixed strike put. To establish the claim, it suffices to show that

$$P_{fix}(\widehat{S},\widehat{m},\tau;K) = K - \widehat{m}. \tag{E.3}$$

Combining the results in (E.1) and (E.2), we obtain $P_{fix}(\widehat{S},\widehat{m},\tau;K) \leq K-\widehat{m}$. Since the option value of the American fixed strike put cannot fall below its exercise payoff, the result in (E.3) is then established.

Next, we take the limit $\alpha \to \infty$ and observe that

$$\lim_{\alpha\to\infty} \eta^*(\tau;\alpha) = 1 \quad \text{and} \quad \lim_{\alpha\to\infty} S_P^*\left(\tau;\frac{K}{\alpha}\right) = 0$$

for all τ. As $\alpha \to \infty$, R_α shrinks to an infinitesimally small triangular wedge with the oblique side $S = m$. Hence, we can deduce that as $S \to 0^+$ and for all values of τ, all the exercise boundaries $m^*(S,\tau;K)$ tend to the oblique asymptotic line $S = m$.

Finally, we consider the case where $q = 0$. We add the parameter q in the price function $P_{fix}(S,m,\tau;K,q)$ and exercise boundary $m^*(S,\tau;q)$, and write the corresponding stopping region as $\mathcal{S}(q)$ with dependence on q. From the pricing property

$$P_{fix}(S,m,\tau;K,0) \leq P_{fix}(S,m,\tau;K,q),$$

we deduce that

$$\mathcal{S}(q) \subset \mathcal{S}(0), \quad q > 0.$$

Hence, we have $m^*(S,\tau;0) \geq m^*(S,\tau;q)$ so that

$$\frac{m^*(S,\tau;q)}{S} \leq \frac{m^*(S,\tau;0)}{S} \leq 1, \quad q > 0.$$

Since we have established that $\frac{m^*(S,\tau;q)}{S} \to 1$ as $S \to 0$, it follows that $\lim_{S\to 0^+} \frac{m^*(S,\tau;0)}{S} = 1$. □

Appendix F. Proof of Proposition 3.4.

(iii) We show only the proof of

$$\lim_{S_1\to 0^+} \frac{S_2^*(S_1,\tau)}{S_1} = 1.$$

The proof of the other limiting property in (3.14) can be pursued in a similar manner. Following an approach similar to that in Appendix E, we employ the inequality

$$(K - \min(S_1,S_2))^+ \leq (K-\alpha S_2)^+ + (\alpha S_2 - \min(S_1,S_2))^+ \tag{F.1}$$

and examine the stopping region $\widehat{\mathcal{S}}_\alpha$ of the American two-asset option with payoff $(\alpha S_2 - \min(S_1,S_2))^+$. Also, we let $S_{2,P}^*$ be the critical asset price of the American put with payoff $(\frac{K}{\alpha} - S_2)^+$. By applying inequality (F.1) and following an argument similar to that presented in Appendix E, one can show that the stopping region of the two-asset American minimum put option is contained inside

$$\overline{R}_\alpha = \left\{(S_1,S_2) : (S_1,S_2) \in \widehat{\mathcal{S}}_\alpha \quad \text{and} \quad S_2 \leq S_{2,P}^*\left(\tau; \frac{K}{\alpha}\right)\right\}.$$

The asymptotic behavior of $S_2^*(S_1,\tau)$ at infinitesimally small values of S_1 is established once we can show that the boundaries of $\overline{R}_\alpha$ are bounded by the line $S_1 = S_2$ as $\alpha \to \infty$.

Let V_α denote the price function of the American two-asset option with payoff $(\alpha S_2 - \min(S_1,S_2))^+$, $\alpha \geq 1$. We let $x = S_1/S_2$ and define $W_\alpha = V_\alpha/S_2$. The exercise boundary of the American option model $W_\alpha(x,\tau)$ has two branches; let them be denoted by $x_h^*(\tau)$ and $x_\ell^*(\tau)$. The continuation region is represented by $\{(x,\tau) : x_\ell^*(\tau) < x < x_h^*(\tau),\ 0 \leq \tau < \infty\}$. The linear complementarity formulation of $W_\alpha(x,\tau)$ is given by

$$\frac{\partial W_\alpha}{\partial \tau} - \frac{1}{2}(\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2)x^2\frac{\partial^2 W_\alpha}{\partial x^2} - (q_2 - q_1)x\frac{\partial W_\alpha}{\partial x} + q_2 W_\alpha = 0,$$
$$x_\ell^*(\tau) < x < x_h^*(\tau), \quad \tau > 0,$$

with auxiliary conditions

$$W_\alpha(x_\ell^*,\tau) = \alpha - x_\ell^*, \qquad \frac{\partial W_\alpha}{\partial x}(x_\ell^*,\tau) = -1,$$
$$W_\alpha(x_h^*,\tau) = \alpha - 1, \qquad \frac{\partial W_\alpha}{\partial x}(x_h^*,\tau) = 0,$$
$$W_\alpha(x,0) = \begin{cases} \alpha - x & \text{if } x \leq 1, \\ \alpha - 1 & \text{if } x > 1. \end{cases}$$

For $q_2 > 0$, one can show that $x_\ell^*(\tau)$ and $x_h^*(\tau)$ are monotonic functions of τ. Also, $x_\ell^*(0^+) = x_h^*(0^+) = 1$ when $\alpha > \frac{q_1}{q_2}$. Similarly to property (ii) in Proposition 2.2, we would like to establish the asymptotic results

$$\lim_{\alpha\to\infty} x_\ell^*(\tau;\alpha) = 1 \quad \text{and} \quad \lim_{\alpha\to\infty} x_h^*(\tau;\alpha) = 1 \tag{F.2}$$

so that the boundary of $\overline{R}_\alpha$ will be bounded by $S_1 = S_2$ as $\alpha \to \infty$. By virtue of the monotonicity properties of $x_\ell^*(\tau)$ and $x_h^*(\tau)$ with respect to τ, the asymptotic properties in (F.2) are valid if we can show that

$$\lim_{\alpha\to\infty} x_\ell^*(\infty;\alpha) = 1 \quad \text{and} \quad \lim_{\alpha\to\infty} x_h^*(\infty;\alpha) = 1. \tag{F.3}$$

When $q_2 = 0$, $x_\ell^*(\tau)$ does not exist, but $\lim_{\alpha\to\infty} x_h^*(\tau;\alpha) = 1$ remains valid. The arguments in the proof presented below have to be modified slightly for this degenerate case.

The proof of (F.3) requires the solution of $W_\alpha^\infty(x)$, the perpetual limit of $W_\alpha(x,\tau)$. The governing equation for $W_\alpha^\infty(x)$ is given by

$$\frac{1}{2}(\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2)x^2\frac{d^2W_\alpha^\infty}{dx^2} + (q_2 - q_1)x\frac{dW_\alpha^\infty}{dx} - q_2 W_\alpha^\infty = 0, \qquad x_\ell^*(\infty) < x < x_h^*(\infty),$$

with auxiliary conditions

$$W_\alpha^\infty(x_\ell^*(\infty)) = \alpha - x_\ell^*(\infty), \qquad \frac{dW_\alpha^\infty}{dx}(x_\ell^*(\infty)) = -1,$$
$$W_\alpha^\infty(x_h^*(\infty)) = \alpha - 1, \qquad \frac{dW_\alpha^\infty}{dx}(x_h^*(\infty)) = 0.$$

By following an approach similar to that in Appendix B, we can show that

$$\lim_{\alpha\to\infty} \frac{x_h^*(\infty;\alpha)}{x_\ell^*(\infty;\alpha)} = 1,$$

and hence the relations in (F.3) are established. □

REFERENCES

[1] H. Brezis and A. Friedman, *Estimates on the support of solutions of parabolic variational inequalities*, Illinois J. Math., 20 (1976), pp. 82–97.

[2] M. Broadie and J. Detemple, *The valuation of American options on multiple assets*, Math. Finance, 7 (1997), pp. 241–286.

[3] M. Dai, *A closed-form solution for perpetual American floating strike lookback options*, J. Comput. Finance, 4 (2000), pp. 63–68.

[4] M. Dai, H. Y. Wong, and Y. K. Kwok, *Quanto lookback options*, Math. Finance, 14 (2004), pp. 445–467.

[5] J. Detemple, S. Feng, and W. Tian, *The valuation of American call options on the minimum of two dividend-paying assets*, Ann. Appl. Probab., 13 (2003), pp. 953–983.

[6] J. B. Detemple, *American options: Symmetry properties*, in Option Pricing, Interest Rates and Risk Management, J. Cvitanic, E. Jouini, and M. Musiela, eds., Cambridge University Press, Cambridge, UK, 2001, pp. 67–104.

[7] J. D. Duffie and J. M. Harrison, *Arbitrage pricing of Russian options and perpetual lookback options*, Ann. Appl. Probab., 3 (1993), pp. 641–651.

[8] H. U. Gerber and G. Pafumi, *Pricing dynamic investment fund protection*, N. Am. Actuar. J., 4 (2000), pp. 28–41.

[9] H. Gerber and E. Shiu, *Martingale approach to pricing perpetual American options on two stocks*, Math. Finance, 3 (1996), pp. 87–106.
[10] M. B. Goldman, H. B. Sosin, and M. A. Gatto, *Path dependent options: Buy at the low, sell at the high*, J. Finance, 34 (1979), pp. 1111–1127.
[11] J. Imai and P. Boyle, *Dynamic fund protection*, N. Am. Actuar. J., 5 (2001), pp. 31–51.
[12] P. Jaillet, D. Lamberton, and B. Lapeyre, *Variational inequalities and the pricing of American options*, Acta Appl. Math., 21 (1990), pp. 263–289.
[13] L. S. Jiang, *Analysis of pricing American options on the maximum (minimum) of two risky assets*, Interfaces Free Bound., 4 (2002), pp. 27–46.
[14] J. Kampen, *On American derivatives and related obstacle problems*, Int. J. Theor. Appl. Finance, 6 (2003), pp. 565–591.
[15] T. L. Lai and T. W. Lim, *Exercise regions and effective valuation of American lookback options*, Math. Finance, 14 (2004), pp. 249–269.
[16] T. L. Lai and T. W. Lim, *Efficient Valuation of American Floating-Strike Lookback Options Using a Decomposition Technique*, working paper, Stanford University, Stanford, CA, 2004.
[17] G. Peskir, *The Russian option: Finite horizon*, Finance Stoch., 9 (2005), p. 251–267.
[18] L. A. Shepp and A. N. Shiryaev, *The Russian option: Reduced regret*, Ann. Appl. Probab., 3 (1993), pp. 631–640.
[19] K. Tan and K. Vetzal, *Early exercise regions for exotic options*, J. Derivatives, 3 (1995), pp. 42–56.
[20] S. Villeneuve, *Exercise regions of American options on several assets*, Finance Stoch., 3 (1999), pp. 295–322.
[21] P. Wilmott, J. Dewynne, and J. Howison, *Option Pricing: Mathematical Models and Computation*, Oxford Financial Press, Oxford, UK, 1993.
[22] H. Yu, Y. K. Kwok, and L. Wu, *Early exercise policies of American floating and fixed strike lookback options*, Nonlinear Anal., 47 (2001), pp. 4591–4602.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 228–236

VOLUME OF SUSPENSION THAT FLOWS THROUGH A SMALL ORIFICE BEFORE IT CLOGS*

GUILLERMO H. GOLDSZTEIN†

Abstract. We consider the following experiment. A container is filled with a suspension consisting of particles immersed in an incompressible liquid. An opening is made on the container wall and the suspension flows through the opening. We develop a mathematical model to compute the expected volume of suspension extracted before particles clog the opening. Our studies are relevant to the understanding of clogging of pore throats in porous media, which plays an important role in geomaterials, biological systems, and industrial applications.

Key words. clogging, suspension flow, porous media, mathematical modeling

AMS subject classifications. 76S05, 76M99

DOI. 10.1137/040616164

1. Introduction.

The migration of fines, i.e., small particles, in porous media plays an important role in several engineering applications including oil production, soil erosion, ground water pollution, and the operation of filter beds. Accordingly, this topic is an active area of research in a number of disciplines including petroleum, geotechnical, chemical, environmental, and hydraulic engineering (see [6]).

Soil mass is an example of a porous medium. The particles that hold the material together form what is known as the load carrying skeleton. Fines are small particles that do not form part of the load-carrying skeleton. Rocks are other examples of porous media with fines present in them. A typical size of these fines, which can be of inorganic, organic, or biological nature, is 1 μm, and they may have an electric surface charge. If liquid flows through the porous medium, fines attached to pore surfaces may be released due to hydrodynamic forces. These fines will move with the flow and be retained at other locations or exit the porous medium. The sites that retain fines are usually pore constrictions or pore throats. If several migrating particles reach a small pore throat simultaneously, the particles may clog the pore throat. More detailed discussions on the physical phenomena that lead to clogging can be found in [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14].

If fines get captured, the porous medium may become plugged. On the other hand, when fines exit the medium, the porous medium may erode, which may result in structural failure. Examples where these phenomena have important consequences include the following: the extraction of petroleum, where plugging is an undesirable effect—if the well completely clogs, it can no longer be used; the containment of contaminants—plugging may help in this situation; the failure of earthen dams and roads, which can be caused by the erosion that results from particle migration.

In this paper we study the following simple experiment that models aspects of clogging at a single pore throat. A container is filled with a suspension made of an incompressible liquid and spherical particles. A circular opening is made in the container wall through which the suspension flows. The particles may or may not clog

*Received by the editors October 1, 2004; accepted for publication (in revised form) May 16, 2005; published electronically November 4, 2005. This research was supported by the NSF.
http://www.siam.org/journals/siap/66-1/61616.html
†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (ggold@math.gatech.edu).

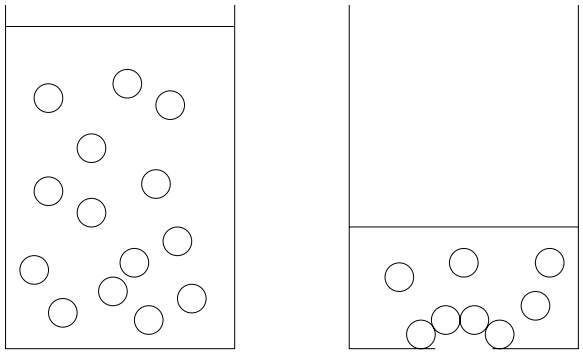


FIG. 1.1. *Suspension in a container. The left-hand image shows the container filled with the suspension before the opening is made. The right-hand image shows the system at the moment the opening clogs.*

the opening. Our goal is to predict the volume of fluid extracted before clogging (if clogging does occur). The experiment described is illustrated in Figure 1.1.

The mathematical modeling of migration of fines in porous media is a complex task that is in its infancy (see [6]). The objective of this paper is to provide a step toward the more ambitious goal of developing reliable models for studying more complex problems where migration of fines in porous media plays an important role.

The rest of the paper proceeds as follows. In section 2 we make our physical assumptions and describe our mathematical model. In section 3 we describe a numerical algorithm to obtain solutions of the model. In section 4 we derive an upper bound on the volume extracted before clogging, and in section 5 we obtain a lower bound. The paper ends in section 6 with examples and conclusions.

2. The model. Our model relies on the following approximations. The liquid is incompressible. The flow is not disturbed by the presence of particles. The center of each particle flows with the same velocity as the fluid. Before the opening is made, the center of each particle is randomly placed inside the container with a uniform probability distribution in space.

Note that the initial location of the centers of the particles are independent random variables, and thus we allow particles to overlap.

For each point x in the container, we denote by $F(x)$ the volume of fluid extracted by the time the element of fluid initially at x reaches the opening. The left-hand image in Figure 2.1 shows a two-dimensional sketch of level sets of the function F. (The actual level sets of F are surfaces within the three-dimensional container.) Due to the incompressibility of the fluid, the region enclosed by the level sets $\{x : F(x) = V+\Delta V\}$ and $\{x : F(x) = V\}$ has volume ΔV.

We denote by A the area of the orifice and by v the volume fraction of particles (i.e., the volume occupied by the particles divided by the volume of the suspension). All the particles have the same radius r.

To motivate our criteria for clogging, assume that the fluid velocity is constant in space across the opening and out of the container. Once the volume of suspension initially in $\{x : V < F(x) \leq V + rA\}$ leaves the container, it forms a cylinder with height r (see the right-hand image in Figure 2.1). Since the centers of particles flow with the fluid, the number of centers of particles that belong to this cylinder is equal to the number of centers of particles initially placed in $\{x : V < F(x) \leq V + rA\}$. We

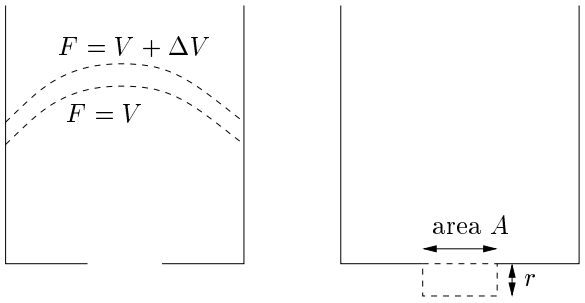


FIG. 2.1. *The left-hand image is a two-dimensional sketch of level sets of F. The region enclosed by the dashed lines is $\{x : V < F(x) \leq V + \Delta V\}$. The right-hand image is a two-dimensional sketch of the three-dimensional cylinder (enclosed by dashed lines) which is formed by the suspension initially in $\{x : V < F(x) \leq V + rA\}$ as soon as it leaves the container.*

denote this number by $k(V)$, i.e.,

$$k(V) = \text{ number of particles initially placed in } \{x : V < F(x) \leq V + rA\}. \tag{2.1}$$

Note that the particles whose centers belong to the dashed cylinder of Figure 2.1 arrive *almost simultaneously* at the opening. Thus, we propose that clogging occurs when $k(V)$, the number of particles arriving almost simultaneously at the opening, exceeds a threshold $k_{\max}$ for the first time. Thus, if the opening clogs, the volume of fluid that is extracted before clogging is

$$V^\star = \min_{\{V : V \geq 0 \text{ and } k(V) > k_{\max}\}} V. \tag{2.2}$$

We define λ to be the ratio of the volume of the dashed cylinder of Figure 2.1 and the volume of a particle, i.e.,

$$\lambda = \frac{3A}{4\pi r^2}. \tag{2.3}$$

Since the number of centers of particles that can belong to the cylinder under the condition that the particles do not overlap increases linearly with λ, we assume that $k_{\max}$ is of the form

$$k_{\max} = \gamma\lambda, \tag{2.4}$$

where γ is a material parameter to be experimentally determined. Given a realization of initial distribution of centers of particles inside the container, (2.1)–(2.4) determine the extracted volume $V^\star$.

3. Algorithm to compute the extracted volume. Assume that the suspension has volume $\mathcal{V}$ and contains a large but finite number N of particles. Then, the volume fraction of particles is $v = N4\pi r^3/(3\mathcal{V})$. The initial location of the center of each particle is a random variable with uniform probability distribution. This fact along with the incompressibility of the fluid implies that the volume extracted by the time a center of a particle reaches the opening is also a random variable with uniform probability distribution. As a consequence, if V_i is the volume extracted when the ith particle reaches the opening, these volumes V_i are the result of ordering N

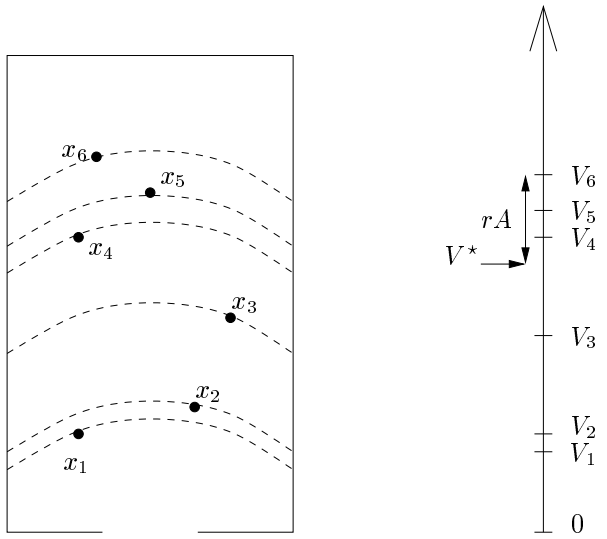


FIG. 3.1. *The initial location of the center of the ith particle to reach the opening is x_i. The dashed lines are the level sets of F. $V_i = F(x_i)$ is the volume extracted when the center of the ith particle reaches the opening. $V^\star$ is the volume extracted before clogging.*

numbers selected independently with uniform probability distribution in the interval $[0, \mathcal{V}]$. This is illustrated in Figure 3.1.

Our criterion for clogging (described in section 2) is illustrated in Figure 3.1. If we place a segment of length rA on top of the vertical volume axis of Figure 3.1 with the left end at 0, then move the segment in the upward direction, and stop as soon as the segment covers more than $k_{\max}$ particles simultaneously, the location of the lower end of the segment is the extracted volume $V^\star$.

The above paragraph can be precisely described as follows. For each i we define n_i to be the largest integer such that $V_{i-n_i+1} > V_i - rA$ subjected to the restriction $n_i \le i$. If there exists $i \in [1, N]$ such that $n_i > k_{\max}$, clogging occurs. Assuming that this is the case, let $i^\star = \min\{i : n_i > k_{\max}\}$. Then

$$V^\star = \begin{cases} 0 & \text{if } V_{i^\star} < rA, \\ V_{i^\star} - rA & \text{if } V_{i^\star} \ge rA. \end{cases} \tag{3.1}$$

The present discussion leads to the following algorithm to compute $V^\star$ for a given realization:

$i \leftarrow 1$
$n \leftarrow 1$
While $n \le k_{\max}$ and $i < N$
    $i \leftarrow i + 1$
    $n \leftarrow n + 1$
    While $V_{i-n+1} \le V_i - rA$
        $n \leftarrow n - 1$
    end
end
If $i = N$ and $n \le k_{\max}$ then
    "No clogging"

else
$V^\star \leftarrow \max\{0, V_i - rA\}$
end

The expected volume extracted before clogging, $E(V^\star)$, is computed by averaging the values of $V^\star$ obtained for a large number of different realizations. Note that the complexity of this algorithm in $O(N)$.

4. Upper bound on the expected extracted volume of suspension before clogging. Given a realization, the extracted volume before clogging $V^\star$, assuming that clogging does occur, is the minimum of the function $f(V) = V$ over the set $\{V : V \geq 0 \text{ and } k(V) > k_{\max}\}$ (see (2.2)). We define U to be the minimum of the same function $f(V) = V$ over a smaller set. More precisely,

$$U = \min_{\{V : V = irA,\ i \text{ integer},\ i \geq 0,\ k(V) > k_{\max}\}} V. \tag{4.1}$$

Since U and $V^\star$ are the minimum of the same function $f(V) = V$, but the set where f is minimized to obtain U is a subset of the set where f is minimized to obtain $V^\star$, we have

$$V^\star \leq U. \tag{4.2}$$

Thus, $E(V^\star)$ and $E(U)$, the expected values of $V^\star$ and U, respectively, satisfy

$$E(V^\star) \leq E(U). \tag{4.3}$$

In Appendix A we show that, if $rA \ll \mathcal{V}$ ($\mathcal{V}$ is the initial volume of the suspension),

$$E(U) = \frac{\mu}{1-\mu} rA, \quad \text{where} \quad \mu = e^{-\lambda v} \sum_{i=0}^{k_{\max}} \frac{(\lambda v)^i}{i!}, \tag{4.4}$$

where we recall that v is the volume fraction of the particles and λ was defined in (2.3). In particular, we also show in Appendix B that, in the parameter regime $\lambda v \ll 1$, we have

$$E(U) \simeq \frac{([k_{\max}] + 1)!}{(\lambda v)^{[k_{\max}]+1}} rA, \tag{4.5}$$

where $[k_{\max}]$ is the integral part of $k_{\max}$, i.e., the largest integer that is not greater than $k_{\max}$.

5. Lower bound on the expected extracted volume of suspension before clogging. Let M be the positive integer that satisfies

$$(M-1)rA \leq V^\star < MrA. \tag{5.1}$$

The sets $\{x : (M-1)rA < F(x) \leq MrA\}$ and $\{x : MrA < F(x) \leq (M+1)rA\}$ are disjoint, and their union contains $\{x : V^\star < F(x) \leq V^\star + rA\}$. Thus, since the number of particle centers initially placed in $\{x : V^\star < F(x) \leq V^\star + rA\}$ is larger than $k_{\max}$, the number of particle centers initially placed in one of the sets $\{x : (M-1)rA < F(x) \leq MrA\}$ or $\{x : MrA < F(x) \leq (M+1)rA\}$ is larger than $k_{\max}/2$. In other words, $\max\{k((M-1)rA), k(MrA)\} > k_{\max}/2$.

We define

$$L = \min_{\{V:V=irA,\ i \text{ integer},\ i\geq 0,\ k(V)>k_{\max}/2\}} V. \tag{5.2}$$

Given the above discussion, we have that $L \leq MrA$. Thus, (5.1) implies

$$L - rA \leq V^\star \tag{5.3}$$

and thus

$$E(L) - rA \leq E(V^\star), \tag{5.4}$$

where, as in the previous section, $E(.)$ denotes the expected value of the expression between brackets.

Following the same arguments to compute the upper bound, we obtain that, in the regime $rA \ll \mathcal{V}$,

$$E(L) = \frac{\eta}{1-\eta} rA, \quad \text{where} \quad \eta = e^{-\lambda v} \sum_{i=0}^{k_{\max}/2} \frac{(\lambda v)^i}{i!}. \tag{5.5}$$

In particular, in the parameter regime $\lambda v \ll 1$, we have

$$E(L) \simeq \frac{([k_{\max}/2]+1)!}{(\lambda v)^{[k_{\max}/2]+1}} rA \tag{5.6}$$

(as before [.] is the integral part of the argument).

6. Examples and conclusions. As an illustrative example, in Figure 6.1 we show a plot of the expected extracted volume $E(V^\star)$ and the upper and lower bounds $E(U)$ and $E(L) - rA$ versus the volume fraction v. The parameter values chosen are $\lambda = 3$ and $\gamma = 1$ (and thus, $k_{\max} = \lambda$). The expected extracted volumes were

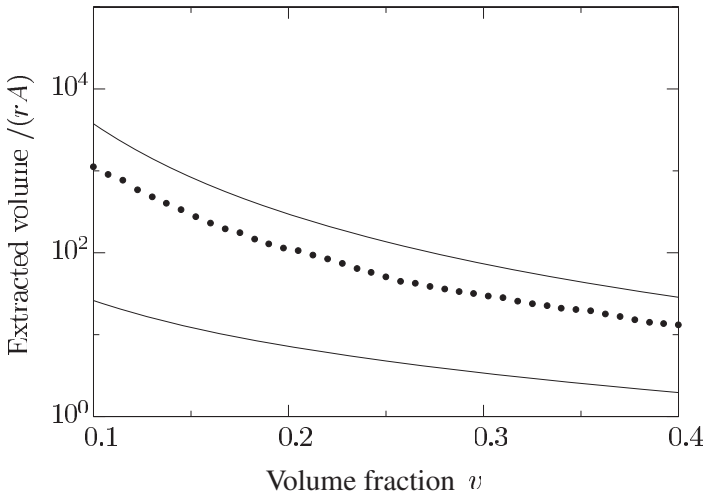


FIG. 6.1. *Normalized expected extracted volume $E(V^\star)/(rA)$ (dotted line), normalized upper bound $E(U)/(rA)$ (upper solid line), and normalized lower bound $(E(L) - rA)/(rA)$ (lower solid line) versus volume fraction v (for $\lambda = 3$ and $\gamma = 1$).*

numerically computed with the method described in this paper. Note that, for a circular opening, $\lambda = 3$ when the radius of the orifice is twice the radius of the particles.

We have developed and analyzed a simple mathematical model to predict the volume of suspension extracted through a small orifice before it clogs. Our model leads to a simple and efficient numerical algorithm as well as analytic expressions for lower and upper bounds on the volume extracted. From the expressions of the bounds, our model reflects the sensitivity of the volume extracted to two key parameters: the volume fraction of particles, v, and λ, which reflects the ratio between the size of the orifice and the size of the particles.

A next step will be to validate the model (or relax some of the physical assumptions made) by comparing the predictions with experimental measurements. After the necessary adjustments, a more ambitious goal is to use the results obtained here as a building block to address more complex problems. These issues will be pursued in the future.

Appendix A. The expected value of the upper bound. To compute the upper bound on the expected volume of suspension extracted before clogging, we need the observations that we describe next. As in the rest of this paper, N is the number of particles initially placed in the container and $\mathcal{V}$ is the initial volume of the suspension.

OBSERVATION 1. *Let Ω be a region inside the container, and let $|\Omega|$ be its volume. The probability that the centers of exactly i of the N particles were initially placed in Ω is*

$$p(i,|\Omega|) = \frac{N!}{i!(N-i)!}\left(\frac{|\Omega|}{\mathcal{V}}\right)^i\left(1-\frac{|\Omega|}{\mathcal{V}}\right)^{N-i}. \tag{A.1}$$

In particular, this probability depends only on i, N, and the volume of Ω. Moreover, the asymptotic value $p(i,|\Omega|)$ in the regime $N \gg i$ and $\mathcal{V} \gg |\Omega|$ is

$$p(i,|\Omega|) \simeq \frac{1}{i!}\left(\frac{N|\Omega|}{\mathcal{V}}\right)^i e^{-\frac{N|\Omega|}{\mathcal{V}}}. \tag{A.2}$$

In particular, if $|\Omega| = rA$,

$$p(i,rA) \simeq \frac{1}{i!}(\lambda v)^i e^{-\lambda v} \tag{A.3}$$

(where v is the volume fraction of the particles and λ was defined in (2.3)).

This observation results from the fact that the centers of the particles are placed in the container randomly with uniform probability distribution, from basic probability arguments (see any probability text book), and from the equality $rAN = \lambda v\mathcal{V}$.

We use the standard notation $P(z)$ for the probability that the event z is true.

OBSERVATION 2. *If i is an integer that satisfies $i \ll N$, and if $rA \ll \mathcal{V}$, then for any $0 \le V \le \mathcal{V} - rA$ we have*

$$P(k(V) = i) \simeq \frac{1}{i!}(\lambda v)^i e^{-\lambda v}. \tag{A.4}$$

Note, in particular, that $P(k(V) = i)$ is independent of V.

This observation results from the definition of the function $k = k(V)$ (see (2.1)), the fact that the volume of the set $\{x : V < F(x) \le V + rA\}$ is rA, and equation (A.3).

OBSERVATION 3. *Let Ω_1 and Ω_2 be two disjoint regions inside the container. Assume that $\mathcal{V} \gg \max\{|\Omega_1|, |\Omega_2|\}$. Let i_1 and i_2 be two nonnegative integers that satisfy $N \gg \max\{i_1, i_2\}$. The probability of having placed exactly i_1 centers of particles in Ω_1 and i_2 centers of particles in Ω_2 is asymptotically equal to $p(i_1, |\Omega_1|)p(i_2, |\Omega_2|)$.*

The validity of this observation is a consequence of the fact that the placements of exactly i_1 centers of particles in Ω_1 and i_2 centers of particles in Ω_2 are asymptotically independent events in the regime $N \gg \max\{i_1, i_2\}$ and $\mathcal{V} \gg \max\{|\Omega_1|, |\Omega_2|\}$.

OBSERVATION 4. *Let i and j be two different nonnegative integers, $i \neq j$. The random variables $k(irA)$ and $k(jrA)$ are asymptotically independent.*

This observation results from the definition of the function $k = k(V)$, the fact that the sets $\{x : jrA < F(x) \leq (j+1)rA\}$ and $\{x : irA < F(x) \leq (i+1)rA\}$ are disjoint, and Observation 3.

Let m be a nonnegative integer. From the definition of U (see (4.1)), we have $U = mrA$ if $k(jrA) \leq k_{\max}$ for $0 \leq j < m$ and $k(mrA) > k_{\max}$. Thus,

$$P(U = mrA) = P(k(jrA) \leq k_{\max} \text{ for } j < m \text{ and } k(mrA) > k_{\max}). \tag{A.5}$$

Given Observation 4, the $m+1$ events $k(jrA) \leq k_{\max}$ (for $0 \leq j < m$) and $k(mrA) > k_{\max}$ are asymptotically independent (more precisely, in the parameter regime $mk_{\max} \ll N$). Thus, (A.5) reduces to

$$P(U = mrA) \simeq P(k(mrA) > k_{\max}) \prod_{j=0}^{m-1} P(k(jrA) \leq k_{\max}). \tag{A.6}$$

From Observation 2 and the definition of the parameter μ in (4.4), we have that

$$P(k(jrA) \leq k_{\max}) \simeq \mu \quad \text{and} \quad P(k(mrA) > k_{\max}) \simeq 1 - \mu. \tag{A.7}$$

Equations (A.6) and (A.7) imply that

$$P(U = mrA) \simeq (1-\mu)\mu^m, \tag{A.8}$$

and thus, in the parameter regime $N \gg k_{\max}$, the expected value of U is

$$E(U) \simeq \sum_{m=0}^{\infty} mrAP(U = mrA) \simeq rA \sum_{m=0}^{\infty} m(1-\mu)\mu^m = \frac{\mu}{1-\mu} rA, \tag{A.9}$$

which shows the validity of (4.4).

Appendix B. The upper bound in the regime $\lambda v \ll 1$. Given the definition of μ (see (4.4)), we have

$$e^{\lambda v}(1-\mu) = e^{\lambda v} - \sum_{i=0}^{k_{\max}} \frac{(\lambda v)^i}{i!} = \sum_{i=0}^{\infty} \frac{(\lambda v)^i}{i!} - \sum_{i=0}^{k_{\max}} \frac{(\lambda v)^i}{i!} = \sum_{i=[k_{\max}]+1}^{\infty} \frac{(\lambda v)^i}{i!}. \tag{B.1}$$

Thus, we have

$$1 - \mu = e^{-\lambda v} \frac{(\lambda v)^{[k_{\max}]+1}}{([k_{\max}]+1)!} \qquad \text{if} \quad \lambda v \ll 1. \tag{B.2}$$

Since we clearly have

$$\mu = e^{-\lambda v} \quad \text{if} \quad \lambda v \ll 1, \tag{B.3}$$

the validity of (4.5) follows.

Acknowledgments. The author thanks Professor Santamarina for introducing the author to this area of research and for stimulating discussions. This work was motivated by experiments of Santamarina's research group.

REFERENCES

[1] Y. Bigno, M. B. Oyeneyin, and J. M. Peden, *Investigation of pore-blocking mechanism in gravel packs in the management and control of fines migration*, in Proceedings of the SPE International Symposium for Damage Control, Society of Petroleum Engineers, 1994, pp. 29–40.

[2] A. Bouhroum and F. Civan, *Study of particulates migration in gravel pack*, in Proceedings of the SPE International Symposium for Damage Control, Society of Petroleum Engineers, 1994, pp. 75–91.

[3] C. Gruesbeck and E. Collins, *Entrainment and deposition of fine particles in porous media*, J. Society of Petroleum Engineers, 22 (1982), pp. 847–856.

[4] C. R. Ison and K. J. Ives, *Removal mechanisms in deep bed filtration*, Chemical Engineering Science, 24 (1969), pp. 717–729.

[5] T. C. Kenney, R. Chahal, E. Chiu, G. I. Ofoegbu, G. N. Omange, and C. A. Ume, *Controlling constriction sizes of granular filters*, Canadian Geotech. J., 22 (1985), pp. 32–43.

[6] K. C. Khilar and H. S. Fogler, *Migration of Fines in Porous Media*, Kluwer Academic Publishers, Boston, MA, 1998.

[7] T. W. Muecke, *Formation fines and factors controlling their movement in porous media*, J. Petroleum Tech., 31 (1979), pp. 144–150.

[8] M. B. Oyeneyin, J. M. Peden, A. Hosseini, and R. Ren, *Factors to consider in the effective management and control of fines migration in high permeability sands*, J. Society of Petroleum Engineers, 35 (1995), pp. 355–360.

[9] V. B. Pandya, S. Bhuniya, and K. C. Khilar, *Existence of a critical particle concentration in plugging of a packed bed*, AIChE J., 44 (1998), pp. 978–981.

[10] R. Sakthivadivel and H. A. Einstein, *Clogging of porous column of spheres by sediment*, ASCE J. Hydraulic Engineering, HY2 (1970), pp. 461–472.

[11] M. M. Sharma and Y. C. Yortsos, *Fines migration in porous media*, Amer. Inst. Chem. Engrg. J., 33 (1987), pp. 1654–1662.

[12] J. L. Sherard, L. P. Dunnigan, and J. R. Talbot, *Basic properties of sand and gravel filters*, J. Geotech. Engineering, 110 (1984), pp. 684–701.

[13] J. L. Sherard, L. P. Dunnigan, and J. R. Talbot, *Filters for silts and clays*, J. Geotech. Engineering, 110 (1984), pp. 701–718.

[14] J. R. Valdes, *Fines Migration and Formation Damage–Microscale Studies*, Ph.D. Dissertation, Department of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, 2002.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 237–265

EXISTENCE AND STABILITY OF TRAVELING WAVES IN BUFFERED SYSTEMS*

JE-CHIANG TSAI† AND JAMES SNEYD‡

Abstract. We study wave propagation in the buffered bistable equation, i.e., the bistable equation where the diffusing species reacts with immobile buffers that restrict its diffusion. Such a model describes wave front propagation in excitable systems where the diffusing species is buffered; in particular, the study of the propagation of waves of increased calcium concentration in a variety of cell types depends directly upon the analysis of such buffered excitability. However, despite the biological importance of these types of equations, there have been almost no analytical studies of their properties.

Here, we study the question of whether or not the inclusion of multiple buffers can eliminate propagated waves. First, we prove that a unique (up to translation) traveling wave front exists. Moreover, the wave speed is also unique. Then we prove that this traveling wave front is stable, i.e., that any initial condition which vaguely resembles a traveling wave front (in a way we make precise) evolves to the unique wave front.

We thus prove that multiple stationary buffers cannot prevent the existence of a traveling wave front in the buffered bistable equation and may not eliminate stable wave fronts. This suggests (although we do not prove) that the same result is true for more complex and realistic models of calcium wave propagation, a result of direct physiological relevance.

Key words. calcium, reaction-diffusion equations, traveling wave, bistable equation, FitzHugh–Nagumo equations, stability

AMS subject classifications. 34A34, 34A12, 35K57

DOI. 10.1137/040618291

1. Introduction.

Wave propagation in excitable systems has been the subject of a vast number of mathematical studies over the last 50 years. The basic mathematical theory can be used to describe wave propagation in a wide array of biological and chemical systems, from action potentials in neurons to chemical waves in the Belousov–Zhabotinskii reaction to combustion waves [4, 9, 24, 26].

One of the more recent applications of the theory has been to the study of waves of increased calcium concentration that travel both within and between cells. Such waves are observed in a wide array of cell types [19, 24]. Although their precise physiological function is not always clear, it is widely accepted that they are an important way in which cells can transmit an intracellular signal or coordinate the behavior of a large number of cells. They have thus been studied extensively, both by experimentalists and theoreticians.

As a general rule, models for calcium waves have been based on reaction-diffusion mechanisms. Thus, if we let u denote the concentration of free cytosolic calcium, a

*Received by the editors November 4, 2004; accepted for publication (in revised form) June 9, 2005; published electronically November 4, 2005. This work was partially supported by the National Science Council of the Republic of China under grant NSC 93-2119-M-019-007 and by the Marsden Fund of the Royal Society of New Zealand.
http://www.siam.org/journals/siap/66-1/61829.html
†Corresponding author. Department of Computer Science, National Taiwan Ocean University, 2, Pei-Ning Road, Keelung 202, Taiwan (tsaijc@mail.ntou.edu.tw).
‡Department of Mathematics, University of Auckland, Private Bag 92019, Auckland, New Zealand (sneyd@math.auckland.ac.nz).

typical model for intracellular calcium waves can be expressed in the following generic form:

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + f(u), \tag{1.1}$$

where $D > 0$ is the diffusion coefficient of the free calcium and $f(u)$ describes the kinetics of calcium transport into and out of the cytosol. For typical examples, the reader is referred to Sneyd, Keizer, and Sanderson [24].

Despite the complexity of all realistic models of calcium wave propagation, simpler models still play a useful role. Although the FitzHugh–Nagumo (FHN) model was originally designed as a simple model of an action potential, the mathematical similarities between action potential propagation and calcium wave propagation mean that one may gain much understanding of the mechanisms underlying calcium waves from a study of these earlier, simple models. Even simpler models, such as the bistable equation, are also useful, despite their almost complete lack of physiological details. Such, of course, is the power of a mathematical model, to abstract the general from the particular, so that one is not restricted to always dealing with special cases.

However, despite the important similarities with other excitable systems, the study of calcium waves has some crucial differences, the most important of which, at first glance, is the existence of calcium buffers. A large fraction of cytosolic calcium (at least 99%) is bound to large proteins that act as calcium buffers. Not only do these buffers restrict the diffusion of free calcium, they also affect the kinetics of calcium release and uptake, and thus they would be expected to have an important effect on the properties of traveling calcium waves. Terms describing buffering do not occur in simpler excitable models such as FHN, nor even in more complex models such as Hodgkin–Huxley. Thus their effect on wave propagation is not at all well understood. Despite this, there have been few analytic investigations of the effects of buffers on traveling waves.

A simple way to model buffering is to assume that calcium (Ca^{2+}) reacts with buffers according to the following reaction scheme:

$$Ca^{2+} + B_i \rightleftharpoons CaB_i, \quad i = 1, \ldots, n, \tag{1.2}$$

where B_i denotes the ith buffer in its unbound form, and CaB_i denotes the ith buffer that is bound to calcium. Let v_i denote the concentration $[B_i]$ of the ith buffer and b_0^i denote the total amount of the ith buffer. Also note that $b_0^i = [B_i] + [CaB_i]$. We shall assume that b_0^i is a constant. It follows that the rate of change of u due to buffering is given by

$$\frac{du}{dt} = \sum_{i=1}^{n} [k_-^i (b_0^i - v_i) - k_+^i u v_i], \tag{1.3}$$

where k_+^i and k_-^i denote the forward and reverse rate constants of the ith reaction (1.2), respectively. Hence if we assume that all of the buffers are stationary, then we have the buffered reaction-diffusion system

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + f(u) + \sum_{i=1}^{n} [k_-^i (b_0^i - v_i) - k_+^i u v_i], \tag{1.4}$$

$$\frac{\partial v_i}{\partial t} = k_-^i (b_0^i - v_i) - k_+^i u v_i, \ (x,t) \in \mathbf{R} \times (0, +\infty), \tag{1.5}$$

with the initial data

$$u(x,0)=\phi(x),\quad v_i(x,0)=\psi_i(x),\ x\in\mathbf{R},\quad \text{for}\ \ i=1,\dots,n. \tag{1.6}$$

Although there are many different types of calcium buffers, each with widely different rate constants, reasonable values for a typical endogenous buffer are $k_+ = 50\ \mu\text{M}^{-1}\text{s}^{-1}$, $k_- = 500\ \text{s}^{-1}$, and $b_0 = 100\ \mu\text{M}$. Typical exogeneous buffers (i.e., ones added experimentally) are BAPTA ($k_+ = 600\ \mu\text{M}^{-1}\text{s}^{-1}$ and $k_- = 100\ \text{s}^{-1}$) or EGTA ($k_+ = 1.5\ \mu\text{M}^{-1}\text{s}^{-1}$ and $k_- = 0.3\ \text{s}^{-1}$). Because BAPTA and EGTA are added exogeneously, their total concentration (b_0) can take any desired value.

There have been a large number of purely numerical studies including calcium buffers (see, for instance, [3, 8, 16, 20]). However, numerical studies suffer from the disadvantage that one can never know how much the results are dependent solely on the particular values chosen for the parameters. It is far preferable to obtain analytical results, applicable to as wide an array of buffers as possible, so that the effects of buffers on calcium dynamics can be understood in the most general possible context.

Some of the early analytical work on buffers was that of Wagner and Keizer [27], who showed that if the buffer is assumed to have fast kinetics (relative to the other reactions in the model), the full buffered model could be reduced to a single equation for the diffusion of calcium, an equation in which the effective diffusion coefficient of calcium was now dependent on the calcium concentration. This is the so-called rapid buffering approximation (RBA). This early work was extended, and put into a more direct experimental context, by Naraghi and Neher [13], Naraghi, Muller, and Neher [14], and Neher [15]. The RBA was also used as the basis of analytical investigations of profiles around the mouth of an open calcium channel [21] and was later derived as one case in a more general asymptotic expansion of the full equation [22].

However, these previous studies left open the question of whether or not multiple buffers, not necessarily fast ones, could eliminate wave propagation. Of course, in the limit of infinitely slow buffers, the answer is intuitively clear; if the wave exists in the absence of buffers, then it will exist in the presence of an infinitely slow buffer. Conversely, Sneyd, Dale, and Duffy [25] showed that if the buffer is infinitely fast, then waves cannot be eliminated by the buffers. However, nothing was known about the intermediate cases when the buffer is neither infinitely fast nor infinitely slow. It is all these intermediate cases that we address here. They are, of course, vastly more difficult. Previous studies also neglected to study the stability of the buffered waves, something we also address here.

As in previous studies [25] we shall study only the buffered bistable equation here, i.e., when the reaction function f takes the particularly simple form,

$$f(u)=u(u-a)(1-u),$$

for some constant $a>0$. A more realistic model of calcium wave propagation would include a recovery variable, as in the FHN equations. However, if recovery is slow, the bistable equation provides a good description of the traveling wave in the FHN equations, and thus our results are also applicable to traveling waves in the FHN equations. This is confirmed numerically in Sneyd, Dale, and Duffy [25]. A similar analysis of a realistic model of calcium waves of much greater complexity is simply not possible at this stage. Nevertheless, as discussed above, previous work suggests that results for the FHN model carry over to more complex calcium wave models.

Description of results. Our results here can be summarized concisely:

1. A unique wave front solution exists (as long as it exists in the absence of buffers). Moreover, the wave speed is also unique.
2. Furthermore, if some technical constraints are made on the initial values, then this unique wave front is also stable with respect to this class of initial values.

Thus, our results complete the picture of how stationary buffers affect wave existence in the bistable equation.

In order to state our results more precisely, we give the definition of a traveling wave solution of (1.4)–(1.5). First, for each solution $(u, \mathbf{v}) = (u, v_1, \dots, v_n)$ of (1.4)–(1.6), we introduce the moving coordinate $z = x - ct$ and set $\tilde{u}(z,t) = u(z+ct,t)$ and $\tilde{v}_i(z,t) = v_i(z+ct,t)$, $i = 1, \dots, n$, where c is a negative constant (wave speed). Then $\tilde{u}$ and $\tilde{v}_i$ satisfy the following system:

$$\frac{\partial \tilde{u}}{\partial t} = D\frac{\partial^2 \tilde{u}}{\partial z^2} + c\frac{\partial \tilde{u}}{\partial z} + f(\tilde{u}) + \sum_{i=1}^{n}[k_-^i(b_0^i - \tilde{v}_i) - k_+^i \tilde{u}\tilde{v}_i], \tag{1.7}$$

$$\frac{\partial \tilde{v}_i}{\partial t} = c\frac{\partial \tilde{v}_i}{\partial z} + k_-^i(b_0^i - \tilde{v}_i) - k_+^i \tilde{u}\tilde{v}_i,\ (z,t) \in \mathbf{R} \times (0, +\infty), \tag{1.8}$$

with the initial data

$$\tilde{u}(z,0) = \phi(z) \ \text{ and } \ \tilde{v}_i(z,0) = \psi_i(z),\ z \in \mathbf{R}, \ \text{ for } \ i = 1, \dots, n. \tag{1.9}$$

Then a set of nonnegative functions $(\mathcal{U}(\xi), \mathbf{\Pi}(\xi)) = (\mathcal{U}(\xi), \Pi_1(\xi), \dots, \Pi_n(\xi)) \in \mathbf{C}^2(\mathbf{R}) \times \mathbf{C}^1(\mathbf{R}) \times \cdots \times \mathbf{C}^1(\mathbf{R})$ are said to be a traveling wave solution of (1.4)–(1.5) with wave speed c if they satisfy

$$D\ddot{\mathcal{U}} + c\dot{\mathcal{U}} + f(\mathcal{U}) + \sum_{i=1}^{n}[k_-^i(b_0^i - \Pi_i) - k_+^i \mathcal{U}\Pi_i] = 0,$$
$$c\dot{\Pi}_i + k_-^i(b_0^i - \Pi_i) - k_+^i \mathcal{U}\Pi_i = 0, \ \ \xi = x - ct \in \mathbf{R},\ i = 1, \dots, n,$$

with the boundary conditions

$$(\mathcal{U}(+\infty), \Pi_1(+\infty), \dots, \Pi_n(+\infty)) = (1, k_-^1 b_0^1/(k_+^1 + k_-^1), \dots, k_-^n b_0^n/(k_+^n + k_-^n))$$

and

$$(\mathcal{U}(-\infty), \Pi_1(-\infty), \dots, \Pi_n(-\infty)) = (0, b_0^1, \dots, b_0^n),$$

where $\cdot$ denotes $d/d\xi$. Note that this definition implies that a traveling wave solution $(\mathcal{U}(\xi), \mathbf{\Pi}(\xi))$ of (1.4)–(1.5) with wave speed c is a steady state solution of (1.7)–(1.8). Also note that, physiologically, the concentration of the ith buffer v_i shall decrease from b_0^i to some constant concentration as time evolves for $i = 1, \dots, n$. Thus we impose such boundary conditions.

It is well known that traveling waves with negative speed in the bistable equation exist only when $a \in (0, 1/2)$ or, more generally, when $\int_0^1 f(u)du > 0$ (see Fife and McLeod [5] and Britton [2]). Moreover, the wave speed is unique. On the other hand, it is interesting to point out that we can prove that given $a \in (0, 1/2)$, then there exists a unique $c := c(a) < 0$ such that a unique (up to translation) traveling wave solution $(\mathcal{U}, \mathbf{\Pi})$ of our buffered bistable equations (1.4)–(1.5) with wave speed c exists. Moreover, $(\mathcal{U}, \mathbf{\Pi})$ satisfies that $\dot{\mathcal{U}} > 0$ and $\dot{\Pi}_i < 0$, $i = 1, \dots, n$, on $\mathbf{R}$. Furthermore,

a traveling wave solution $(\mathcal{U}, \mathbf{\Pi})$ of (1.4)–(1.5) with negative wave speed, such that $\dot{\mathcal{U}} > 0$ and $\dot{\Pi}_i < 0$, $i = 1, \ldots, n$, on $\mathbf{R}$, exists only if $a \in (0, 1/2)$. Comparing this result with the unbuffered equation (1.1), we may conclude that stationary buffers cannot eliminate wave activity. Throughout the remainder of this section, $(\mathcal{U}, \mathbf{\Pi})$ will denote such a unique traveling wave solution of (1.4)–(1.5) with speed c and $\mathcal{U}(0) = 1/2$.

Regarding the stability of this traveling wave solution of (1.4)–(1.5), we need to put some technical constraint on the initial values. More precisely, let $a \in (0, 1/2)$ hold and $(\tilde{u}, \tilde{\mathbf{v}})$ be the solution of (1.7)–(1.8) with the initial condition $(\phi, \psi_1, \ldots, \psi_n)$, which satisfies the following conditions:

(1) ϕ and ψ_i are sufficiently smooth, and $\phi' \geq 0$, $\psi_i' \leq 0$ on $\mathbf{R}$ for $i = 1, \ldots, n$;
(2) $\sup_{z\in\mathbf{R}} |\phi(z)| + \sup_{z\in\mathbf{R}} |\phi'(z)| + \sup_{z\in\mathbf{R}} |\phi''(z)| + \sup_{z\in\mathbf{R}} |\phi'''(z)| < +\infty$;
(3) $1 - \phi_2 > 0$, $\psi_{2i} - k_-^i b_0^i/(k_+^i + k_-^i) > 0$, $\phi_0 > 0$, and $b_0^i - \psi_{0i} > 0$ are sufficiently small for $i = 1, \ldots, n$;
(4) $\tilde{u}_t(z, 0) \geq 0$ and $\tilde{v}_{i,t}(z, 0) \leq 0$ for all $z \in \mathbf{R}$ and $i = 1, \ldots, n$,

where

(1.10)
$$\phi_2 = \lim_{z\to+\infty} \phi(z),\ \psi_{2i} = \lim_{z\to+\infty} \psi_i(z),\ \phi_0 = \lim_{z\to-\infty} \phi(z),\ \text{ and }\ \psi_{0i} = \lim_{z\to-\infty} \psi_i(z);$$

then there exists $z_0 \in \mathbf{R}$ such that

$$\lim_{t\to+\infty} |\tilde{u}(z, t) - \mathcal{U}(z - z_0)| = 0,\ \ \lim_{t\to+\infty} |\tilde{v}_i(z, t) - \Pi_i(z - z_0)| = 0,\ i = 1, \ldots, n,$$

uniformly with respect to $z \in \mathbf{R}$. Roughly speaking, this implies that a solution of (1.4)–(1.5) which vaguely resembles a traveling front $(\mathcal{U}, \Pi_1, \ldots, \Pi_n)$ at initial time will develop into a translate of such a traveling front as $t \to +\infty$. Therefore, we may conclude that, physiologically, a unique stable traveling wave front exists, as long as it exists in the absence of buffers. Finally, we will consider the case for mobile buffers in the future.

We will modify the method of Klaasen and Troy [10] (also see Hastings [7] and Fife and McLeod [5]) to prove our results. However, note that the assumptions made on the reaction terms in [10] are different from the one here. This paper is organized as follows. In section 2, we first give some preliminary results. In section 3, we will adapt the method of [10, 7] to prove the existence and uniqueness of a traveling wave solution of (1.4)–(1.5), and the uniqueness of wave speed is also considered. In section 4, some theorems on partial differential equations and the results of stability for our traveling wave solution will be stated, and then we will modify the method of [10, 5] to prove the theorem of stability for our traveling wave solution of (1.4)–(1.5). Finally, the proofs of some technical lemmas are deferred to the appendix.

2. Preliminaries. First, for ease of use we set up some notation.

DEFINITION 1.

(1) $\gamma_i(u) := k_+^i u + k_-^i$ *and* $k_i(u) := k_-^i b_0^i/(k_+^i u + k_-^i)$ *for* $i = 1, \ldots, n$.
(2) $F(u, \mathbf{v}) := u(u-a)(1-u) + \sum_{i=1}^n [k_-^i b_0^i - (k_+^i u + k_-^i) v_i]$, *where* $\mathbf{v} = (v_1, \ldots, v_n)$.
(3) $G_i(u, \mathbf{v}) := k_-^i b_0^i - (k_+^i u + k_-^i) v_i = \gamma_i(u)(k_i(u) - v_i)$, *where* $\mathbf{v} = (v_1, \ldots, v_n)$ *for* $i = 1, \ldots, n$.
(4) $a_0 = 0$, $a_1 = a$, *and* $a_2 = 1$.
(5) $\pi_0 = (a_0, 0, \mathbf{b_0})^t = (a_0, 0, b_0^1, \ldots, b_0^n)^t = (0, 0, b_0^1, \ldots, b_0^n)^t$,
$\pi_1 = (a_1, 0, \mathbf{b_1})^t = (a_1, 0, b_1^1, \ldots, b_1^n)^t = (a, 0, k_-^1 b_0^1/(k_+^1 a + k_-^1), \ldots, k_-^n b_0^n/(k_+^n a + k_-^n))^t$,
and

$\pi_2 = (a_2, 0, \mathbf{b_2})^t = (a_2, 0, b_2^1, \ldots, b_2^n)^t = (1, 0, k_-^1 b_0^1/(k_+^1 + k_-^1), \ldots, k_-^n b_0^n/(k_+^n + k_-^n))^t$, *where the* t *denotes transposition of a vector.*

(6) *For two vectors* $\mathbf{c} = (c_1, \ldots, c_n)$ *and* $\mathbf{d} = (d_1, \ldots, d_n)$, *the symbol* $\mathbf{c} < \mathbf{d}$ *means* $c_i < d_i$ *for* $i = 1, \ldots, n$, *and* $\mathbf{c} \leq \mathbf{d}$ *means* $c_i \leq d_i$ *for* $i = 1, \ldots, n$.

(7) $\mathbf{R}^+$ *stands for the set of all of the positive real numbers.*

In terms of this notation, we can rewrite (1.4)–(1.5) as the following system:

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + u(u-a)(1-u) + \sum_{i=1}^{n} \gamma_i(u)(k_i(u) - v_i), \tag{2.1}$$

$$\frac{\partial v_i}{\partial t} = \gamma_i(u)(k_i(u) - v_i),\ i = 1, \ldots, n. \tag{2.2}$$

The following lemma follows from the associated definitions and a simple calculation.

LEMMA 2.1.

(1) F *and* G_i *are analytic on* $[0,1] \times [0, \infty)$, *and* $\gamma_i(u)$, $k_i(u)$ *are analytic and positive on* $[0,1]$ *for* $i = 1, \ldots, n$.

(2) $F_u(u, \mathbf{v}) = u(2-3u) + a(2u-1) - \sum_{i=1}^n k_+^i v_i$, $F_{v_i}(u, \mathbf{v}) = -\gamma_i(u) < 0$, $G_{i,u}(u, \mathbf{v}) = -k_+^i v_i \leq 0$, *and* $G_{i,v_i}(u, \mathbf{v}) = -\gamma_i(u) < 0$ *for all* $(u, \mathbf{v}) \in [0,1] \times [0, \infty)^n$.

(3) *If* $a \in (0, 1/2)$, *then* $\int_{a_0}^{a_2} F(u, k_1(u), k_2(u), \ldots, k_n(u))du > 0$.

3. Existence and uniqueness of the traveling wave solution.

3.1. Outline of the proof. First, we set up our problem. Namely, we shall look for a traveling wave solution of (2.1)–(2.2) in the form $u = \mathcal{U}(\xi), v_i = \Pi_i(\xi), i = 1, \ldots, n$, with $\xi = x - ct$. Therefore, $(\mathcal{U}, \mathbf{\Pi})$ satisfies

$$\ddot{\mathcal{U}} = (-c\dot{\mathcal{U}} - F(\mathcal{U}, \mathbf{\Pi}))/D, \tag{3.1}$$

$$\dot{\Pi}_i = -G_i(\mathcal{U}, \mathbf{\Pi})/c,\ i = 1, \ldots, n, \tag{3.2}$$

with the boundary conditions

$$(\mathcal{U}(-\infty), \dot{\mathcal{U}}(-\infty), \mathbf{\Pi}(-\infty)) = (a_0, 0, \mathbf{b_0}) \tag{3.3}$$

and

$$(\mathcal{U}(+\infty), \dot{\mathcal{U}}(+\infty), \mathbf{\Pi}(+\infty)) = (a_2, 0, \mathbf{b_2}), \tag{3.4}$$

where $\mathbf{\Pi} = (\Pi_1, \ldots, \Pi_n)$ and $\cdot = d/d\xi$.

By setting $\tau = (-\xi)/c$, $' = d/d\tau$, $\theta = c^2/D$, and $\mathcal{U}' = \mathcal{Z}$, we can rewrite the above problem as the following first-order system of differential equations:

$$\mathcal{U}' = \mathcal{Z}, \tag{3.5}$$

$$\mathcal{Z}' = \theta(\mathcal{Z} - F(\mathcal{U}, \mathbf{\Pi})), \tag{3.6}$$

$$\Pi_i' = G_i(\mathcal{U}, \mathbf{\Pi}),\ i = 1, \ldots, n, \tag{3.7}$$

with the boundary conditions

$$(\mathcal{U}(-\infty), \mathcal{Z}(-\infty), \mathbf{\Pi}(-\infty)) = (a_0, 0, \mathbf{b_0}) \tag{3.8}$$

and

$$(\mathcal{U}(+\infty), \mathcal{Z}(+\infty), \mathbf{\Pi}(+\infty)) = (a_2, 0, \mathbf{b_2}). \tag{3.9}$$

Therefore, in order to solve the problem of existence of traveling wave solutions of (2.1)–(2.2), it suffices to consider the problem (3.5)–(3.9). Note that from the definitions of F and G_i, $i = 1, \dots, n$, it follows that (3.5)–(3.7) have three equilibrium solutions: π_0, π_1, and π_2.

Our strategy is to use the shooting method. We briefly describe the procedure as follows. First, we let $(\mathcal{U}_\theta, \mathcal{Z}_\theta, \mathbf{\Pi}_\theta)$ denote a solution of (3.5)–(3.8) and $(-\infty, T_\theta)$ be the corresponding maximal existence interval of $(\mathcal{U}_\theta, \mathcal{Z}_\theta, \mathbf{\Pi}_\theta)$ (we will omit the subscript θ in the following subsections if there is no ambiguity). Thus our goal is to choose suitable θ to make $(\mathcal{U}_\theta, \mathcal{Z}_\theta, \mathbf{\Pi}_\theta)$ satisfy (3.9). In section 3.2, we will analyze the behavior of $(\mathcal{U}_\theta, \mathcal{Z}_\theta, \mathbf{\Pi}_\theta)$ around $\tau = -\infty$ (see Lemma 3.1) and obtain the uniqueness (up to translation) of the traveling wave solution of (2.1)–(2.2) with given negative wave speed c if it exists. Next, in section 3.3, we will prove two technical lemmas and use one of them to prove a necessary condition for the existence of a monotone traveling wave solution of (2.1)–(2.2) with negative wave speed. Then, in section 3.4, we will give a proof of the existence of the traveling wave solution of (2.1)–(2.2). Roughly speaking, we will consider the following two sets:

$$\mathcal{P}_1 = \{\theta > 0 \mid \mathcal{U}'_\theta > 0 \text{ on } (-\infty, \hat{\tau}] \text{ and } \mathcal{U}_\theta(\hat{\tau}) = a_2 \text{ for some finite } \hat{\tau}\}$$

and

$$\mathcal{P}_2 = \{\theta > 0 \mid \mathcal{U}'_\theta(\tau) = 0 \text{ for some } \tau \in \mathbf{R} \text{ and } \mathcal{U}_\theta(\tau) \in (a_1, a_2]\}.$$

Then we will show that $\mathcal{P}_1$ is open and contains $(\theta_1, +\infty)$ for some $\theta_1 > 0$ (see Step 2 in the proof of Lemma 3.5) and that $\mathcal{P}_2$ is open and contains $(0, \theta_2)$ for some $\theta_2 > 0$ (see Step 3 in the proof of Lemma 3.5). Therefore, $\theta^* := \sup \mathcal{P}_2$ exists and $\theta^* \in \mathbf{R}^+ \setminus (P_1 \cup P_2)$, and so $(\mathcal{U}_{\theta^*}, \mathbf{\Pi}_{\theta^*})$ is our desired solution (see Step 4 in the proof of Lemma 3.5).

Finally, we will use the comparison method to prove the uniqueness of the wave speed of the traveling wave front in section 3.5.

3.2. The behavior of $(\mathcal{U}, \mathcal{Z}, \mathbf{\Pi})$ around $\tau = -\infty$. Linearizing (3.5)–(3.7) around the constant solution π_0, we obtain the equation $d\mathbf{X}/d\tau = A_0\mathbf{X}$, where $\mathbf{X} = (\mathcal{U} - a_0, \mathcal{Z}, \Pi_1 - b_0^1, \dots, \Pi_n - b_0^n)^t$,

$$A_0 = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ -\theta F_u & \theta & -\theta F_{v_1} & \cdots & -\theta F_{v_n} \\ G_{1,u} & 0 & G_{1,v_1} & \cdots & G_{1,v_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ G_{n,u} & 0 & G_{n,v_1} & \cdots & G_{n,v_n} \end{bmatrix} \quad \text{with } G_{i,v_j} = \partial G_i/\partial v_j, \tag{3.10}$$

and all the values are evaluated at $(\mathcal{U}, \mathcal{Z}, \mathbf{\Pi}) = \pi_0$. Thus the associated characteristic polynomial with A_0 is

$$p_n(\lambda) = \det \begin{bmatrix} -\lambda & 1 & 0 & \cdots & 0 \\ -\theta F_u & \theta - \lambda & -\theta F_{v_1} & \cdots & -\theta F_{v_n} \\ G_{1,u} & 0 & G_{1,v_1} - \lambda & \cdots & G_{1,v_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ G_{n,u} & 0 & G_{n,v_1} & \cdots & G_{n,v_n} - \lambda \end{bmatrix}. \tag{3.11}$$

Note that $\partial G_i/\partial v_j = 0$ if $i \neq j$ and $G_{i,v_i}(a_0, \mathbf{b_0}) = -k_-^i$ for $i = 1, \dots, n$. Then, by Lemma 5.1 in the appendix, we may assume that the eigenvalues $\lambda_1, \dots, \lambda_{n+2}$ of A_0 satisfy $\lambda_{n+2} < \cdots < \lambda_2 < 0 < \lambda_1$ (some of the negative ones may be equal).

Let $\bar{\mathbf{X}}$ be an eigenvector of A_0 corresponding to λ_1. Then there is a nonconstant solution Γ of (3.5)–(3.7) which tends to π_0 as $\tau \to -\infty$ and whose tangent vector at $\tau = -\infty$ is the eigenvector $\bar{\mathbf{X}}$ or $-\bar{\mathbf{X}}$. Set $\bar{\mathbf{X}} = (X_1, \ldots, X_{n+2})$. From the equation $A_0\bar{\mathbf{X}} = \lambda_1\bar{\mathbf{X}}$ it follows that

$$X_2 = \lambda_1 X_1, \tag{3.12}$$

$$-\theta(F_u(a_0, \mathbf{b_0})X_1 + \sum_{i=1}^{n} F_{v_i}(a_0, \mathbf{b_0})X_{i+2}) = (\lambda_1 - \theta)X_2, \tag{3.13}$$

$$G_{i,u}(a_0, \mathbf{b_0})X_1 = (\lambda_1 - G_{i,v_i}(a_0, \mathbf{b_0}))X_{i+2} \tag{3.14}$$

for $i = 1, \ldots, n$. Noting that $\lambda_1 > 0$, $G_{i,v_i}(a_0, \mathbf{b_0}) = -k_-^i < 0$, and using (3.12) and (3.14), we can assume that $X_1 > 0$ and $X_2 > 0$. Moreover, from $\lambda_1 > 0$, $G_{i,u}(a_0, \mathbf{b_0}) = -k_+^i b_0^i < 0$, $G_{i,v_i}(a_0, \mathbf{b_0}) = -k_-^i < 0$, and (3.14), it follows that $X_{i+2} < 0$ for $i = 1, \ldots, n$.

In the remainder of this section, we assume that a solution $(\mathcal{U}, \mathcal{Z}, \mathbf{\Pi})$ of (3.5)–(3.8) satisfies the condition that the tangent vector $\bar{\mathbf{X}}$ to $(\mathcal{U}, \mathcal{Z}, \mathbf{\Pi})$ at $\tau = -\infty$ has the properties as discussed above. Hence near $\tau = -\infty$, $(\mathcal{U}, \mathcal{Z}, \mathbf{\Pi})$ satisfies $\mathcal{U} > a_0$, $\mathcal{Z} = \mathcal{U}' > 0$, and $\Pi_i' < 0$ for $i = 1, \ldots, n$. From this it follows that the following definition is well defined.

DEFINITION 2. *Let $(\mathcal{U}, \mathcal{Z}, \mathbf{\Pi})$ be a solution of* (3.5)–(3.8). *Let $\tau_0 = \tau_0(\theta)$ be the first zero of $\mathcal{U}'$ if it exists; and set $\tau_0 = T$ if $\mathcal{U}' > 0$ on $(-\infty, T)$. We also set $\bar{u} = \mathcal{U}(\tau_0)$ ($\bar{u}$ may be $+\infty$).*

Furthermore, by this definition and the above discussion, we have the following lemma.

LEMMA 3.1. *Let $(\mathcal{U}, \mathcal{Z}, \mathbf{\Pi})$ be a solution of* (3.5)–(3.8). *Then $\mathcal{U} > 0$ and $\mathcal{U}' > 0$ on $(-\infty, \tau_0)$, and a traveling wave solution of* (2.1)–(2.2) *for given negative speed c is unique (up to translation) if it exists.*

3.3. Two auxiliary lemmas and a necessary condition. First, we will make a transformation for (3.5)–(3.7) which is useful for our discussion. Since $\mathcal{U}' > 0$ on $(-\infty, \tau_0)$, we can express $\mathcal{Z}$ and Π_i, $i = 1, \ldots, n$, as functions of $\mathcal{U}$ for $\mathcal{U} \in (a_0, \bar{u})$. Let $Z(\mathcal{U}) = \mathcal{Z}(\tau(\mathcal{U}))$ and $V_i(\mathcal{U}) = \Pi_i(\tau(\mathcal{U}))$, $i = 1, \ldots, n$, for $\mathcal{U} \in (a_0, \bar{u})$. Set $\mathbf{V} = (V_1, \ldots, V_n)$. Then Z and V_i satisfy the following equations:

$$Z' := \frac{dZ}{d\mathcal{U}} = \theta\left(1 - \frac{F(\mathcal{U}, \mathbf{V})}{Z}\right), \tag{3.15}$$

$$V_i' := \frac{dV_i}{d\mathcal{U}} = \frac{G_i(\mathcal{U}, \mathbf{V})}{Z},\ i = 1, \ldots, n, \tag{3.16}$$

for $\mathcal{U} \in (a_0, \bar{u})$ with the initial conditions

$$Z(a_0) = 0,\ \ V_i(a_0) = b_0^i \ \text{ for } \ i = 1, \ldots, n. \tag{3.17}$$

Note that $Z(u) > 0$ for $u \in (a_0, \bar{u})$ and $Z(\bar{u}) = 0$ if $\bar{u}$ is finite.

LEMMA 3.2. *Let $(\mathcal{U}, \mathcal{Z}, \mathbf{\Pi})$ be a solution of* (3.5)–(3.8). *Then $k_i(\mathcal{U}) < V_i(\mathcal{U}) < b_0^i$ for all $\mathcal{U} \in (a_0, \bar{u})$ and $i = 1, \ldots, n$. Moreover, $\Pi_i'(\tau) < 0$ for all $\tau \in (-\infty, \tau_0)$ and $i = 1, \ldots, n$.*

Proof. Fix $i \in \{1, \ldots, n\}$. Noting that $\Pi_i' < 0$ near $\tau = -\infty$, $\mathcal{U}(\tau) > 0$ for $\tau \in (-\infty, \tau_0)$ and using (3.7), we have $V_i(\mathcal{U}) > k_i(\mathcal{U})$ for all $\mathcal{U} \in (a_0, u_3)$ for some $u_3 \in (a_0, \bar{u}]$. If $u_3 < \bar{u}$ and $V_i(u_3) = k_i(u_3)$, then $V_i'(u_3) \le k_i'(u_3) < 0$, where we

have used the fact that k_i is decreasing on $\mathbf{R}$. On the other hand, by (3.16) we have $V_i'(u_3) = 0$, a contradiction. Therefore, $V_i(\mathcal{U}) > k_i(\mathcal{U})$ for all $\mathcal{U} \in (a_0, \bar{u})$. Combining this with Lemma 3.1, we have $V_i' < 0$ on $(a_0, \bar{u})$, and so $\Pi_i' < 0$ on $(-\infty, \tau_0)$. Moreover, from $V_i(a_0) = b_0^i$ it follows that $V_i(\mathcal{U}) < b_0^i$ for all $\mathcal{U} \in (a_0, \bar{u})$. This completes the proof. ∎

In the next lemma, we will show that $a \in (0, 1/2)$ is a necessary condition for the existence of a monotone traveling wave solution of (2.1)–(2.2) with negative wave speed.

LEMMA 3.3 (necessary condition). *Let $a > 0$, and suppose that* (3.1)–(3.2) *with the boundary conditions* (3.3)–(3.4) *and $c < 0$ has a solution $(\mathcal{U}, \mathbf{\Pi})$ with $\dot{\mathcal{U}} > 0$ on $\mathbf{R}$. Then we have $a \in (0, 1/2)$.*

Proof. Rewrite (3.1)–(3.2) as the following:

$$\begin{aligned} -D\ddot{\mathcal{U}} &= c\dot{\mathcal{U}} + F(\mathcal{U}, \mathbf{\Pi}), \\ 0 &= c\dot{\Pi}_i + G_i(\mathcal{U}, \mathbf{\Pi}),\ i = 1, \dots, n. \end{aligned}$$

Then multiplying these two equations by $\dot{\mathcal{U}}$, integrating from $-\infty$ to $+\infty$, and summing the second equation from $i = 1$ to n, and then subtracting the first equation from the resulting second one, we get

$$-c\left(\int_{-\infty}^{+\infty} \dot{\mathcal{U}}^2(\xi)d\xi - \sum_{i=1}^{n}\int_{-\infty}^{+\infty} \dot{\mathcal{U}}(\xi)\dot{\Pi}_i(\xi)d\xi\right) = \int_{a_0}^{a_2} u(u-a)(1-u)du.$$

By Lemma 3.2, $\dot{\Pi}_i(\xi) = (-1/c)\Pi_i'(\tau) < 0$ on $\mathbf{R}$ for $i = 1, \dots, n$. Thus the left-hand side of the above equation is positive. From this and Lemma 2.1 it follows that $a \in (0, 1/2)$. Hence the proof is completed. ∎

LEMMA 3.4. *There exists no solution $(\mathcal{U}, \mathbf{\Pi})$ of* (3.5)–(3.7) *with the condition* (3.8) *satisfying that $\mathcal{U}(\tau_0) = a_2$, $\mathcal{U}'(\tau_0) = 0$ and $\mathcal{U}''(\tau_0) \le 0$ for some finite τ_0.*

Proof. Suppose that there is such a solution. Thus either $\mathcal{U}''(\tau_0) = 0$ or $\mathcal{U}''(\tau_0) < 0$. In the first case, it follows from Lemma 3.2 that $\Pi_i'(\tau_0) \le 0$ for $i = 1, \dots, n$. Therefore, noting the fact that if the sum of a sequence of nonnegative numbers is zero, then every number in this sequence must be zero, using (3.6)–(3.7), and the definitions of F and G_i, $i = 1, \dots, n$, it follows that $\Pi_i'(\tau_0) = 0$ for $i = 1, \dots, n$, a contradiction to the uniqueness theorem for the differential equations. For the latter case, by (3.6)–(3.7) again and a similar argument as the above, we have $\Pi_{i_0}'(\tau_0) > 0$ for some $i_0 \in \{1, \dots, n\}$, a contradiction to Lemma 3.2. This completes the proof. ∎

3.4. Final proof. Now we are ready to prove the existence of a traveling wave solution of (2.1)–(2.2) under the assumption $a \in (0, 1/2)$. The proof consists of four steps.

LEMMA 3.5. *If $a \in (0, 1/2)$, then there exists $c := c(a) < 0$ such that there is a unique solution (up to translation) $(\mathcal{U}(\tau), \mathcal{Z}(\tau), \mathbf{\Pi}(\tau))$ of* (3.5)–(3.9) *satisfying that $\mathcal{U}' > 0$ and $\Pi_i' < 0$, $i = 1, \dots, n$, on $\mathbf{R}$.*

Proof. Step 1. *We claim that $\bar{u} > a_1$, $Z(a_1) \ge \theta(a_1 - a_0)$, and $V_i(a_1) > k_i(a_1)$ for $i = 1, \dots, n$.* Indeed, by Lemma 3.2 and Lemma 3.1 we have $V_i' < 0$ and $G_i(\mathcal{U}, \mathbf{V}(\mathcal{U})) < 0$ for all $\mathcal{U} \in (a_0, \bar{u})$ and $i = 1, \dots, n$. Also note that $f(u) < 0$ for $u \in (a_0, a_1)$. Combining these two facts with the definitions of F and G_i, $i = 1, \dots, n$, we obtain that $F(\mathcal{U}, \mathbf{V}(\mathcal{U})) < 0$ and $G_i(\mathcal{U}, \mathbf{V}(\mathcal{U})) < 0$ for all $\mathcal{U} \in (a_0, \min\{\bar{u}, a_1\})$ and $i = 1, \dots, n$. Thus if $\bar{u} \le a_1$, then from (3.15) it follows that $Z' \ge \theta$ on $(a_0, \bar{u})$. This

implies that $Z(\bar{u}) > 0$, a contradiction. Therefore, we have $\bar{u} > a_1$. Moreover, noting that $F(\mathcal{U}, \mathbf{V}(\mathcal{U})) < 0$ and $Z > 0$ on $(a_0, a_1]$, and using (3.15) again, it follows that $Z' \geq \theta$ on $(a_0, a_1]$. Hence $Z(a_1) \geq \theta(a_1 - a_0)$ and $V_i(a_1) > k_i(a_1)$ for $i = 1, \ldots, n$.

Step 2. *We claim that if θ is sufficiently large, then there exists a finite $\hat{\tau}$ such that $\mathcal{U}' > 0$ on $(-\infty, \hat{\tau}]$ and $\mathcal{U}(\hat{\tau}) = a_2$.* By Lemma 3.2, we have $k_i(\mathcal{U}) < V_i(\mathcal{U}) < b_0^i$ for all $\mathcal{U} \in (a_0, \bar{u})$ and $i = 1, \ldots, n$. Let

$$B = \sup\{|\, F(\mathcal{U}, \mathbf{V}) \,|\, |\, a_0 \leq \mathcal{U} \leq a_2, k_i(\mathcal{U}) \leq V_i \leq b_0^i, i = 1, \ldots, n\}.$$

Choose θ_0 such that $\theta_0 > 4B/(a_1 - a_0)$. Then we claim that $\bar{u} > a_2$ for all $\theta > \theta_0$. Suppose that the claim does not hold for some $\tilde{\theta} > \theta_0$. Let $(Z_{\tilde{\theta}}, V_{\tilde{\theta},1}, \ldots, V_{\tilde{\theta},n})$ be the corresponding solution of (3.15)–(3.17). Then, since $Z_{\tilde{\theta}}(a_1) \geq \tilde{\theta}(a_1 - a_0)$ by Step 1, we have $Z_{\tilde{\theta}}(\mathcal{U}) > \tilde{\theta}(a_1 - a_0)/2$ for all $\mathcal{U} \in [a_1, \hat{u})$ for some $\hat{u} \in (a_1, a_2)$ and $Z_{\tilde{\theta}}(\hat{u}) = \tilde{\theta}(a_1 - a_0)/2$. On the other hand, from (3.15) it follows that

$$\begin{aligned} Z'_{\tilde{\theta}}(\mathcal{U}) &= \tilde{\theta}\left(1 - \frac{F(\mathcal{U}, \mathbf{V}_{\tilde{\theta}})}{Z_{\tilde{\theta}}}\right) \\ &\geq \tilde{\theta}(1 - B/[\tilde{\theta}(a_1 - a_0)/2]) \\ &> \tilde{\theta}/2 \end{aligned}$$

for all $\mathcal{U} \in [a_1, \hat{u}]$. This implies that $Z_{\tilde{\theta}}(\mathcal{U}) \geq \tilde{\theta}(a_1 - a_0)$ for all $\mathcal{U} \in [a_1, \hat{u}]$, a contradiction. Therefore, if $\theta > \theta_0$, we have $\bar{u} > a_2$, and so $Z(\mathcal{U}) > 0$ for all $\mathcal{U} \in (a_0, a_2]$. Hence there exists a finite $\hat{\tau}$ such that $\mathcal{U}' > 0$ on $(-\infty, \hat{\tau}]$ and $\mathcal{U}(\hat{\tau}) = a_2$. Moreover, if we let $(\mathcal{U}_\theta, \mathcal{Z}_\theta, \mathbf{\Pi}_\theta)$ be a solution of (3.5)–(3.8) and

$$\mathcal{P}_1 = \{\theta > 0 \mid \mathcal{U}'_\theta > 0 \text{ on } (-\infty, \hat{\tau}] \text{ and } \mathcal{U}_\theta(\hat{\tau}) = a_2 \text{ for some finite } \hat{\tau}\},$$

then $\mathcal{P}_1$ is nonempty. Furthermore, $\mathcal{P}_1$ is open by continuous dependence on the parameter θ.

Step 3. *We show that for sufficiently small $\theta > 0$, there is a finite τ_0 with $\mathcal{U}'_\theta(\tau_0) = 0$ and $\mathcal{U}_\theta(\tau_0) \in (a_1, a_2)$.* If not, then there exists a sequence $\{\theta_i\}_{i \in \mathbf{N}}$ with $\lim_{i \to \infty} \theta_i = 0$ such that the corresponding solutions $(\mathcal{U}_i, \mathbf{\Pi}_i)$ of (3.5)–(3.8) satisfy that $\mathcal{U}'_i > 0$ on $(-\infty, \hat{\tau}_i)$ and $\mathcal{U}_i(\hat{\tau}_i) = a_2$ for some $\hat{\tau}_i$ ($\hat{\tau}_i$ may be infinite). Let $\mathbf{V}_i = (V_{i,1}, \ldots, V_{i,n})$. Multiplying (3.15) with Z_i and integrating from a_0 to $\mathcal{U}$, we obtain

$$\frac{Z_i(\mathcal{U})^2}{2} = \theta_i \int_{a_0}^{\mathcal{U}} (Z_i(t) - F(t, \mathbf{V}_i(t)))dt \tag{3.18}$$

for all $\mathcal{U} \in [a_0, a_2]$. By Lemma 3.2, we have $k_j(\mathcal{U}) \leq V_{i,j}(\mathcal{U}) \leq b_0^j$ for all $\mathcal{U} \in [a_0, a_2]$ and $j = 1, \ldots, n$. Note that $Z_i(\mathcal{U}) > 0$ for all $\mathcal{U} \in (a_0, a_2)$. Thus if we let $A_i = \sup_{a_0 \leq \mathcal{U} \leq a_2} Z_i(\mathcal{U})$, then it follows from (3.18) that $A_i^2/2 \leq \theta_i(A_i(a_2 - a_0) + B(a_2 - a_0))$, and so

$$\lim_{i \to \infty} A_i = 0. \tag{3.19}$$

Solving (3.16) with integration by parts, we obtain that

$$V_{i,j}(\mathcal{U}) = k_j(\mathcal{U}) - \int_{a_0}^{\mathcal{U}} k'_j(t) \exp\left[-\int_t^{\mathcal{U}} (\gamma_j(s)/Z_i(s))ds\right] dt$$

for all $\mathcal{U} \in [a_0, a_2]$ and $j = 1, \dots, n$. Thus it follows from (3.19) that

$$\lim_{i\to\infty} V_{i,j}(\mathcal{U}) = k_j(\mathcal{U}) \tag{3.20}$$

uniformly for $\mathcal{U} \in [a_0, a_2]$ and $j = 1, \dots, n$. From (3.18) it follows that

$$\int_{a_0}^{a_2} F(t, \mathbf{V}_i(t))dt \le \int_{a_0}^{a_2} Z_i(t)dt. \tag{3.21}$$

Letting $i \to \infty$ in (3.21), using (3.19) and (3.20), we obtain that

$$\int_{a_0}^{a_2} F(s, k_1(t), \dots, k_n(t))dt \le 0,$$

a contradiction to Lemma 2.1.

Step 4. *We reach the conclusion.* Let $(\mathcal{U}_\theta, \mathcal{Z}_\theta, \mathbf{\Pi}_\theta)$ be a solution of (3.5)–(3.8) and

$$\mathcal{P}_2 = \{\theta > 0 \mid \mathcal{U}'_\theta(\tau) = 0 \text{ for some } \tau \in \mathbf{R} \text{ and } \mathcal{U}_\theta(\tau) \in (a_1, a_2]\}.$$

Thus $\mathcal{P}_2$ is nonempty by Step 3. For each $\theta \in \mathcal{P}_2$, recall that $\tau_0 = \tau_0(\theta)$ denotes the first zero of $\mathcal{U}'_\theta$. Then we have $\mathcal{U}''_\theta(\tau_0) \le 0$. Moreover, by Lemma 3.4 and Step 1, we have $\mathcal{U}_\theta(\tau_0) \in (a_1, a_2)$. Next we claim that $\mathcal{U}''_\theta(\tau_0) < 0$. If not, then $\mathcal{U}''_\theta(\tau_0) = 0$. By (3.5)–(3.6), we have $F(\mathcal{U}_\theta(\tau_0), \mathbf{\Pi}_\theta(\tau_0)) = 0$. Also recall from Lemma 3.2 that $\Pi'_{\theta,i}(\tau_0) \le 0$ for $i = 1, \dots, n$. Combining these two facts with the fact that $f(u) > 0$ for $u \in (a_1, a_2)$, we obtain that $\Pi'_{\theta,i_0}(\tau_0) < 0$ for some $i_0 \in \{1, \dots, n\}$. From this, (3.6), and part (2) of Lemma 2.1 it follows that $\mathcal{U}'''_\theta(\tau_0) = -\theta \sum_{i=1}^n F_{v_i}(\mathcal{U}_\theta(\tau_0), \mathbf{\Pi}_\theta(\tau_0))\Pi'_{\theta,i}(\tau_0) < 0$, a contradiction to the definition of τ_0. Thus we have $\mathcal{U}''_\theta(\tau_0) < 0$, and this implies that $\mathcal{P}_2$ is open.

By Step 2 and the above discussion, $\mathcal{P}_2$ is nonempty, open, and bounded above. Therefore $\theta^* := \sup \mathcal{P}_2$ exists and $\theta^* \in \mathbf{R}^+ \setminus (P_1 \cup P_2)$. Let $(\mathcal{U}_{\theta^*}, \Pi_{\theta^*,1}, \dots, \Pi_{\theta^*,n})$ be the corresponding solution of (3.5)–(3.8). Then $\mathcal{U}'_{\theta^*} > 0$ on $\mathbf{R}$ and $\mathcal{U}_{\theta^*}(\tau) \to a_2$ as $\tau \to +\infty$. Moreover, by Lemma 3.2 we have $\Pi'_{\theta^*,i} < 0$ on $\mathbf{R}$ for $i = 1, \dots, n$. Now we claim that $\Pi_{\theta^*,i}(\tau) \to b_2^i$ as $\tau \to +\infty$ for $i = 1, \dots, n$. Indeed, fix $i \in \{1, \dots, n\}$; using Lemma 3.2 and noting that $\mathcal{U}_{\theta^*} \in (a_0, a_2)$ on $\mathbf{R}$, we have $\Pi_{\theta^*,i} \in (b_2^i, b_0^i)$ on $\mathbf{R}$. Using this fact and noting that $\Pi'_{\theta^*,i} < 0$ on $\mathbf{R}$, it follows that there exists $l \in [b_2^i, b_0^i)$ such that $\Pi_{\theta^*,i}(\tau) \to l$ as $\tau \to +\infty$. Hence we can choose a sequence $s_1 < s_2 < \cdots < s_m < \cdots$ with $s_m \to +\infty$ as $m \to +\infty$ satisfying that $\mathcal{U}_{\theta^*}(s_m) \to a_2$, $\Pi_{\theta^*,i}(s_m) \to l$, and $\Pi'_{\theta^*,i}(s_m) \to 0$ as $m \to +\infty$. From this and (3.7) it follows that $l = b_2^i$. Hence the proof is completed. □

Finally, combining Lemma 3.3 with Lemma 3.5, we obtain the following theorem.

THEOREM 1. *There exists $c := c(a) < 0$ such that a unique (up to translation) traveling wave solution $(\mathcal{U}, \mathbf{\Pi})$ of our buffered bistable equations* (1.4)–(1.5) *with wave speed c, such that $\dot{\mathcal{U}} > 0$ and $\ddot{\Pi}_i < 0$, $i = 1, \dots, n$, on $\mathbf{R}$, exists if and only if $a \in (0, 1/2)$.*

3.5. Uniqueness of wave speed. In this subsection, we will concern ourselves with the uniqueness of wave speed. First, throughout the remainder of this section, $(\mathcal{U}, \mathbf{\Pi})$ will denote such a unique traveling wave solution of (1.4)–(1.5) with speed c and $\mathcal{U}(0) = 1/2$, which is described in Theorem 1. Since the proof is based on the use of supersolution (subsolution) of (2.1)–(2.2) and the comparison principle, we state the definition of supersolution (subsolution) as follows.

DEFINITION 3. *A set of bounded functions* $(u, \mathbf{v})$ *is called a subsolution of* (2.1)–(2.2) *in* $\mathbf{R}\times\mathbf{R}^+$, *if* $u, v_i \in C^{2,1}(\mathbf{R}\times\mathbf{R}^+)$ *and* $C^{1,1}(\mathbf{R}\times\mathbf{R}^+)$, *respectively, and* $(u, \mathbf{v})$ *satisfies that* $u_t - Du_{xx} \le F(u, \mathbf{v})$, $v_{i,t} \ge G_i(u, \mathbf{v})$, *and*

$$u \ge -\min_{j=1,\dots,n}\{k_-^j/(2k_+^j)\}, \quad v_i(z,t) \ge b_2^i/2 \tag{3.22}$$

on $\mathbf{R}\times\mathbf{R}^+$ *for* $i = 1, \dots, n$. *Supersolution is defined by reversing the inequalities with* (3.22) *held.*

The next lemma is a comparison theorem for the supersolution and subsolution, whose proof is similar to Lemma 5.3 in the appendix.

LEMMA 3.6 (comparison principle). *Let* $(u_1, \mathbf{v}_1)$ *and* $(u_2, \mathbf{v}_2)$ *be the subsolution and supersolution of* (2.1)–(2.2) *on* $\mathbf{R}\times\mathbf{R}^+$, *respectively, with* $u_1(x,0) \le u_2(x,0)$ *and* $\mathbf{v}_2(x,0) \le \mathbf{v}_1(x,0)$ *for all* $x \in \mathbf{R}$. *Then the following statement holds:*

$$u_1(x,t) \le u_2(x,t) \ \text{ and } \ \mathbf{v}_2(x,t) \le \mathbf{v}_1(x,t) \ \text{ for all } \ (x,t) \in \mathbf{R}\times\mathbf{R}^+.$$

Next, from the above lemma (also see the proof of Lemma 5.3 in the appendix), it is easily seen that the following lemma holds.

LEMMA 3.7. *There exist positive constants* d_1, μ_0, *and* k_{0i}, $i = 1, \dots, n$, *which are independent of* $(\mathcal{U}, \mathbf{\Pi})$, *and a positive constant* ν *such that, for any* $d \in (0, d_1]$ *and* $x_0 \in \mathbf{R}$, *the functions* $(w^+, \mathbf{p}^+)$ *and* $(w^-, \mathbf{p}^-)$ *defined by*

$$\begin{aligned} w^{\pm}(x,t) &:= \mathcal{U}(x - ct + x_0 \pm \nu d(1 - e^{-\mu_0 t})) \pm de^{-\mu_0 t}, \\ p_i^{\pm}(x,t) &:= \Pi_i(x - ct + x_0 \pm \nu d(1 - e^{-\mu_0 t})) \mp dk_{0i}e^{-\mu_0 t}, \ i = 1, \dots, n, \end{aligned} \tag{3.23}$$

are a supersolution and a subsolution of (2.1)–(2.2), *respectively.*

Now we are ready to prove the uniqueness of wave speed of a traveling wave front of (2.1)–(2.2).

LEMMA 3.8 (uniqueness). *Given* $a \in (0, 1/2)$, *for any traveling wave solution* $(\tilde{\mathcal{U}}, \tilde{\mathbf{\Pi}})$ *of* (2.1)–(2.2) *with wave speed* $\tilde{c}$ *such that* $\tilde{\mathcal{U}} \in [0,1]$ *and* $\tilde{\mathbf{\Pi}} \in [\mathbf{b_2}, \mathbf{b_0}]$ *on* $\mathbf{R}$, *we have* $\tilde{c} = c$ *and* $(\tilde{\mathcal{U}}(\cdot), \tilde{\mathbf{\Pi}}(\cdot)) = (\mathcal{U}(\cdot + \xi_0), \mathbf{\Pi}(\cdot + \xi_0))$ *for some* $\xi_0 \in \mathbf{R}$.

Proof. We assume that there exists another traveling wave solution $(\tilde{\mathcal{U}}, \tilde{\mathbf{\Pi}})$ of (2.1)–(2.2) with wave speed $\tilde{c}$. Note that $(\tilde{\mathcal{U}}, \tilde{\mathbf{\Pi}})$ is not necessarily monotone. The proof is divided into two steps.

Step 1. *We claim that* $c = \tilde{c}$. We will follow X. Chen's arguments in [1] to prove this claim, where Chen studied the existence, uniqueness, and stability of a single nonlocal equation of the form

$$u_t(x,t) = \mathcal{A}[u(\cdot,t)](x), \ x \in \mathbf{R}, \ t > 0.$$

Since $(\tilde{\mathcal{U}}(\xi), \tilde{\mathbf{\Pi}}(\xi))$ and $(\mathcal{U}(\xi), \mathbf{\Pi}(\xi))$ satisfy the same boundary condition at $\pm\infty$, we can choose a sufficiently large ξ_1 such that the following hold:

$$\mathcal{U}(\cdot - \xi_1) - d_1 < \tilde{\mathcal{U}}(\cdot) \ \text{ and } \ \Pi_i(\cdot - \xi_1) + d_1 k_{0i} > \tilde{\Pi}_i(\cdot) \ \text{ on } \ \mathbf{R} \ \text{ for } \ i = 1, \dots, n. \tag{3.24}$$

By a translation if necessary, we may assume that $\xi_1 = 0$. Next we choose a sufficiently large ξ_2 such that the inequalities

$$\mathcal{U}(\cdot) - d_1 < \tilde{\mathcal{U}}(\cdot) < \mathcal{U}(\cdot + \xi_2) + d_1 \ \text{ and } \ \Pi_i(\cdot + \xi_2) - d_1 k_{0i} < \tilde{\Pi}_i(\cdot) < \Pi_i(\cdot) + d_1 k_{0i}$$

hold on $\mathbf{R}$ for $i = 1, \dots, n$. Using the above inequalities and applying Lemma 3.6 to $(\tilde{\mathcal{U}}(x-\tilde{c}t), \tilde{\mathbf{\Pi}}(x-\tilde{c}t))$ and $(w^{\pm}(x,t), \mathbf{p}^{\pm}(x,t))$ in (3.23) (with $x_0 = 0$ for $(w^-, \mathbf{p}^-)$ and $x_0 = \xi_2$ for $(w^+, \mathbf{p}^+)$), we obtain that

$$\begin{aligned} &\mathcal{U}(x - ct - \nu d_1(1 - e^{-\mu_0 t})) - d_1 e^{-\mu_0 t} \le \tilde{\mathcal{U}}(x - \tilde{c}t) \\ &\qquad\qquad \le \mathcal{U}(x - ct + \xi_2 + \nu d_1(1 - e^{-\mu_0 t})) + d_1 e^{-\mu_0 t} \end{aligned} \tag{3.25}$$

and

$$\begin{aligned} \Pi_i(x - ct + \xi_2 + \nu d_1(1 - e^{-\mu_0 t})) - d_1 k_{0i} e^{-\mu_0 t} &\le \tilde{\Pi}_i(x - \tilde{c}t) \\ &\le \Pi_i(x - ct - \nu d_1(1 - e^{-\mu_0 t})) \\ &\quad + d_1 k_{0i} e^{-\mu_0 t} \end{aligned} \tag{3.26}$$

for all $(x,t) \in \mathbf{R} \times \mathbf{R}^+$ and $i = 1, \dots, n$. Note that $\mathcal{U}(+\infty) = 1$ and that $\tilde{\mathcal{U}} \in [0,1]$ is not identical to the constant 1. Letting $x - \tilde{c}t$ be fixed and $t \to +\infty$ in the first inequality of (3.25), we obtain $c \ge \tilde{c}$. A similar argument for the second inequality of (3.25) leads to the conclusion $c \le \tilde{c}$. Hence, to summarize, we have $c = \tilde{c}$. This proves this claim.

Step 2. *We claim that* $(\mathcal{U}(\cdot), \mathbf{\Pi}(\cdot)) \equiv (\tilde{\mathcal{U}}(\cdot + \xi_0), \tilde{\mathbf{\Pi}}(\cdot + \xi_0))$ *for some* $\xi_0 \in \mathbf{R}$. Indeed, since we have $\tilde{c} = c$ by Step 1, then our claim follows from the final statement of Lemma 3.1. This completes the proof. ∎

Now we can summarize what we have done. Indeed, by Lemma 3.1, Theorem 1, and Lemma 3.8, we obtain the following theorem.

THEOREM 2. *If* $a \in (0, 1/2)$, *then there exists a unique* $c := c(a) < 0$ *such that a unique (up to translation) traveling wave solution* $(\mathcal{U}, \mathbf{\Pi})$ *of our buffered bistable equations* (1.4)–(1.5) *with wave speed* c *exists. Moreover,* $(\mathcal{U}, \mathbf{\Pi})$ *satisfies that* $\dot{\mathcal{U}} > 0$ *and* $\dot{\Pi}_i < 0$, $i = 1, \dots, n$, *on* $\mathbf{R}$. *Furthermore, a traveling wave solution* $(\mathcal{U}, \mathbf{\Pi})$ *of* (1.4)–(1.5) *with negative wave speed, such that* $\dot{\mathcal{U}} > 0$ *and* $\dot{\Pi}_i < 0$, $i = 1, \dots, n$, *on* $\mathbf{R}$, *exists only if* $a \in (0, 1/2)$.

Comparing this result with the unbuffered equation (1.1), we may conclude that stationary buffers cannot eliminate wave activity.

4. The proof of stability of the traveling wave solution.

4.1. Statement of stability of the traveling wave solution. Now we investigate the stability of the traveling wave solution of the problem

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + f(u) + \sum_{i=1}^{n} [k_-^i(b_0^i - v_i) - k_+^i u v_i], \tag{4.1}$$

$$\frac{\partial v_i}{\partial t} = k_-^i(b_0^i - v_i) - k_+^i u v_i, \ (x,t) \in \mathbf{R} \times \mathbf{R}^+, \ i = 1, \dots, n, \tag{4.2}$$

with the initial data

$$u(x,0) = \phi(x) \ \text{ and } \ v_i(x,0) = \psi_i(x), \ x \in \mathbf{R}, \ \text{ for } \ i = 1, \dots, n.$$

First, for each solution $(u, \mathbf{v})$ of (4.1)–(4.2), we introduce the moving coordinate $z = x - ct$ and set $\tilde{u}(z,t) = u(z + ct, t)$ and $\tilde{v}_i(z,t) = v_i(z + ct, t)$, $i = 1, \dots, n$, where c is a constant (wave speed). Let $\tilde{\mathbf{v}} = (\tilde{v}_1, \dots, \tilde{v}_n)$. Then $\tilde{u}$ and $\tilde{v}_i$ satisfy the following system:

$$L_1[\tilde{u}, \mathbf{v}] \equiv \tilde{u}_t - D\tilde{u}_{zz} - c\tilde{u}_z = F(\tilde{u}, \tilde{\mathbf{v}}), \tag{4.3}$$

$$L_{2i}[\tilde{u}, \tilde{\mathbf{v}}] \equiv \tilde{v}_{i,t} - c\tilde{v}_{i,z} = G_i(\tilde{u}, \tilde{\mathbf{v}}), \ i = 1, \dots, n, \tag{4.4}$$

with the initial data

$$\tilde{u}(z,0)=\phi(z) \text{ and } \tilde{v}_i(z,0)=\psi_i(z),\ z\in\mathbf{R}, \text{ for } i=1,\dots,n, \tag{4.5}$$

where $\tilde{v}_{i,t}=\partial\tilde{v}_i/\partial t$ and $\tilde{v}_{i,z}=\partial\tilde{v}_i/\partial z$. Hereafter for notational convenience we shall suppress the tilde.

We will briefly discuss the existence of the global solution of (4.3)–(4.5) and its associated properties. Indeed, we assume throughout the remainder of this paper that $\phi(z)$ and $\psi_i(z)$, $i=1,\dots,n$, are sufficiently smooth and satisfy that $a_0\le\phi(z)\le a_2$ and $b_2^i\le\psi_i(z)\le b_0^i$ for all $z\in\mathbf{R}$ and $i=1,\dots,n$. Then a similar argument as in Rauch and Smoller [17] shows that the problem (4.3)–(4.5) has a unique smooth solution $(u(z,t),\mathbf{v}(z,t))$ on $\mathbf{R}\times[0,t_0]$ for some $t_0>0$. Furthermore, we can show that $(u(z,t),\mathbf{v}(z,t))$ is a global solution of (4.3)–(4.5). In fact, we have the following lemma.

LEMMA 4.1 (invariance region). *Let $(u,\mathbf{v})$ be the solution of* (4.3)–(4.5). *Then $(u,\mathbf{v})$ exists for all $t>0$. Moreover, $a_0\le u(z,t)\le a_2$ and $b_2^i\le v_i(z,t)\le b_0^i$ for all $(z,t)\in\mathbf{R}\times\mathbf{R}^+$ and $i=1,\dots,n$.*

Proof. The proof follows from Theorem 14.11 on p. 203 and Corollary 14.9 on p. 202 of Smoller [23]. Also see Redheffer and Walter [18]. The outer normal conditions on the boundary of the set $\{(u,\mathbf{v})\mid a_0\le u\le a_2, b_2^i\le v_i\le b_0^i, i=1,\dots,n\}$ follow from the definitions of F and G_i, $i=1,\dots,n$. □

Recall that for each $D>0$ and $a\in(0,1/2)$, there exists a unique $c<0$ such that there is a unique solution $(\mathcal{U}(z),\ \mathbf{\Pi}(z))$ of the steady state equation

$$D\mathcal{U}_{zz}+c\mathcal{U}_z+F(\mathcal{U},\mathbf{\Pi})=0, \tag{4.6}$$

$$c\Pi_{i,z}+G_i(\mathcal{U},\mathbf{\Pi})=0 \tag{4.7}$$

satisfying that

$$\mathcal{U}(0)=(a_0+a_2)/2,\ \mathcal{U}(-\infty)=a_0,\ \Pi_i(-\infty)=b_0^i,\ \mathcal{U}(+\infty)=a_2,\ \Pi_i(+\infty)=b_2^i \tag{4.8}$$

and that

$$\mathcal{U}'>0,\ \ \Pi_i'<0 \text{ on } \mathbf{R} \text{ for } i=1,\dots,n. \tag{4.9}$$

Thus $(\mathcal{U}(x-ct),\mathbf{\Pi}(x-ct))$ is the unique traveling wave solution of (4.1)–(4.2) with wave speed c, which was shown to exist in section 3. Hereafter, throughout the remainder of this paper, $(\mathcal{U},\mathbf{\Pi})$ will denote such a unique solution of (4.6)–(4.9).

Now we formulate the theorem about the stability of the traveling wave solution of (4.1)–(4.2) which will be shown later.

THEOREM 3. *Let $a\in(0,1/2)$ and $(u,\mathbf{v})$ be the solution of* (4.3)–(4.5) *satisfying the following conditions:*

(1) *ϕ and ψ_i are sufficiently smooth, and $\phi'\ge0$, $\psi_i'\le0$ on $\mathbf{R}$ for $i=1,\dots,n$;*
(2) *$\sup_{z\in\mathbf{R}}|\phi(z)|+\sup_{z\in\mathbf{R}}|\phi'(z)|+\sup_{z\in\mathbf{R}}|\phi''(z)|+\sup_{z\in\mathbf{R}}|\phi'''(z)|<+\infty$;*
(3) *$a_2-\phi_2>0$, $\psi_{2i}-b_2^i>0$, $\phi_0-a_0>0$, and $b_0^i-\psi_{0i}>0$ are sufficiently small for $i=1,\dots,n$, where ϕ_2, ψ_{2i}, ϕ_0, and ψ_{0i} are defined by* (1.10);
(4) *$u_t(z,0)\ge0$ and $v_{i,t}(z,0)\le0$ for all $z\in\mathbf{R}$ and $i=1,\dots,n$;*

then there exists $z_0\in\mathbf{R}$ such that

$$\lim_{t\to+\infty}|u(z,t)-\mathcal{U}(z-z_0)|=0,\ \ \lim_{t\to+\infty}|v_i(z,t)-\Pi_i(z-z_0)|=0,\ i=1,\dots,n,$$

uniformly with respect to $z\in\mathbf{R}$.

4.2. Outline of the proof of Theorem 3. Since the proof of Theorem 3 is lengthy, we outline our proof as follows.

The plan of the proof follows Fife [4]. First, we need the following compactness lemma, whose proof can be found in the appendix.

LEMMA 4.2. *Let* $(u, \mathbf{v})$ *be the solution of* (4.3)–(4.5). *Then, under the assumption of Theorem* 3, $\{u(\cdot,t) \mid t \geq t_0 > 0\}$ *and* $\{v_i(\cdot,t) \mid t \geq t_0 > 0\}$, $i = 1, \ldots, n$, *are relatively compact, considered as subsets of* $C^2(\mathbf{R})$ *and* $C^1(\mathbf{R})$, *respectively, for each* $t_0 > 0$.

Once we have Lemma 4.2, it follows that there exist a sequence of $\{t_p\}$ with $\lim_{p\to+\infty} t_p = +\infty$ and a set of functions $(\hat{u}(z), \hat{\mathbf{v}}(z)) = (\hat{u}(z), \hat{v}_1(z), \ldots, \hat{v}_n(z)) \in C^2(\mathbf{R}) \times C^1(\mathbf{R}) \times \cdots \times C^1(\mathbf{R})$ such that

$$(u(z,t_p), \mathbf{v}(z,t_p)) \to (\hat{u}(z), \hat{\mathbf{v}}(z))$$

as $p \to +\infty$ uniformly in z, together with their associated partial derivatives. Then what is left for us to show are the following two questions.

(1) $(\hat{u}, \hat{\mathbf{v}})$ is a traveling wave front of (4.1)–(4.2) (this also implies that $(\hat{u}(\cdot), \hat{\mathbf{v}}(\cdot)) \equiv (\mathcal{U}(\cdot - z_0), \mathbf{\Pi}(\cdot - z_0))$ for some $z_0 \in \mathbf{R}$ since $(\mathcal{U}, \mathbf{\Pi})$ is unique up to translation), and

(2) the convergence of $(u, \mathbf{v})$ to $(\hat{u}, \hat{\mathbf{v}})$ is more than just along a sequence of t-values.

To answer the first question, we will make use of a Lyapunov function, which is an extension of the one in [10, 5] (see section 4.3). Regarding the second question, we will use Lemma 5.6 (local stability), which says that once u, v_i come close to $\mathcal{U}, \Pi_i$, it remains close; i.e., we have the following:

$$u(z,t) \to \mathcal{U}(z - z_0) \ \text{ and } \ v_i(z,t) \to \Pi_i(z - z_0) \ \text{ as } t \to +\infty$$

uniformly in z, together with their associated partial derivatives for $i = 1, \ldots, n$. We shall prove the first question and Lemma 5.6 in section 4.3 and the appendix, respectively.

4.3. Proof. Now we can reach the conclusion of Theorem 3. The proof will be divided into five steps. Steps 1–4 contain the associated properties of the Lyapunov function, and Step 5 will finish our treatment of the stability of the traveling wave solution. We also need two technical lemmas, i.e., Lemmas 5.4 and 5.5, whose proofs are deferred to the appendix.

Proof.

Step 1. *Construction of auxiliary functions.* First, we will truncate $(u, \mathbf{v})$ for large z and t. In fact, let $\eta > 0$ satisfy $|\kappa_j|\eta - \mu < 0$, where κ_j, μ are defined in Lemma 5.5 for $j = 1, 2$. Define the functions $\bar{u}(z,t)$ and $\bar{v}_i(z,t)$ by

$$\bar{u}(z,t) = \begin{cases} a_0 & \text{for } z \leq -\eta t - 1, \\ u(z,t) & \text{for } |z| \leq \eta t, \\ a_2 & \text{for } z \geq \eta t + 1 \end{cases}$$

and

$$\bar{v}_i(z,t) = \begin{cases} b_0^i & \text{for } z \leq -\eta t - 1, \\ v_i(z,t) & \text{for } |z| \leq \eta t, \\ b_2^i & \text{for } z \geq \eta t + 1 \end{cases}$$

for $i = 1, \dots, n$. From (5.13) and (5.14) defined in Lemma 5.5, it follows that $\bar{u}$ and $\bar{v}_i$ may be smoothed so that $\bar{u}(z,t)$ and $\bar{v}_i(z,t)$ also satisfy (5.13) and (5.14) for $i = 1, \dots, n$. Consider the function

$$M(t) = -\int_0^t \int_{-\infty}^{+\infty} H(z,s)dzds,$$

where

$$H(z,t) = \sum_{j=1}^{2} e^{\kappa_j z}\left[(D\bar{u}_{zz} + c\bar{u}_z + F(\bar{u},\bar{\mathbf{v}}))\bar{u}_t + \sum_{i=1}^{n}(c\bar{v}_{i,z} + G_i(\bar{u},\bar{\mathbf{v}}))\bar{v}_{i,t}\right]\chi_j(z)$$

for all $(z,t) \in \mathbf{R}\times\mathbf{R}^+$, $\bar{\mathbf{v}} = (\bar{v}_1, \dots, \bar{v}_n)$, χ_1 is a characteristic function on $(-\infty, 0]$, and χ_2 is a characteristic function on $[0, +\infty)$. Noting that

$$H \equiv 0 \ \text{ for } (z,t) \in \{|z| \le \eta t\}\bigcup\{|z| \ge \eta t + 1\}, \tag{4.10}$$

it follows that $\int_{-\infty}^{+\infty} H(z,s)dz$ converges for all $s \ge 0$. Therefore, $M(t)$ is well defined for all $t \ge 0$. From now on, combining with Lemmas 5.4 and 5.5, we can follow a similar calculation of [10, 5] to complete the proof.

Step 2. *We claim that* $|M(t)|$ *is bounded independently of* t. Indeed, noting that $\bar{u}$, $\bar{v}_i$, a_j, and b_j^i are uniformly bounded for $i = 1, \dots, n$, and $j = 0, 2$, and that ∇F and $\nabla G_i, i = 1, \dots, n$, are continuous, it follows that there exists a constant $C > 0$ such that

$$\begin{aligned} |F(\bar{u},\bar{\mathbf{v}})| &= |F(\bar{u},\bar{\mathbf{v}}) - F(a_j,\mathbf{b_j})| \le C\max_{1\le l\le n}\{|\bar{u}-a_j|, |\bar{v}_l - b_j^l|\},\\ |G_i(\bar{u},\bar{\mathbf{v}})| &= |G_i(\bar{u},\bar{\mathbf{v}}) - G_i(a_j,\mathbf{b_j})| \le C\max_{1\le l\le n}\{|\bar{u}-a_j|, |\bar{v}_l - b_j^l|\} \end{aligned} \tag{4.11}$$

on $\mathbf{R}\times[0,+\infty)$ for $i = 1, \dots, n$ and $j = 0, 2$. Throughout the proof, C always denotes a constant, which may be different from sentence to sentence, but they depend only on D, k_+^i, k_-^i, b_0^i, κ_i, σ_i, μ, and η. From these inequalities, the definitions of $\bar{u}$ and $\bar{\mathbf{v}}$, (4.10), and (5.13)–(5.14), it follows that there exists a positive constant C such that the following inequality holds:

$$\begin{aligned} |M(t)| \le\ & \left[\int_0^t\int_{-\eta s}^{0} e^{\kappa_1 z}(u_t^2(z,s) + \sum_{i=1}^{n} v_{i,t}^2(z,s))dzds\right.\\ &\left. +\ \int_0^t\int_0^{\eta s} e^{\kappa_2 z}(u_t^2(z,s) + \sum_{i=1}^{n} v_{i,t}^2(z,s))dzds\right]\\ &+\ \left[C\int_0^t\int_{-\eta s-1}^{-\eta s}(e^{2\sigma_1 z} + e^{\kappa_1 z - 2\mu t} + 2e^{((\kappa_1/2)+\sigma_1)z-\mu t})dzds\right.\\ &\left. +\ C\int_0^t\int_{\eta s}^{\eta s+1}(e^{-2\sigma_2 z} + e^{\kappa_2 z - 2\mu t} + 2e^{((\kappa_2/2)-\sigma_2)z-\mu t})dzds\right]\\ :=\ & I(t) + II(t). \end{aligned}$$

Noting that $|\kappa_j|\eta - 2\mu = (|\kappa_j|\eta - \mu) - \mu < 0$, $\kappa_j/2 + \sigma_j > 0$, and $\kappa_j/2 - \sigma_j < 0$, $j = 1, 2$, then it follows that $II(t)$ is bounded by some positive constant C. Recalling that $u_t(\cdot, 0) \ge 0$ and $v_{i,t}(\cdot, 0) \le 0$ on $\mathbf{R}$, it follows from Lemma 5.4 that $u_t(z,t) \ge 0$

and $v_{i,t}(z,t) \le 0$ for all $(z,t) \in \mathbf{R} \times \mathbf{R}^+$. Using this fact and Fubini's theorem, we can estimate $I(t)$ as follows:

$$\begin{aligned} I(t) \quad &\le 2C_1 \int_0^t \int_{-\eta s}^0 e^{\kappa_1 z} \left(u_t(z,s) - \sum_{i=1}^n v_{i,t}(z,s) \right) dz ds \\ &\quad + 2C_1 \int_0^t \int_0^{\eta s} e^{\kappa_2 z} \left(u_t(z,s) - \sum_{i=1}^n v_{i,t}(z,s) \right) dz ds \\ &= 2C_1 \left[\int_{-\eta t}^0 e^{\kappa_1 z} \left(\int_{-z/\eta}^t (u_t(z,s) - \sum_{i=1}^n v_{i,t}(z,s)) ds \right) dz \right] \\ &\quad + 2C_1 \left[\int_0^{\eta t} e^{\kappa_2 z} \left(\int_{z/\eta}^t (u_t(z,s) - \sum_{i=1}^n v_{i,t}(z,s)) ds \right) dz \right] \\ &= 2C_1 \left[\int_{-\eta t}^0 e^{\kappa_1 z} \left(u(z,t) - \sum_{i=1}^n v_i(z,t) - u(z,-z/\eta) + \sum_{i=1}^n v_i(z,-z/\eta) \right) dz \right] \\ &\quad + 2C_1 \left[\int_0^{\eta t} e^{\kappa_2 z} \left(u(z,t) - \sum_{i=1}^n v_i(z,t) - u(z,z/\eta) + \sum_{i=1}^n v_i(z,z/\eta) \right) dz \right], \end{aligned}$$

where C_1 is defined in Lemma 5.5. Noting that $\kappa_j < 0$, $\kappa_j/2 + \sigma_j > 0$, $\kappa_j/2 - \sigma_j < 0$, and $|\kappa_j|\eta - \mu < 0$ for $j = 1, 2$, and using the above inequality, (5.13)–(5.14), we obtain that $I(t)$ is also bounded by some positive constant C. Since $I(t)$, $II(t)$ are bounded for all $t \ge 0$, it follows that $|M(t)| \le C$ for all $t \ge 0$ and some constant $C > 0$. This proves our claim.

Step 3. *We claim that* $\lim_{t \to +\infty} |\dot{M(t)} + Q[\bar{u}, \bar{\mathbf{v}}](t)| = 0$, *where*

$$Q[\bar{u}, \bar{\mathbf{v}}](t) = \sum_{j=1}^2 \int_{-\infty}^{+\infty} e^{\kappa_j z} \left[(D\bar{u}_{zz} + c\bar{u}_z + F(\bar{u}, \bar{\mathbf{v}}))^2 + \sum_{i=1}^n (c\bar{v}_{i,z} + G_i(\bar{u}, \bar{\mathbf{v}}))^2 \right] \chi_j(z) dz.$$

Indeed, differentiating $M(t)$, we obtain

$$\dot{M(t)} = -\sum_{j=1}^2 \int_{-\infty}^{+\infty} e^{\kappa_j z} \left[(D\bar{u}_{zz} + c\bar{u}_z + F(\bar{u}, \bar{\mathbf{v}}))\bar{u}_t + \sum_{i=1}^n (c\bar{v}_{i,z} + G_i(\bar{u}, \bar{\mathbf{v}}))\bar{v}_{i,t} \right] \chi_j(z) dz,$$

and so we have

$$\begin{aligned} \dot{M(t)} + Q[\bar{u}, \bar{\mathbf{v}}](t) = &- \int_{-\infty}^0 e^{\kappa_1 z} [(D\bar{u}_{zz} + c\bar{u}_z + F(\bar{u}, \bar{\mathbf{v}}))(L_1[\bar{u}, \bar{\mathbf{v}}] - F(\bar{u}, \bar{\mathbf{v}})) \\ &\qquad + \sum_{i=1}^n (c\bar{v}_{i,z} + G_i(\bar{u}, \bar{\mathbf{v}}))(L_{2i}[\bar{u}, \bar{\mathbf{v}}] - G(\bar{u}, \bar{\mathbf{v}}))] dz \\ &- \int_0^{+\infty} e^{\kappa_2 z} [(D\bar{u}_{zz} + c\bar{u}_z + F(\bar{u}, \bar{\mathbf{v}}))(L_1[\bar{u}, \bar{\mathbf{v}}] - F(\bar{u}, \bar{\mathbf{v}})) \\ &\qquad + \sum_{i=1}^n (c\bar{v}_{i,z} + G_i(\bar{u}, \bar{\mathbf{v}}))(L_{2i}[\bar{u}, \bar{\mathbf{v}}] - G(\bar{u}, \bar{\mathbf{v}}))] dz. \end{aligned}$$

Noting that $L_1[\bar{u}, \bar{\mathbf{v}}] - F(\bar{u}, \bar{\mathbf{v}}) \equiv L_{2i}[\bar{u}, \bar{\mathbf{v}}] - G(\bar{u}, \bar{\mathbf{v}}) \equiv 0$, $i = 1, \ldots, n$, for $(z,t) \in \{|z| \le \eta t\} \bigcup \{|z| \ge \eta t + 1\}$, and using (4.11) and (5.13)–(5.14), we obtain that there

exists a positive constant C such that

$$\begin{aligned}|\dot{M(t)} + Q[\bar{u},\bar{\mathbf{v}}](t)| &\le C\int_{-\eta t-1}^{-\eta t}(e^{2\sigma_1 z} + e^{\kappa_1 z - 2\mu t} + e^{(\kappa_1/2)z+\sigma_1 z-\mu t})dz \\ &\quad + C\int_{\eta t}^{\eta t+1}(e^{-2\sigma_2 z} + e^{\kappa_2 z - 2\mu t} + e^{(\kappa_2/2)z-\sigma_2 z-\mu t})dz \\ &\le C\left(\frac{e^{-2\sigma_1\eta t}}{2\sigma_1} - \frac{e^{-\kappa_1(\eta t+1)-2\mu t}}{\kappa_1} + \frac{e^{-(\kappa_1/2+\sigma_1)\eta t-\mu t}}{((\kappa_1/2)+\sigma_1)}\right) \\ &\quad + C\left(\frac{e^{-2\sigma_2\eta t}}{2\sigma_2} - \frac{e^{\kappa_2\eta t-2\mu t}}{\kappa_2} + \frac{e^{(\kappa_2/2-\sigma_2)\eta t-\mu t}}{(-(\kappa_2/2)+\sigma_2)}\right).\end{aligned}$$

Using this inequality and the fact that $|\kappa_j|\eta - 2\mu < 0$, $\kappa_j/2+\sigma_j > 0$, and $\kappa_j/2-\sigma_j < 0$ for $j = 1, 2$, it follows that the claim holds.

Step 4. *We claim that there exists a sequence* $\{t_p\}_{p\in\mathbf{N}}$ *with* $\lim_{p\to\infty} t_p = +\infty$ *such that*

$$\lim_{p\to\infty} Q[\bar{u},\bar{\mathbf{v}}](t_p) = 0. \tag{4.12}$$

Indeed, noting that $Q[\bar{u},\bar{\mathbf{v}}](t) \ge 0$ for all $t \ge 0$ and that $|M(t)|$ is bounded, it follows that we have $\lim_{t\to+\infty} \sup \dot{M(t)} = 0$. Combining this with Step 3, we obtain (4.12).

Step 5. *We reach our conclusion.* Indeed, Lemma 4.2 implies that there exists a subsequence of $\{t_p\}_{p\in\mathbf{N}}$, say $\{t'_p\}_{p\in\mathbf{N}}$, along which $\bar{u}(\cdot,t'_p)$ and $\bar{v}_i(\cdot,t'_p)$ converge to the limit functions $\hat{u}(z)$ and $\hat{v}_i(z)$ in $C^2(\mathbf{R})$ and $C^1(\mathbf{R})$, respectively, for some functions $\hat{u} \in C^2(\mathbf{R})$ and $\hat{v}_i \in C^1(\mathbf{R})$ and $i = 1,\dots,n$. Set $\hat{\mathbf{v}} = (\hat{v}_1,\dots,\hat{v}_n)$. Using this and (4.12) it follows that for each finite interval $I = [-l, l]$ with $l > 0$, we have

$$\begin{aligned}&\left[\sum_{j=1}^{2}\int_I e^{\kappa_j z}((D\bar{u}_{zz} + c\bar{u}_z + F(\bar{u},\bar{\mathbf{v}}))^2 + \sum_{i=1}^{n}(c\bar{v}_{i,z} + G_i(\bar{u},\bar{\mathbf{v}}))^2)\chi_j(z)dz\right]_{t=t'_p} \\ &\to \left[\sum_{j=1}^{2}\int_I e^{\kappa_j z}((D\hat{u}_{zz} + c\hat{u}_z + F(\hat{u},\hat{\mathbf{v}}))^2 + \sum_{i=1}^{n}(c\hat{v}_{i,z} + G_i(\hat{u},\hat{\mathbf{v}}))^2)\chi_j(z)dz\right] = 0\end{aligned}$$

as $p \to +\infty$. Thus

$$D\hat{u}_{zz} + c\hat{u}_z + F(\hat{u},\hat{\mathbf{v}}) = 0, \quad c\hat{v}_{i,z} + G_i(\hat{u},\hat{\mathbf{v}}) = 0 \ \text{ on } \ \mathbf{R}$$

for $i = 1,\dots,n$. Note that $\hat{u}(-\infty) = a_0$, $\hat{u}(+\infty) = a_2$, $\hat{v}_i(-\infty) = b_0^i$, and $\hat{v}_i(+\infty) = b_2^i$ for $i = 1,\dots,n$. Thus, from the uniqueness of the traveling wave front solution $(\mathcal{U},\mathbf{\Pi})$ of (4.1)–(4.2), it follows that $\hat{u} = \mathcal{U}(z - z_0)$ and $\hat{v}_i = \Pi_i(z - z_0)$, $i = 1,\dots,n$, for some $z_0 \in \mathbf{R}$. Noting that $(u,\mathbf{v}) \equiv (\bar{u},\bar{\mathbf{v}})$ for $|z| \le \eta t$, we obtain that

$$u(z,t'_p) \to \mathcal{U}(z - z_0) \ \text{ and } \ v_i(z,t'_p) \to \Pi_i(z - z_0),\ i = 1,\dots,n,$$

in $C^2(\mathbf{R})$ and $C^1(\mathbf{R})$, respectively, as $p \to \infty$. Finally, an application of Lemma 5.6 with $(\phi(z),\psi_i(z)) = (u(z,t'_p), v_i(z,t'_p))$, $i = 1,\dots,n$, completes the proof. □

5. Appendix.

5.1. The zeros of (3.11). In the following lemma, we will count the positive zero and negative zeros of $p_n(\lambda)$ defined by (3.11).

LEMMA 5.1. *There are only one positive zero and $n+1$ negative zeros for the polynomial $p_n(\lambda)$ which is defined by* (3.11).

Proof. Recall that $p_n(\lambda)$ is a polynomial of degree $n+2$ and is defined by

$$\begin{aligned} p_n(\lambda) &:= p_n(\lambda; k_+^1, \dots, k_+^n, k_-^1, \dots, k_-^n, b_0^1, \dots, b_0^n) \\ &= \det \begin{bmatrix} -\lambda & 1 & 0 & \cdots & 0 \\ -\theta(-a - \sum_{i=1}^n k_+^i b_0^i) & \theta - \lambda & \theta k_-^1 & \cdots & \theta k_-^n \\ -k_+^1 b_0^1 & 0 & -k_-^1 - \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -k_+^n b_0^n & 0 & 0 & \cdots & -k_-^n - \lambda \end{bmatrix}. \end{aligned}$$

Then, by a careful calculation, we have

$$p_n(\lambda) = (\lambda^2 - \theta\lambda - a\theta) \prod_{i=1}^n (-\lambda - k_-^i) + \sum_{i=1}^n \left[\theta(k_+^i b_0^i) \lambda \prod_{j \neq i}^n (-\lambda - k_-^j) \right]. \tag{5.1}$$

First, we claim that if the off rates $\{k_-^i\}_{i=1}^n$ are different and $n > 1$, then the conclusion of this lemma holds. Indeed, by (5.1), we have the following equalities:

$$\begin{aligned} p_n(0) &= (-1)^{n+1} \theta a \prod_{i=1}^n k_-^i, \\ p_n(-k_-^i) &= -\theta b_0^i k_+^i k_-^i \prod_{j \neq i}^n (k_-^i - k_-^j), \ i = 1, \dots, n. \end{aligned}$$

Set $k_-^0 = 0$. Then, by assumption, we may assume that $0 < k_-^1 < \cdots < k_-^n$, and so we can conclude that $p_n(-k_-^i) p_n(-k_-^{i+1}) < 0$ for $i = 0, \dots, n-1$. Hence we complete the proof of this claim.

Now we turn to the general case; i.e., the off rates $\{k_-^i\}_{i=1}^n$ are not necessarily different. We will use induction to prove it. When $n = 1$, i.e., there is only one buffer, then we observe that $p_1(-\infty) > 0$, $p_1(-k_-^1) = -\theta k_+^1 k_-^1 b_0^1 < 0$, $p_1(0) = \theta a k_-^1 > 0$, and $p_1(+\infty) < 0$, and so we obtain that the conclusion holds for $n = 1$.

Assume that there are only one positive zero and $m+1$ negative zeros for $p_m(\lambda)$ (some zeros of $p_m(\lambda)$ may have multiplicity bigger than 1) for any positive $k_+^1, \dots, k_+^m$, $k_-^1, \dots, k_-^m$, and $b_0^1, \dots, b_0^m$.

Now we consider the case for $n = m+1$. If the off rates $\{k_-^i\}_{i=1}^{m+1}$ are different, then from the above claim it follows that there are only one positive zero and $m+2$ negative zeros for $p_{m+1}(\lambda)$. If the off rates $\{k_-^i\}_{i=1}^{m+1}$ are not necessarily different, then we may assume that $k_-^m = k_-^{m+1}$, and so from (5.1) it follows that

$$\begin{aligned} & p_{m+1}(\lambda) \\ &= p_{m+1}(\lambda; k_+^1, \dots, k_+^{m+1}, k_-^1, \dots, k_-^{m+1}, b_0^1, \dots, b_0^{m+1}) \\ &= (-\lambda - k_-^m) p_m(\lambda; k_+^1, \dots, k_+^m, k_-^1, \dots, k_-^m, b_0^1, \dots, b_0^{m-1}, b_0^m + (k_+^{m+1} b_0^{m+1} / k_+^m)) \\ &= (-\lambda - k_-^m) q_m(\lambda). \end{aligned}$$

By induction hypothesis, $q_m(\lambda)$ has only one positive zero and $m+1$ negative zeros (some of them may have multiplicity bigger than 1). Note that k_-^m is positive. Therefore, $p_{m+1}(\lambda)$ has only one positive zero and $m+2$ negative zeros. The completes the induction, and so the proof is completed. ∎

5.2. Proof of Lemma 4.2. *Plan of the proof of Lemma* 4.2.

The proof of Lemma 4.2 is based on the estimation of the L^∞ norm of $u(\cdot,t)$, $v_i(\cdot,t)$ and their associated partial derivatives. Roughly speaking, we use the invariance principle to show that our solution $(u,\mathbf{v})$ of (4.3)–(4.5) is bounded by the traveling wave solution $(\mathcal{U},\mathbf{\Pi})$ (see Lemma 5.3). Then we use the assumptions of Theorem 3, the L^p-regularity theorem, and Schauder estimates to estimate the partial derivatives of u and v_i for $i=1,\dots,n$ (see Lemma 5.5). Once Lemmas 5.3 and 5.5 are proved, then we can use standard arguments to obtain Lemma 4.2 (see the proof of Lemma 4.6 in Klaasen and Troy [10] for details).

Now we turn to our proof. First, for each $\delta>0$, we define the following two sets:

$$R_\delta^0=\{(u,\mathbf{v})\mid |u-a_0|\le\delta, |v_i-b_0^i|\le\delta \ \text{ for } \ i=1,\dots,n\} \tag{5.2}$$

and

$$R_\delta^2=\{(u,\mathbf{v})\mid |u-a_2|\le\delta, |v_i-b_2^i|\le\delta \ \text{ for } \ i=1,\dots,n\}. \tag{5.3}$$

Then these two sets have the following property.

LEMMA 5.2. *If $a\in(0,1)$, then there exist $\delta_1>0$ and constants α_{11}^i, $\alpha_{11,j}^i$, $\alpha_{12,j}^i$, $\alpha_{21,j}^i$, $\alpha_{22,j}^i$, $i=0,2$, $j=1,\dots,n$, such that for all $\delta\in(0,\delta_1)$, $i=0,2$, and $j=1,\dots,n$, we have*

(1) *$F_u<\alpha_{11}^i=\sum_{j=1}^n\alpha_{11,j}^i<0$, $\alpha_{12,j}^i<F_{v_j}<0$, $\alpha_{21,j}^i<G_{j,u}<0$, and $G_{j,v_j}<\alpha_{22,j}^i<0$ on R_δ^i, where $\alpha_{11,j}^i<0$;*

(2) *$\alpha_{11,j}^i\alpha_{22,j}^i-\alpha_{12,j}^i\alpha_{21,j}^i>0$;*

(3) *if both $(u,\mathbf{v})$ and $(u-q_1,\mathbf{v}+\mathbf{q_2})$ belong to R_δ^i for some $i\in\{0,2\}$, and $q_1>0$, $q_{2j}>0$, $j=1,\dots,n$, then*

$$F(u,\mathbf{v})-F(u-q_1,\mathbf{v}+\mathbf{q_2})\le\alpha_{11}^iq_1-\sum_{j=1}^n\alpha_{12,j}^iq_{2j},$$
$$G_j(u,\mathbf{v})-G_j(u-q_1,\mathbf{v}+\mathbf{q_2})\ge\alpha_{21,j}^iq_1-\alpha_{22,j}^iq_{2j}$$

for $j=1,\dots,n$, where $\mathbf{q_2}=(q_{21},\dots,q_{2n})$.

Proof. For simplicity, we set $K_i=k_-^i/k_+^i$, $i=1,\dots,n$.

(1) By a simple calculation, the quantities

$$F_u(a_0,\mathbf{b_0}),F_{v_j}(a_0,\mathbf{b_0}),F_u(a_2,\mathbf{b_2}),F_{v_j}(a_2,\mathbf{b_2}),$$
$$G_{j,u}(a_0,\mathbf{b_0}),G_{j,v_j}(a_0,\mathbf{b_0}),G_{j,u}(a_2,\mathbf{b_2}),G_{j,v_j}(a_2,\mathbf{b_2})$$

are negative if $a\in(0,1)$. From this it follows that there exist $\alpha_{11}^i<0$, $\alpha_{11,j}^i<0$, $\alpha_{12,j}^i<0$, $\alpha_{21,j}^i<0$, and $\alpha_{22,j}^i<0$, $i=0,2$, $j=1,\dots,n$, such that part (1) of the conclusion is true.

(2) Moreover, noting that

$$F_u(a_0,\mathbf{b_0})=-\sum_{j=1}^n(a/n+k_+^jb_0^j),$$
$$F_u(a_2,\mathbf{b_2})=-\sum_{j=1}^n[(1-a)/n+k_-^jb_0^j/(K_j+1)],$$

the identities

$$(-a/n-k_+^jb_0^j)G_{j,v_j}(a_0,\mathbf{b_0})-F_{v_j}(a_0,\mathbf{b_0})G_{j,u}(a_0,\mathbf{b_0})=ak_-^j/n>0,$$

and

$$[-(1-a)/n - k_-^j b_0^j/(K_j+1)]G_{j,v_j}(a_2,\mathbf{b_2}) - F_{v_j}(a_2,\mathbf{b_2})G_{j,u}(a_2,\mathbf{b_2})$$
$$= (1-a)(k_+^j + k_-^j)/n > 0$$

if $a \in (0,1)$, it follows that we can choose $\alpha_{11}^i < 0$, $\alpha_{11,j}^i < 0$, $\alpha_{12,j}^i < 0$, $\alpha_{21,j}^i < 0$, and $\alpha_{22,j}^i < 0$ such that part (1) of the conclusion and

$$\alpha_{11}^i = \sum_{j=1}^n \alpha_{11,j}^i, \quad \alpha_{11,j}^i \alpha_{22,j}^i - \alpha_{12,j}^i \alpha_{21,j}^i > 0$$

for $i = 0, 2$, and $j = 1, \ldots, n$ hold.

(3) The proof follows from the mean value theorem and part (1). □

Next we estimate the bound of the L^∞ norm of the solution $(u(\cdot,t), \mathbf{v}(\cdot,t))$ of (4.3)–(4.5). We will adapt the method of [10] to prove it.

LEMMA 5.3. *Let $(u, \mathbf{v})$ be the solution of* (4.3)–(4.5). *If $a_2 - \phi_2 > 0$, $\psi_{2i} - b_2^i > 0$, $\phi_0 - a_0 > 0$, and $b_0^i - \psi_{0i} > 0$ are sufficiently small, then there exist constants z_1, $z_2 \in \mathbf{R}$, $\hat{k}_1 > 0$, $k_{2i} > 0$, $i = 1, \ldots, n$, and $\mu > 0$ such that*

$$\mathcal{U}(z - z_1) - \hat{k}_1 e^{-\mu t} \le u(z,t) \le \mathcal{U}(z - z_2) + \hat{k}_1 e^{-\mu t}, \tag{5.4}$$

and

$$\Pi_i(z - z_2) - k_{2i} e^{-\mu t} \le v_i(z,t) \le \Pi_i(z - z_1) + k_{2i} e^{-\mu t} \tag{5.5}$$

for all $(z,t) \in \mathbf{R} \times [0, +\infty)$ and $i = 1, \ldots, n$.

Proof. We prove the left-hand side of (5.4) and the right-hand side of (5.5), since the remaining inequalities follow in a similar way.

The idea is to use the invariance principle to choose suitable functions $\epsilon(t)$, $q_1(t)$, and $\mathbf{q_2}(t)$ such that $\mathcal{U}(z - \epsilon(t)) - q_1(t) \le u(z,t)$ and $\mathbf{\Pi}(z - \epsilon(t)) + \mathbf{q_2}(t) \ge \mathbf{v}(z,t)$ for all $(z,t) \in \mathbf{R} \times \mathbf{R}^+$ and then use the monotone properties of $\mathcal{U}$ and $\mathbf{\Pi}$ to obtain our conclusion. The proof will be divided into three steps.

Step 1. *Construction of $\tilde{u}$ and $\tilde{\mathbf{v}}$.* Recall that the functions $u(z,t)$ and $\mathbf{v}(z,t)$ satisfy

$$L_1[u, \mathbf{v}] \equiv u_t - D u_{zz} - c u_z = F(u, \mathbf{v}),$$
$$L_{2i}[u, \mathbf{v}] \equiv v_{i,t} - c v_{i,z} = G_i(u, \mathbf{v}),\ i = 1, \ldots, n,$$

and

$$u(z,0) = \phi(z), \quad v_i(z,0) = \psi_i(z)$$

for all $z \in \mathbf{R}$ and $i = 1, \ldots, n$. Let $(\underline{u}, \underline{\mathbf{v}}) = (u, \underline{v}_1, \ldots, \underline{v}_n)$ be defined by

$$\underline{u}(z,t) \equiv \mathcal{U}(z - \epsilon(t)) - q_1(t)$$

and

$$\underline{v}_i(z,t) \equiv \Pi_i(z - \epsilon(t)) + q_{2i}(t),\ i = 1, \ldots, n, \quad \text{on } \mathbf{R} \times [0, +\infty),$$

where $\epsilon(t)$, $q_1(t)$, and $q_{2i}(t)$ are positive functions with the properties that

$$\underline{u}(z,t) \ge -\min_{j=1,\ldots,n} \{k_-^j/(2k_+^j)\}, \quad \underline{v}_i(z,t) \ge b_2^i/2 \tag{5.6}$$

for all $(z,t) \in \mathbf{R} \times [0,+\infty)$ and $i = 1,\dots,n$. We remark that $(\underline{u}, \underline{\mathbf{v}})$ will be a subsolution of (4.3)–(4.4). Now we will find the differential equations for which $(\underline{u}, \underline{\mathbf{v}})$ satisfy. Setting $\tau = z - \epsilon(t)$ and using the fact that $D\mathcal{U}'' + c\mathcal{U}' + F(\mathcal{U}, \mathbf{\Pi}) = 0$ and $c\Pi_i' + G_i(\mathcal{U}, \mathbf{\Pi}) = 0$, $i = 1,\dots,n$, we find that for each $i = 1,\dots,n$,

$$\begin{aligned} L_1[\underline{u}, \underline{\mathbf{v}}] &= -\epsilon'(t)\mathcal{U}'(\tau) - q_1'(t) - D\mathcal{U}''(\tau) - c\mathcal{U}'(\tau) \\ &= -\epsilon'(t)\mathcal{U}'(\tau) - q_1'(t) + F(\mathcal{U}(\tau), \mathbf{\Pi}(\tau)) \end{aligned}$$

and

$$\begin{aligned} L_{2i}[\underline{u}, \underline{\mathbf{v}}] &= -\epsilon'(t)\Pi_i'(\tau) + q_{2i}'(t) - c\Pi_i'(\tau) \\ &= -\epsilon'(t)\Pi_i'(\tau) + q_{2i}'(t) + G_i(\mathcal{U}(\tau), \mathbf{\Pi}(\tau)) \end{aligned}$$

for all $\tau \in \mathbf{R}$.

Set $(\tilde{u}, \tilde{\mathbf{v}}) = (u - \underline{u}, \mathbf{v} - \underline{\mathbf{v}})$. Then the variables $(\tilde{u}, \tilde{\mathbf{v}})$ satisfy the following system:

$$\begin{aligned} L_1[\tilde{u}, \tilde{\mathbf{v}}] &= F(u, \mathbf{v}) + \epsilon'(t)\mathcal{U}'(\tau) + q_1'(t) - F(\mathcal{U}(\tau), \mathbf{\Pi}(\tau)) \\ &= [F(u, \mathbf{v}) - F(\underline{u}, \underline{\mathbf{v}})] + [\epsilon'(t)\mathcal{U}'(\tau) + q_1'(t) + F(\underline{u}, \underline{\mathbf{v}}) - F(\mathcal{U}(\tau), \mathbf{\Pi}(\tau))] \\ &\equiv \left[f_1(z,t)\tilde{u} + \sum_{j=1}^{n} f_{1j}(z,t)\tilde{v}_j \right] + N_1(z,t), \end{aligned} \tag{5.7}$$

$$\begin{aligned} L_{2i}[\tilde{u}, \tilde{\mathbf{v}}] &= G_i(u, \mathbf{v}) + \epsilon'(t)\Pi_i'(\tau) - q_{2i}'(t) - G_i(\mathcal{U}(\tau), \mathbf{\Pi}(\tau)) \\ &= [G_i(u, \mathbf{v}) - G_i(\underline{u}, \underline{\mathbf{v}})] + [\epsilon'(t)\Pi_i'(\tau) - q_{2i}'(t) + G_i(\underline{u}, \underline{\mathbf{v}}) - G_i(\mathcal{U}(\tau), \mathbf{\Pi}(\tau))] \\ &\equiv [g_{2i}(z,t)\tilde{u} + \tilde{g}_{2i}(z,t)\tilde{v}_i] + N_{2i}(z,t), \end{aligned} \tag{5.8}$$

together with the initial data

$$\begin{aligned} \tilde{u}(z,0) &= u(z,0) - \underline{u}(z,0), \\ \tilde{v}_i(z,0) &= v_i(z,0) - \underline{v}_i(z,0), \end{aligned}$$

where

$$\begin{aligned} f_1(z,t) &= F_u(\theta_1 u + (1-\theta_1)\underline{u}, \theta_1 \mathbf{v} + (1-\theta_1)\underline{\mathbf{v}})(z,t), \\ f_{1j}(z,t) &= F_{v_j}(\theta_1 u + (1-\theta_1)\underline{u}, \theta_1 \mathbf{v} + (1-\theta_1)\underline{\mathbf{v}})(z,t), \\ g_{2i}(z,t) &= G_{i,u}(\theta_{2i} u + (1-\theta_{2i})\underline{u}, \theta_{2i} \mathbf{v} + (1-\theta_{2i})\underline{\mathbf{v}})(z,t), \\ \tilde{g}_{2i}(z,t) &= G_{i,v_i}(\theta_{2i} u + (1-\theta_{2i})\underline{u}, \theta_{2i} \mathbf{v} + (1-\theta_{2i})\underline{\mathbf{v}})(z,t), \end{aligned}$$

for some $\theta_1 = \theta_1(u, \mathbf{v}, \underline{u}, \underline{\mathbf{v}}) \in (0,1)$, $\theta_{2i} = \theta_{2i}(u, \mathbf{v}, \underline{u}, \underline{\mathbf{v}}) \in (0,1)$, and $i, j = 1,\dots,n$.

Step 2. *We claim that for suitably chosen $\epsilon(t)$, $q_1(t)$, and $q_{2i}(t)$, $i = 1,\dots,n$, the region $\{\tilde{u} \geq 0, \tilde{v}_{2i} \leq 0, i = 1,\dots,n\}$ is invariant under the flow* (5.7)–(5.8). Indeed, from (5.6), Lemma 4.1, and the definitions of F and G_i, it follows that $f_{1i}(z,t) < 0$ and $g_{2i}(z,t) < 0$ for all $(z,t) \in \mathbf{R} \times [0,+\infty)$ and $i = 1,\dots,n$. Therefore, if we can choose $\epsilon(t)$, $q_1(t)$, and $q_{2i}(t)$ satisfying that (5.6) holds and that $N_1(z,t) \geq 0$ and $N_{2i}(z,t) \leq 0$ for all $(z,t) \in \mathbf{R} \times [0,+\infty)$ and $i = 1,\dots,n$, then, by Theorem 14.11 on p. 203 of Smoller [23] we have $\tilde{u}(z,t) \geq 0$ and $\tilde{\mathbf{v}}(z,t) \leq \mathbf{0} = (0,\dots,0)$ for all $(z,t) \in \mathbf{R} \times [0,+\infty)$ if $\tilde{u}(\cdot,0) \geq 0$ and $\tilde{\mathbf{v}}(\cdot,0) \leq \mathbf{0}$ on $\mathbf{R}$.

Next we try to find the desired functions $\epsilon(t)$, $q_1(t)$, and $q_{2i}(t)$, $i = 1,\dots,n$. Indeed, let $\delta \in (0, \delta_1)$ with δ_1 being defined in Lemma 5.2. Also recall the definitions of $R_\delta^i, i = 0, 2$, given by (5.2)–(5.3) and note that $\mathcal{U}' > 0$ and $\Pi_j' < 0$ on $\mathbf{R}$ for $j = 1,\dots,n$. Set $\mathbf{q_2}(t) = (q_{21}(t),\dots,q_{2n}(t))$ and assume $\epsilon' \geq 0$. If for some $i \in \{0, 2\}$

both $(\mathcal{U}(\tau), \mathbf{\Pi}(\tau))$ and $(\mathcal{U}(\tau) - q_1(t), \mathbf{\Pi}(\tau) + \mathbf{q_2}(t))$ belong to R^i_δ, then part (3) of Lemma 5.2 implies

$$\begin{aligned} N_1(z,t) &\geq q_1' - \alpha^i_{11} q_1 + \sum_{l=1}^n \alpha^i_{12,l} q_{2l}, \\ N_{2j}(z,t) &\leq -q_{2j}' - \alpha^i_{21,j} q_1 + \alpha^i_{22,j} q_{2j} \end{aligned} \tag{5.9}$$

for $j = 1, \ldots, n$. On the other hand, from Lemma 5.2 it follows that there exist positive constants μ_0, k_{0j}, $j = 1, \ldots, n$, such that for all $\mu \in (0, \mu_0)$, $i = 0, 2$, and $j = 1, \ldots, n$, we have

$$\frac{\mu}{n} + \alpha^i_{11,j} < 0,\ \mu + \alpha^i_{22,j} < 0,\ \left(\frac{\mu}{n} + \alpha^i_{11,j}\right)(\mu + \alpha^i_{22,j}) > \alpha^i_{12,j}\alpha^i_{21,j},$$

and

$$\frac{\alpha^i_{21,j}}{\mu + \alpha^i_{22,j}} < k_{0j} < \frac{(\mu/n) + \alpha^i_{11,j}}{\alpha^i_{12,j}}.$$

This implies that

$$\begin{aligned} -\mu/n - \alpha^i_{11,j} + k_{0j}\alpha^i_{12,j} > 0, \\ \mu k_{0j} - \alpha^i_{21,j} + k_{0j}\alpha^i_{22,j} < 0 \end{aligned}$$

for all $\mu \in (0, \mu_0)$, $i = 0, 2$, and $j = 1, \ldots, n$. Summing the first inequality from $j = 1$ to n, we obtain that

$$-\mu - \alpha^i_{11} + \sum_{j=1}^n k_{0j}\alpha^i_{12,j} > 0.$$

Hence for all $\mu \in (0, \mu_0)$ and $d > 0$, the functions q_1 and q_{2j} defined by

$$q_1(t) = de^{-\mu t},\ \ q_{2j}(t) = dk_{0j}e^{-\mu t},\ j = 1, \ldots, n, \tag{5.10}$$

satisfy

$$q_1' - \alpha^i_{11} q_1 + \sum_{l=1}^n \alpha^i_{12,l} q_{2l} > 0,\ \ -q_{2j}' - \alpha^i_{21,j} q_1 + \alpha^i_{22,j} q_{2j} < 0 \tag{5.11}$$

for all $t \geq 0$ and $j = 1, \ldots, n$. Therefore, by the above discussion, (5.9), and (5.11), we can choose a sufficiently small $d_0 > 0$ and a sufficiently large $M > 0$ such that if $d \in (0, d_0)$ and $|\tau| > M$, then both $(\mathcal{U}(\tau), \mathbf{\Pi}(\tau))$ and $(\mathcal{U}(\tau) - q_1(t), \mathbf{\Pi}(\tau) + \mathbf{q_2}(t))$ belong to one of R^0_δ and R^2_δ, and $N_1(z,t) > 0$, $N_{2i}(z,t) < 0$ for $i = 1, \ldots, n$.

We focus our attention on the intermediate case: $|\tau| \leq M$. Since $\mathcal{U}' > 0$ and $\Pi_i' < 0$ on $\mathbf{R}$, there exist positive constants β, $\bar\gamma_1$, and $\bar\gamma_{2i}$ such that

$$\begin{aligned} &\mathcal{U}'(\tau) > \beta,\ \ \Pi_i'(\tau) < -\beta, \\ &F(\mathcal{U}(\tau) - q_1(t), \mathbf{\Pi}(\tau) + \mathbf{q_2}(t)) - F(\mathcal{U}(\tau), \mathbf{\Pi}(\tau)) \geq -\bar\gamma_1 q_1(t) - \sum_{l=1}^n \bar\gamma_{2l} q_{2l}(t), \\ &G_i(\mathcal{U}(\tau) - q_1(t), \mathbf{\Pi}(\tau) + \mathbf{q_2}(t)) - G_i(\mathcal{U}(\tau), \mathbf{\Pi}(\tau)) \leq \bar\gamma_1 q_1(t) + \bar\gamma_{2i} q_{2i}(t) \end{aligned}$$

for all $\tau \in [-M, M]$, $t \geq 0$, and $i = 1, \ldots, n$. Thus, for each $i = 1, \ldots, n$, we have

$$N_1(z,t) \geq \epsilon'\beta + q_1' - \bar{\gamma}_1 q_1 - \sum_{l=1}^{n} \bar{\gamma}_{2l} q_{2l},$$
$$N_{2i}(z,t) \leq -\epsilon'\beta - q_{2i}' + \bar{\gamma}_1 q_1 + \sum_{l=1}^{n} \bar{\gamma}_{2l} q_{2l}.$$

Setting $\hat{\gamma}_1 = \bar{\gamma}_1 + \mu$ and $\hat{\gamma}_{2i} = \bar{\gamma}_{2i} + \mu$, $i = 1, \ldots, n$, then from (5.10) it follows that for each $i = 1, \ldots, n$ we have

$$N_1(z,t) \geq \epsilon'\beta - \hat{\gamma}_1 q_1 - \sum_{l=1}^{n} \hat{\gamma}_{2l} q_{2l},$$
$$N_{2i}(z,t) \leq -\epsilon'\beta + \hat{\gamma}_1 q_1 + \sum_{l=1}^{n} \hat{\gamma}_{2l} q_{2l}.$$

Let $\epsilon' = (\hat{\gamma}_1 q_1 + \sum_{l=1}^{n} \hat{\gamma}_{2l} q_{2l})/\beta$ and $\epsilon(0) = z^*$, where z^* is a constant to be determined later. Hence

$$\epsilon(t) = \left[z^* + \frac{d}{\mu\beta} \left(\hat{\gamma}_1 + \sum_{l=1}^{n} \hat{\gamma}_{2l} k_{0l} \right) \right] - \left[\frac{d}{\mu\beta} \left(\hat{\gamma}_1 + \sum_{l=1}^{n} \hat{\gamma}_{2l} k_{0l} \right) \right] e^{-\mu t}.$$

With this choice of $\epsilon(t)$ it follows from the above discussion that $N_1(z,t) \geq 0$ and $N_{2i}(z,t) \leq 0$ for all $(z,t) \in \mathbf{R} \times \mathbf{R}^+$ and $i = 1, \ldots, n$. Now we will choose suitable d such that $\tilde{u}(\cdot, 0) \geq 0$ and $\tilde{\mathbf{v}}(\cdot, 0) \leq \mathbf{0}$. Once these are done, then by the discussion right before finding $\epsilon(t)$, $q_1(t)$, and $\mathbf{q_2}(t)$, we can conclude that

$$\tilde{u}(z,t) \geq 0 \ \text{ and } \ \tilde{v}_i(z,t) \leq 0 \tag{5.12}$$

for all $(z,t) \in \mathbf{R} \times \mathbf{R}^+$ and $i = 1, \ldots, n$. Indeed, let $d_1 \in (0, d_0)$, $a_2 - \phi_2 > 0$, and $\psi_{2i} - b_2^i > 0$ be sufficiently small satisfying that

$$a_2 - d < \phi_2, \quad b_2^i + dk_{0i} > \psi_{2i}$$

and that $\underline{u}(z,t)$ and $\underline{v}_i(z,t)$ satisfy (5.6) for all $(z,t) \in \mathbf{R} \times \mathbf{R}^+$, for all $d \in (0, d_1)$, and $i = 1, \ldots, n$. Hence, if $z^* > 0$ is chosen sufficiently large, it follows that for $i = 1, \ldots, n$, we have

$$\mathcal{U}(z - z^*) - d < \phi(z), \ \ \Pi_i(z - z^*) + dk_{0i} > \psi_i(z) \ \text{ for all } z \in \mathbf{R},$$

and so $\tilde{u}(\cdot, 0) \geq 0$ and $\tilde{\mathbf{v}}(\cdot, 0) \leq \mathbf{0}$. This completes the proof of our claim.

Step 3. *Finally, we will reach our conclusion.* In fact, using (5.12) and noting that $\mathcal{U}$ is monotone increasing and Π_i, $i = 1, \ldots, n$, is monotone decreasing, we obtain that

$$\mathcal{U}(z - z_1) - de^{-\mu t} \leq \underline{u}(z,t) \leq u(z,t) \ \text{ and } \ v_i(z,t) \leq \underline{v}_i(z,t) \leq \Pi_i(z - z_1) + dk_{0i}e^{-\mu t}$$

for all $(z,t) \in \mathbf{R} \times [0, +\infty)$ and $i = 1, \ldots, n$. The proof is completed. □

Before estimating the partial derivatives of the solution $(u, \mathbf{v})$ of (4.3)–(4.5), we need the following lemma.

LEMMA 5.4. *Let* $(u, \mathbf{v})$ *be the solution of* (4.3)–(4.5) *satisfying that* $u_t(z,0) \geq 0$, $v_{i,t}(z,0) \leq 0$, $u_z(z,0) \geq 0$, *and* $v_{i,z}(z,0) \leq 0$ *for all* $z \in \mathbf{R}$ *and* $i = 1, \ldots, n$. *Then*

$u_t(z,t) \geq 0$, $v_{i,t}(z,t) \leq 0$, $u_z(z,t) \geq 0$, *and* $v_{i,z}(z,t) \leq 0$ *for all* $(z,t) \in \mathbf{R} \times [0,+\infty)$ *and* $i = 1, \ldots, n$.

Proof. We prove only that $u_t \geq 0$ and $v_{i,t} \leq 0$, $i = 1, \ldots, n$, on $\mathbf{R} \times [0,+\infty)$, since the remaining inequalities follow in a similar way. Define $p(z,t) = u_t(z,t)$ and $h_i(z,t) = -v_{i,t}(z,t)$, $i = 1, \ldots, n$, on $\mathbf{R} \times [0,+\infty)$. Then p and h_i satisfy the following system:

$$\begin{aligned} L_1[p, h_1, \ldots, h_n] &= p_t - Dp_{zz} - cp_z \\ &= F_u(u, \mathbf{v})(z,t)p - \sum_{j=1}^{n} F_{v_j}(u, \mathbf{v})(z,t)h_j \\ &\equiv N_1(z, t, p, h_1, \ldots, h_n), \\ L_{2i}[p, h_1, \ldots, h_n] &= h_{i,t} - ch_{i,z} \\ &= -G_{i,u}(u, \mathbf{v})(z,t)p + G_{i,v_i}(u, \mathbf{v})(z,t)h_i \\ &\equiv N_2(z, t, p, h_1, \ldots, h_n), \end{aligned}$$

together with the initial data

$$p(z,0) = u_t(z,0) \geq 0 \quad \text{and} \quad h_i(z,0) = -v_t(z,0) \geq 0,$$

where $h_{i,t} = \partial h_i / \partial t$ and $h_{i,z} = \partial h_i / \partial z$ for $i = 1, \ldots, n$. Recall that $F_{v_i}(u, \mathbf{v}) = -(k_+^i u + k_-^i) < 0$ and $G_{i,u}(u, \mathbf{v}) = -k_+^i v_i < 0$ for all $(u, \mathbf{v}) \in [a_0, a_2] \times [b_2^1, b_0^1] \times \cdots \times [b_2^n, b_0^n]$ and that $(u(z,t), \mathbf{v}(z,t)) \in [a_0, a_2] \times [b_2^1, b_0^1] \times \cdots \times [b_2^n, b_0^n]$ for all $(z,t) \in \mathbf{R} \times [0,+\infty)$ and $i = 1, \ldots, n$. Thus $F_{v_i}(u(z,t), \mathbf{v}(z,t)) < 0$ and $G_{i,u}(u(z,t), \mathbf{v}(z,t)) < 0$ for all $(z,t) \in \mathbf{R} \times [0,+\infty)$ and $i = 1, \ldots, n$. Combining this with the fact that $p(z,0) \geq 0$ and $h_i(z,0) \geq 0$ for all $z \in \mathbf{R}$, it follows from Theorem 14.11 on p. 203 of Smoller [23] that $p(z,t) \geq 0$ and $h_i(z,t) \geq 0$ for all $(z,t) \in \mathbf{R} \times [0,+\infty)$ and $i = 1, \ldots, n$. This completes the proof. □

Next we will estimate the partial derivatives of the solution $(u, \mathbf{v})$ of (4.3)–(4.5).

LEMMA 5.5. *Let* $(u, \mathbf{v})$ *be the solution of* (4.3)–(4.5). *Then, under the hypothesis of Theorem* 3, *there exist constants* $\kappa_j < 0$, $\sigma_j > 0$, *and* $C_1 > 0$ *satisfying that* $\kappa_j/2 + \sigma_j > 0$ *for* $j = 1, 2$, *that*

$$|u - a_0|, |u_t|, |u_z|, |u_{zz}|, |b_0^i - v_i|, |v_{i,t}|, |v_{i,z}| < C_1(e^{(-(\kappa_1/2)+\sigma_1)z} + e^{-\mu t}) \tag{5.13}$$

for $z \leq 0$, $t \geq 0$, *and* $i = 1, \ldots, n$, *and that*

$$|a_2 - u|, |u_t|, |u_z|, |u_{zz}|, |v_i - b_2^i|, |v_{i,t}|, |v_{i,z}| < C_1(e^{-(\kappa_2/2+\sigma_2)z} + e^{-\mu t}) \tag{5.14}$$

for $z > 0$, $t \geq 0$, *and* $i = 1, \ldots, n$, *where* μ *is defined in Lemma* 5.3.

Proof. We consider only the case $z \geq 0$ since the case $z < 0$ follows by analogous arguments.

Step 1. *We claim that* $|a_2 - u|$ *and* $|v_i - b_2^i|$, $i = 1, \ldots, n$, *satisfy* (5.14). Linearizing (3.1)–(3.2) around the constant solution π_2, we obtain the equation $d\mathbf{X}/d\xi = A_0\mathbf{X}$, where $\mathbf{X} = (\mathrm{U} - a_2, \mathrm{Z}, \mathrm{V}_1 - b_2^1, \ldots, \mathrm{V}_n - b_2^n)^t$,

$$A_0 = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ -F_u/D & -c/D & -F_{v_1}/D & \cdots & -F_{v_n}/D \\ -G_{1,u}/c & 0 & -G_{1,v_1}/c & \cdots & -G_{1,v_n}/c \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -G_{n,u}/c & 0 & -G_{n,v_1}/c & \cdots & -G_{n,v_n}/c \end{bmatrix} \quad \text{with } G_{i,v_j} = \partial G_i / \partial v_j, \tag{5.15}$$

and all the values are evaluated at $(\mathrm{U}, \mathrm{Z}, \mathrm{V}_1, \ldots, \mathrm{V}_n) = (a_2, 0, b_2^1, \ldots, b_2^n)$. And so the associated characteristic polynomial with A_0 is

$$p_n(\lambda) = \det \begin{bmatrix} -\lambda & 1 & 0 & \cdots & 0 \\ -F_u/D & -c/D - \lambda & -F_{v_1}/D & \cdots & -F_{v_n}/D \\ -G_{1,u}/c & 0 & -G_{1,v_1}/c - \lambda & \cdots & -G_{1,v_n}/c \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -G_{n,u}/c & 0 & -G_{n,v_1}/c & \cdots & -G_{n,v_n}/c - \lambda \end{bmatrix}. \tag{5.16}$$

Note that $\partial G_i/\partial v_j = 0$ if $i \neq j$ and $G_{i,v_i}(a_2, \mathbf{b_2}) = -(k_+^i + k_-^i) < 0$ for $i = 1, \ldots, n$. Then, by a careful calculation, we have the following equalities:

$$\begin{aligned} p_n(0) &= -\frac{(1-a)}{c^n D} \prod_{i=1}^{n} (k_+^i + k_-^i), \\ p_n(-G_{i,v_i}(a_2, \mathbf{b_2})/c) &= p_n((k_+^i + k_-^i)/c) \\ &= \frac{b_0^i k_+^i k_-^i}{c^n D} \prod_{j \neq i}^{n} [(k_+^j + k_-^j) - (k_+^i + k_-^i)], \ i = 1, \ldots, n. \end{aligned}$$

Set $k_+^0 = k_-^0 = 0$. Therefore, if we assume that $(k_+^1 + k_-^1) < \cdots < (k_+^n + k_-^n)$, then we have $p_n((k_+^i + k_-^i)/c) p_n((k_+^{i+1} + k_-^{i+1})/c) < 0$ for $i = 0, \ldots, n-1$. Recall that $c < 0$. Therefore, we may assume that the eigenvalues $\lambda_1, \ldots, \lambda_{n+2}$ of A_0 satisfy

$$\lambda_{n+2} < (k_+^n + k_-^n)/c < \lambda_{n+1} < \cdots < \lambda_3 < (k_+^1 + k_-^1)/c < \lambda_2 < 0 < \lambda_1. \tag{5.17}$$

On the other hand, if $\{k_+^i + k_-^i\}_{i=1}^n$ do not satisfy this assumption, then we can use a similar argument as in Lemma 5.1 to obtain that the eigenvalues $\lambda_1, \ldots, \lambda_{n+2}$ of A_0 still satisfy (5.17) (" $<$ " may be replaced with " $=$," but $\lambda_2 < 0 < \lambda_1$ must hold).

Now we rewrite $p_n(\lambda)$ in the factored form $p_n(\lambda) = (\lambda^2 + k\lambda + l) \prod_{i=3}^{n+2} (\lambda_i - \lambda)$; then it follows that $\lambda_1 = -k/2 + \tilde{\sigma}$ and $\lambda_2 = -k/2 - \tilde{\sigma}$, where $\tilde{\sigma} = (k^2 - 4l)^{1/2}/2 > 0$. Since $c < 0$ and $\lambda_{i+2} \leq (k_+^i + k_-^i)/c < 0$ for $i = 1, \ldots, n$, we have

$$\lambda_1 + \lambda_2 = -c/D + \sum_{i=1}^{n} [(k_+^i + k_-^i)/c - \lambda_{i+2}] > 0,$$

and so $k < 0$. Since $\lambda_1 > 0$ and $\lambda_i < 0$ for $i = 2, \ldots, n+2$, we have $(\mathcal{U}, \mathbf{\Pi}) \to (a_2, \mathbf{b_2})$ exponentially fast as $z \to +\infty$. Hence there exist constants $\tilde{C}_1 > 0$, $\kappa_2 < 0$, and $0 < \sigma_2 < \tilde{\sigma}$ such that $\kappa_2/2 + \sigma_2 > 0$ and

$$\max_{i=1,\ldots,n} \{|a_2 - \mathcal{U}|, |\Pi_i - b_2^i|\} < \tilde{C}_1 e^{-(\kappa_2/2 + \sigma_2)z}$$

for all $z \geq 0$. Using this inequality and Lemma 5.3, then the inequalities for $u(z,t)$ and $v_i(z,t)$, $i = 1, \ldots, n$, follow.

Step 2. *We claim that there exists* $m_i < 0$ *such that* $m_i(v_i - b_2^i) \leq v_{i,z} \leq 0$ *on* $\mathbf{R} \times \mathbf{R}^+$ *for* $i = 1, \ldots, n$. Define

$$m_i = \inf_{(u,\mathbf{v}) \in [a_0, a_2] \times [b_2^1, b_0^1] \times \cdots \times [b_2^n, b_0^n]} \left\{ \frac{G_{i,v_i}(u, \mathbf{v})}{-c} \right\}.$$

Since $G_{i,v_i}(u, \mathbf{v}) = -(k_+^i u + k_-^i)$, we have $m_i < 0$ for $i = 1, \ldots, n$. Also recall that

$$G_i(u, \mathbf{b_2}) = k_-^i b_0^i - (k_+^i u + k_-^i)[k_-^i b_0^i/(k_+^i + k_-^i)] = k_-^i b_0^i (1-u)/(K_i + 1) \geq 0$$

for all $u \in [a_0, a_2]$ and $i = 1, \ldots, n$, where $K_i = k_-^i/k_+^i$. Noting that $v_{i,t}(\cdot, 0) \le 0$ on $\mathbf{R}$, it follows from Lemma 5.4 that $v_{i,t}(z,t) \le 0$ for all $(z,t) \in \mathbf{R} \times \mathbf{R}^+$. Therefore, using (4.4), the definition of G_i, and the mean-value theorem, and noting that $v_i \in [b_2^i, b_0^i]$ on $\mathbf{R} \times \mathbf{R}^+$, we obtain

$$\begin{aligned} 0 \ge v_{i,t} &= cv_{i,z} + G_i(u, \mathbf{v}) \\ &= cv_{i,z} + G_i(u, b_2^1, \ldots, b_2^{i-1}, v_i, b_2^{i+1}, \ldots, b_2^n) \\ &\ge cv_{i,z} + G_i(u, b_2^1, \ldots, b_2^{i-1}, v_i, b_2^{i+1}, \ldots, b_2^n) - G_i(u, b_2^1, \ldots, b_2^n) \\ &\ge cv_{i,z} - cm_i(v_i - b_2^i) \end{aligned}$$

on $\mathbf{R} \times \mathbf{R}^+$ for $i = 1, \ldots, n$. Combining this with Lemma 5.4, we obtain

$$m_i(v_i - b_2^i) \le v_{i,z} \le 0 \tag{5.18}$$

on $\mathbf{R} \times \mathbf{R}^+$ for $i = 1, \ldots, n$, since $v_{i,z}(\cdot, 0) \le 0$ on $\mathbf{R}$.

Step 3. *Estimates for the partial derivatives of* v_i. Indeed, (5.18) and the estimate obtained in Step 1 for $|v_i - b_2^i|$ lead to the inequality for $|v_{i,z}|$ for $i = 1, \ldots, n$. Rewrite (4.4) as the following:

$$v_{i,t} = cv_{i,z} + k_+^i(a_2 - u)(v_i - b_2^i) + k_+^i b_2^i(a_2 - u) - (k_+^i + k_-^i)(v_i - b_2^i).$$

Combining this with the estimates for $|a_2 - u|$, $|v_i - b_2^i|$, and $|v_{i,z}|$, we obtain the estimate for $|v_{i,t}|$ for $i = 1, \ldots, n$.

Step 4. *Estimates for the partial derivatives of* u. First, we set up some notation. For sufficiently smooth functions $g(z)$ and $h(z,t)$, and $\Omega \subset \mathbf{R} \times [0, +\infty)$, we define $|g|_0 \equiv \sup_{z \in \mathbf{R}} |g(z)|$, $[g]_\delta \equiv \sup_{x \ne y \in \mathbf{R}} |g(x) - g(y)|/|x - y|^\delta$, and $|g|_{2+\delta} \equiv |g|_0 + |g_z|_0 + |g_{zz}|_0 + [g_{zz}]_\delta$; $|h|_{0;\Omega} \equiv \sup_\Omega |h(z,t)|$, $[h]_{\delta,\delta/2;\Omega} \equiv \sup_{(x,t) \ne (y,s) \in \Omega} |h(x,t) - h(y,s)|/(|x - y|^\delta + |t - s|^{\delta/2})$, $|h|_{\delta,\delta/2;\Omega} = |h|_{0;\Omega} + [h]_{\delta,\delta/2;\Omega}$, and $|h|_{2+\delta,1+\delta/2;\Omega} \equiv |h|_{0;\Omega} + |h_z|_{0;\Omega} + |h_{zz}|_{\delta,\delta/2;\Omega} + |h_t|_{\delta,\delta/2;\Omega}$. Given the set $Q = Q_{z_0,t_0} = [z_0, z_0 + 1] \times [t_0, t_0 + 3/2]$ with $z_0 \in \mathbf{R}$ and $t_0 \ge 0$, set $Q' = Q'_{z_0,t_0} = [z_0 - 1, z_0 + 2] \times [t_0 - 1/2, t_0 + 2]$ with $z_0 \in \mathbf{R}$ and $t_0 \ge 1$ and $Q'' = Q''_{z_0} = [z_0 - 1, z_0 + 2] \times [0, 2]$ with $z_0 \in \mathbf{R}$. Rewrite (4.3) as the following:

$$\begin{aligned} (u - a_2)_t &= D(u - a_2)_{zz} + c(u - a_2)_z - u(u - a)(u - a_2) \\ &\quad + \sum_{j=1}^n [k_+^j(a_2 - u)(v_j - b_2^j) + k_+^j b_2^j(a_2 - u) - (k_+^j + k_-^j)(v_j - b_2^j)] \\ &\equiv D(u - a_2)_{zz} + c(u - a_2)_z + \tilde{f}(z,t). \end{aligned} \tag{5.19}$$

Note that $\tilde{f}(z,t)$, $u_z(z,0)$ are uniformly bounded on $\mathbf{R} \times [0, +\infty)$, $\mathbf{R}$, respectively. Then applying Theorem 6.28 and 6.33 of Lieberman [12] to (5.19) on the set Q, we obtain that there exist positive constants $\alpha \in (0,1)$ and c_1, determined by D, c and independent of (z_0, t_0), satisfying that

$$|u - a_2|_{\alpha;Q} \le c_1(|u - a_2|_{0;Q'} + |v - b_2|_{0;Q'}) \text{ if } t_0 \ge 1 \tag{5.20}$$

and

$$|u - a_2|_{\alpha;Q} \le c_1(|u - a_2|_{0;Q''} + |v - b_2|_{0;Q''} + |u_z(\cdot, 0)|_0) \text{ if } t_0 = 0. \tag{5.21}$$

Finally, (5.20), (5.21), and Schauder estimates (see Theorem 5 on p. 64 and Theorem 4 on p. 121 of Friedman [6]) imply the remaining inequalities for the partial derivatives of $u(z,t)$ for all $(z,t) \in \mathbf{R} \times [0, +\infty)$. Hence the proof is completed. □

5.3. Proof of Lemma 5.6. This lemma is a simple extension of Lemma 4.2 of [5] and is the so-called local stability of traveling wave fronts (in the C^0 norm).

LEMMA 5.6. *Let $(u, \mathbf{v})$ be the solution of (4.3)–(4.5). Then there exists a function $\omega(\eta)$, defined for small positive η, such that the following properties hold:*

(1) $\lim_{\eta \to 0^+} \omega(\eta) = 0$*;*

(2) *if there exists $z_0 \in \mathbf{R}$ such that $|\mathcal{U}(z - z_0) - \phi(z)| < \eta$ and $|\Pi_i(z - z_0) - \psi_i(z)| < \eta$ for all $z \in \mathbf{R}$ and $i = 1, \ldots, n$, then we have*

$$|u(z,t) - \mathcal{U}(z - z_0)| < \omega(\eta) \quad \text{and} \quad |\, v_i(z,t) - \Pi_i(z - z_0)| < \omega(\eta)$$

for all $(z,t) \in \mathbf{R} \times [0, +\infty)$ and $i = 1, \ldots, n$.

Proof. Since the proof is standard, we will briefly describe it. Without loss of generality, we may assume that $z_0 = 0$. Recall the definitions of $\underline{u}(z,t)$ and $\underline{\mathbf{v}}(z,t)$ from Lemma 5.3, i.e.,

$$\begin{aligned} \underline{u}(z,t) &= \mathcal{U}(z - \epsilon(t)) - de^{-\mu t}, \\ \underline{v}_i(z,t) &= \Pi_i(z - \epsilon(t)) + dk_{0i}e^{-\mu t}, i = 1, \ldots, n, \end{aligned}$$

where we set $z^* = 0$ and $\epsilon(t) = d\nu(1 - e^{-\mu t})$ for some constant $\nu > 0$. Also recall from the proof of Lemma 5.3 that there exist $d_1 > 0$, $\mu_0 > 0$ such that if $u(z,0) \geq \underline{u}(z,0)$ and $v_i(z,0) \leq \underline{v}_i(z,0)$, $i = 1, \ldots, n$, for all $z \in \mathbf{R}$, then for all $d \in (0, d_1)$ and $\mu \in (0, \mu_0)$ we obtain that $(\underline{u}, \underline{\mathbf{v}})$ satisfies (5.6), and $u(z,t) \geq \underline{u}(z,t)$ and $v_i(z,t) \leq \underline{v}_i(z,t)$ for all $(z,t) \in \mathbf{R} \times [0, +\infty)$ and $i = 1, \ldots, n$. The key points are the above comparison argument and $\epsilon(t) = O(d)$. Now we define $w_1(\eta) \equiv \max_{1 \leq i \leq n}\{\eta/k_{0i}, \eta k_{0i}\}$ with $\eta \in (0, \min_{1 \leq j \leq n}\{d_1, d_1 k_{0j}\})$. Then using the mean value theorem (or following a similar argument as in [10, 5]), we can reach our conclusion. ∎

Acknowledgments. The authors thank Professor Jong-Shenq Guo for his valuable help and the Mathematical Biosciences Institute of The Ohio State University for its support.

REFERENCES

[1] X. CHEN, *Existence, uniqueness, and asymptotic stability of traveling waves in nonlocal evolution equations*, Adv. Differential Equations, 2 (1997), pp. 125–160.

[2] N. F. BRITTON, *Reaction-Diffusion Equations and their Applications to Biology*, Academic Press, London, 1986.

[3] M. FALCKE, *Buffers and oscillations in intracellular Ca^{2+} dynamics*, Biophys. J., 84 (2003), pp. 28–41.

[4] P. FIFE, *Mathematical Aspects of Reacting and Diffusing Systems*, Lecture Notes in Biomath. 28, Springer-Verlag, New York, 1979.

[5] P. C. FIFE AND J. B. MCLEOD, *The approach of solutions of non-linear diffusion equations to traveling front solutions*, Arch. Ration. Mech. Anal., 65 (1977), pp. 335–361.

[6] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.

[7] S. P. HASTINGS, *On traveling wave solutions of the Hodgkin-Huxley equations*, Arch. Rational Mech. Anal., 60 (1976), pp. 229–257.

[8] M. S. JAFRI AND J. KEIZER, *On the roles of Ca^{2+} diffusion, Ca^{2+} buffers and the endoplasmic reticulum in IP_3-induced Ca^{2+} waves*, Biophys. J., 69 (1995), pp. 2139–2153.

[9] J. P. KEENER, *Waves in excitable media*, SIAM J. Appl. Math., 39 (1980), pp. 528–548.

[10] G. A. KLAASEN AND W. C. TROY, *The stability of traveling wave front solutions of a reaction-diffusion system*, SIAM J. Appl. Math., 41 (1981), pp. 145–167.

[11] N. V. KRYLOV, *Lectures on Elliptic and Parabolic Equations in Hölder Spaces*, AMS, Providence, RI, 1996.

[12] G. M. LIEBERMAN, *Second Order Parabolic Differential Equations*, World Scientific, Singapore, 1996.

[13] M. Naraghi and E. Neher, *Linearized buffered Ca^{2+} diffusion in microdomains and its implications for calculation of $[Ca^{2+}]$ at the mouth of a calcium channel*, J. Neurosci., 17 (1997), pp. 6961–73.

[14] M. Naraghi, T. H. Muller, and E. Neher, *Two-dimensional determination of the cellular Ca^{2+} binding in bovine chromaffin cells*, Biophys. J., 75 (1998), pp. 1635–1647.

[15] E. Neher, *Usefulness and limitations of linear approximations to the understanding of Ca^{2+} signals*, Cell Calcium, 24 (1998), pp. 345–57.

[16] M. C. Nowycky and M. J. Pinter, *Time courses of calcium and calcium-bound buffers following calcium influx in a model cell*, Biophys. J., 64 (1993), pp. 77–91.

[17] J. Rauch and J. Smoller, *Qualitative theory of the FitzHugh-Nagumo equations*, Advances in Math., 27 (1978), pp. 12–44.

[18] R. Redheffer and W. Walter, *Invariant sets for systems of partial differential equations* I: *Parabolic equations*, Arch. Rational Mech. Anal., 67 (1978), pp. 41–52.

[19] T. A. Rooney and A. P. Thomas, *Intracellular calcium waves generated by $Ins(1,4,5)P_3$ dependent mechanisms*, Cell Calcium, 14 (1993), pp. 674–690.

[20] F. Sala and A. Hernández-Cruz, *Calcium diffusion modeling in a spherical neuron: Relevance of buffering properties*, Biophys. J., 57 (1990), pp. 313–324.

[21] G. D. Smith, *Analytical steady-state solution to the rapid buffering approximation near an open Ca^{2+} channel*, Biophys. J., 71 (1996), pp. 3064–3072.

[22] G. D. Smith, L. Dai, R. M. Miura, and A. Sherman, *Asymptotic analysis of buffered calcium diffusion near a point source*, SIAM J. Appl. Math., 61 (2001), pp. 1816–1838.

[23] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1994.

[24] J. Sneyd, J. Keizer, and M. J. Sanderson, *Mechanisms of calcium oscillations and waves: A quantitative analysis*, FASEB J., 9 (1995), pp. 1463–1472.

[25] J. Sneyd, P. D. Dale, and A. Duffy, *Traveling waves in buffered systems: Applications to calcium waves*, SIAM J. Appl. Math., 58 (1998), pp. 1178–1192.

[26] J. J. Tyson and J. P. Keener, *Singular perturbation theory of traveling waves in excitable media*, Phys. D, 32 (1988), pp. 327–361.

[27] J. Wagner and J. Keizer, *Effects of rapid buffers on Ca^{2+} diffusion and Ca^{2+} oscillations*, Biophys. J., 67 (1994), pp. 447–456.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 266–285

SECOND HARMONICS EFFECTS IN RANDOM DUFFING OSCILLATORS*

JUAN A. ACEBRÓN† AND RENATO SPIGLER‡

To George Papanicolaou on his 60th birthday

Abstract. We consider a stochastic model for Duffing oscillators, where the nonlinearity and the randomness are scaled in such a way that they interact strongly. A typical feature is the appearance of *second harmonics* effects. An asymptotic statistical analysis for these oscillators is performed in the *diffusion limit*, when a suitable *absorbing boundary* condition is imposed, according to the underlying physical problem. The related Fokker–Planck equation has been numerically solved to obtain the first two moments of the oscillator's displacement from its rest-position. Dependence on the nonlinearity strength and on the location of the absorbing boundary has also been investigated. Such results have been compared with those computed solving the corresponding stochastic Ito differential equations by a Monte Carlo method, where *quasi-random* sequences of numbers have been efficiently used.

Key words. nonlinear random oscillators, Duffing oscillators, Fokker–Planck equation, diffusions with absorbing boundaries, quasi-Monte Carlo methods

AMS subject classifications. 34F05, 60H10, 60H35, 65C05

DOI. 10.1137/S0036139903437084

1. Introduction.

Duffing oscillators are among the simplest types of *nonlinear* oscillators. They are governed by second-order ordinary differential equations, describing, e.g., the free motion that a particle performs around its rest-position, subject to a certain nonlinear restoring force. They have been extensively studied (see, e.g., [25]). On the other hand, *random* harmonic (linear) oscillators have also been studied in the literature (see, e.g., [20]).

In [23], a stochastic model for nonlinear oscillators of the Duffing type was considered, in order to investigate the joint effect of nonlinearity and randomness. However, these two mechanisms were scaled in such a way that the effect of the nonlinearity on the underlying linear random oscillator model turned out to be rather insignificant in the behavior over the long run. The topic of nonlinear random oscillators is still an active area of research, especially in connection with applications to engineering; see [5].

In this paper, a stochastic model for a Duffing-type oscillator is considered where the effect of the nonlinearity is much more important than that analyzed in [23]. The model equation is

$$y_\epsilon'' + 2\epsilon^2\lambda(t)y_\epsilon' + \omega_0^2\left[1 + \epsilon\mu(t) + \epsilon w\nu(t)y_\epsilon^2\right]y_\epsilon = 0, \tag{1.1}$$

$$y_\epsilon(0) = y_1, \quad y_\epsilon'(0) = y_2, \tag{1.2}$$

*Received by the editors October 31, 2003; accepted for publication (in revised form) May 25, 2005; published electronically November 4, 2005. This work was carried out within the framework of the Italian GNFM-INdAM.

http://www.siam.org/journals/siap/66-1/43708.html

†Departamento de Automática, Escuela Politécnica, Universidad de Alcalá, Crta. Madrid-Barcelona, Km 31.600, 28871 Alcalá de Henares, Madrid, Spain (juan.acebron@uah.es).

‡Dipartimento di Matematica, Università di "Roma Tre," Largo S.L. Murialdo 1, 00146 Rome, Italy (spigler@mat.uniroma3.it).

where ϵ represents a small real parameter, w is a real parameter which sizes the nonlinearity, and $\lambda(\cdot)$, $\mu(\cdot)$, $\nu(\cdot)$ are suitable real-valued stochastic processes on some probability space. Therefore, $y_\epsilon(\cdot)$ will be a (real-valued) stochastic process as well. An *absorbing condition* is further imposed to take into account that the representative particle, whose random motion is described by (1.1), is lost whenever its position, $y_\epsilon(t)$, reaches a given value, say $\pm R$. This is the case of the so-called accelerator problem. In fact, when a beam of charged particles turns around in the vacuum, inside a toroidal chamber whose cross section has radius R, being confined by strong static magnetic fields, after a very large number of laps the beam opens up, and some particles are lost when they hit the material wall. We may think that here there is an absorbing boundary, and for a randomly perturbed problem this corresponds to a vanishing probability condition at a certain given radial distance from the center of the chamber. Clearly, such a condition amounts to imposing a boundary condition to the transition probability density obeying the Kolmogorov forward equation (Fokker–Planck). Therefore, the present problem, subject to an absorbing boundary condition, will be shown to possess a diffusion limit solution.

Note that only local existence and uniqueness of solutions to problem (1.1)–(1.2) can be guaranteed. However, the absorbing boundary condition associated to such a problem allows for existence and uniqueness of solutions up to the first time when $y_\epsilon(t)$ attains the value $\pm R$ (for every fixed ϵ and for every chosen realization of the noise terms). On the other hand, as soon as the considered trajectory ends up at the level $\pm R$, the corresponding particle is "killed" and goes out of the problem.

In [23], the nonlinear term $\epsilon w \nu(t) y_\epsilon^2$ was replaced by the "weaker" one, $\epsilon^2 w \nu(\cdot) y_\epsilon^2$, and the solution there was studied in the *diffusion limit*, attained when $\epsilon \to 0$ on a suitably long time scale. As will be shown, the model described by (1.1)–(1.2) exhibits some different features compared to the model in [23]. Among other things, stronger nonlinear effects, such as "second harmonics" effects, can now be observed.

It can be seen that the present problem, without imposing the absorbing boundary condition above, does *not* possess a diffusion limit, since, according to the Feller–Hille theory [7, 9], the formally obtained Kolmogorov forward and backward equations do *not* have *unique* solutions. A comment on this problem is given in section 5.

In section 2, the relevant assumptions on the stochastic processes $\lambda(\cdot)$, $\mu(\cdot)$, and $\nu(\cdot)$ are made. A parabolic differential equation (the Fokker–Planck equation) is then derived, which describes the time evolution of the transition probability density of the limiting-process. Such a process approximates the process y_ϵ in the diffusion limit. From this, the first two moments of the displacement of the oscillator are computed (section 3). In section 4, we describe the numerical treatment performed on such equations to get quantitative information. We also solve the underlying Ito stochastic differential equation for the purpose of comparison. Such a simulation, which is of the Monte Carlo type, has been accomplished by using quasi-random (low discrepancy) sequences of numbers [2, 3, 17]. This choice is an alternative to that of the more common sequences of pseudorandom numbers and here is shown to be effective due to the use of a "scrambling" strategy (a reordering technique [18]). This positive outcome contrasts with the earlier findings of [10], where the authors did not introduce any scrambling; see [1], however. Plots are given to illustrate the dependence of the moments on the various parameters, including the location R of the absorbing barrier and the strength, w, of nonlinearity, and the results are discussed. In section 5, the high points of the paper are summarized.

2. Statistical analysis in the diffusion limit. It is convenient to introduce the van der Pool coordinates, ρ_ϵ, φ_ϵ, defined by

$$y_\epsilon(t) = \rho_\epsilon(t)\cos\left(\omega_0 t + \varphi_\epsilon(t)\right), \quad y'_\epsilon(t) = -\omega_0\,\rho_\epsilon(t)\sin\left(\omega_0 t + \varphi_\epsilon t\right).$$

Note that $\rho_\epsilon^2 = y_\epsilon^2 + \omega_0^{-2} y_\epsilon'^2$ represents the energy of the oscillator. By using this transformation in (1.1)–(1.2), the system

$$\frac{d}{dt}\begin{pmatrix} \rho_\epsilon \\ \varphi_\epsilon \end{pmatrix} = \epsilon \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} + \epsilon^2 \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}, \tag{2.1}$$

where $F_i \equiv F_i(t, \rho_\epsilon, \varphi_\epsilon)$, $G_i \equiv G_i(t, \rho_\epsilon, \varphi_\epsilon)$, is obtained along with the initial conditions

$$\rho_\epsilon(0) = \left(y_1^2 + y_2^2/\omega_0^2\right)^{1/2}, \quad \varphi_\epsilon(0) = -\arctan\left(y_2/\omega_0 y_1\right)^{1/2}, \tag{2.2}$$

where we set, for short,

$$F_1 := \frac{\omega_0}{2}\,\rho_\epsilon\left\{\mu(t)\sin(2\omega_0 t + 2\varphi_\epsilon) + \nu(t)\frac{w}{2}\,\rho_\epsilon^2\left[1 + \cos(2\omega_0 t + 2\varphi_\epsilon)\right]\sin(2\omega_0 t + 2\varphi_\epsilon)\right\},$$

$$F_2 := \frac{\omega_0}{2}\left\{\mu(t)\left[1 + \cos(2\omega_0 t + 2\varphi_\epsilon)\right] + \nu(t)\frac{w}{2}\,\rho_\epsilon^2\left[1 + \cos(2\omega_0 t + \varphi_\epsilon)\right]^2\right\}, \tag{2.3}$$

$$G_1 := -\rho_\epsilon\,\lambda(t)\left[1 - \cos(2\omega_0 t + 2\varphi_\epsilon)\right],$$

$$G_2 := -\lambda(t)\sin(2\omega_0 t + 2\varphi_\epsilon).$$

Below, we shall assume that $\lambda(\cdot)$, $\mu(\cdot)$, and $\nu(\cdot)$ are real-valued, almost surely bounded, wide-sense stationary stochastic processes on some probability space, $(\Omega, \mathcal{A}, P)$. The dependence on the chance variable will be omitted throughout, as is customary. We shall assume that

$$E[\lambda(t)] = \lambda_0, \qquad E[\mu(t)] = 0, \qquad E[\nu(t)] = 0, \tag{2.4}$$

for some constant λ_0, where $E[\cdot]$ denotes taking expected values. Moreover, we shall assume that $\mu(\cdot)$ and $\nu(\cdot)$ satisfy a *mixing* condition in a sufficiently *strong* sense (see, e.g., [19, 21]). As for the stationarity, we shall assume below that $\mu(\cdot)$ and $\nu(\cdot)$ are stationarily correlated (see [4, p. 160], [26, pp. 78–79]), with covariance matrix

$$\begin{pmatrix} E[\mu(s)\mu(\sigma)] & E[\mu(s)\nu(\sigma)] \\ E[\nu(s)\mu(\sigma)] & E[\nu(s)\nu(\sigma)] \end{pmatrix}, \tag{2.5}$$

whose entries will be denoted by $R_{ij}(s - \sigma)$, $i, j = 1, 2$.

Under these hypotheses, we want to investigate whether the process converges weakly to some limiting-process when $\epsilon \to 0$ and $t \to +\infty$, with $\tau := \epsilon^2 t = \text{const.}$, uniformly on $0 \le \tau \le \tau_0$ (diffusion limit). Indeed, under similar hypotheses, this was proved to be true for the Duffing model studied in [23]. In that case, the limiting-process, denoted by $(\rho(\tau), \varphi(\tau))$, turned out to be a Markov process with trajectories continuous with probability 1. Therefore, it could be described by its infinitesimal generator

$$L := \sum_{i,j=1}^{2} a_{ij}(z)\frac{\partial^2}{\partial z_i \partial z_j} + \sum_{i=1}^{2}\left[b_i(z) + c_i(z)\right]\frac{\partial}{\partial z_i}, \tag{2.6}$$

where we set $z := (z_1, z_2)^T$ ($z_1 = \rho$, $z_2 = \varphi$, for us), and a_{ij}, b_i, c_i are given by

$$a_{ij}(z) := \lim_{t\to+\infty} \frac{1}{t} \int_0^t \int_0^s E\left[F_i(s,z)F_j(\sigma,z)\right] ds\, d\sigma, \quad i,j = 1,2,$$

$$b_i(z) := \lim_{t\to+\infty} \frac{1}{t} \int_0^t \int_0^s \sum_{j=1}^{2} E\left[\frac{\partial F_i(s,z)}{\partial z_j} F_j(\sigma,z)\right] ds\, d\sigma, \quad i = 1,2, \tag{2.7}$$

$$c_i(z) := \lim_{t\to+\infty} \frac{1}{t} \int_0^t E\left[G_i(s,z)\right] ds, \quad i = 1,2.$$

Here, we can proceed similarly, taking as quantities F_i, G_i, $i = 1,2$, those defined in (2.3). The result that we shall obtain is of the type of Khas'minskii's [12] (see also [24]), as used in [19, 21].

In our problem, it is possible to evaluate explicitly the quantities in (2.7). Though this is an elementary task, the derivation is quite lengthy; see, e.g., [23]. Therefore, we skip the details and give only the results (obtained using the assumptions on $\lambda(\cdot)$, $\mu(\cdot)$, $\nu(\cdot)$):

$$a_{11}(\rho) = \rho^2 \left[\beta_{11}(2) + \frac{w}{2}\,\rho^2(\beta_{12}(2) + \beta_{21}(2)) + \frac{w^2}{4}\,\rho^4 \left(\beta_{22}(2) + \frac{1}{4}\beta_{22}(4)\right)\right],$$

$$a_{12}(\rho) + a_{21}(\rho) = \frac{w}{2}\,\rho^3(\gamma_{12}(2) - \gamma_{21}(2)),$$

$$a_{22}(\rho) = 2\alpha_{11} + \beta_{11}(2) + \frac{w}{2}\,\rho^2[3(\alpha_{12} + \alpha_{21}) + 2(\beta_{12}(2) + \beta_{21}(2))]$$

$$+ \frac{w^2}{16}\,\rho^4\,(18\alpha_{22} + 16\beta_{22}(2) + \beta_{22}(4)), \tag{2.8}$$

$$b_1(\rho) = \rho \left[3\beta_{11}(2) + \frac{5}{2}\,w\,\rho^2(\beta_{12}(2) + \beta_{21}(2)) + \frac{7}{16}\,w^2\,\rho^4\,(4\beta_{22}(2) + \beta_{22}(4))\right],$$

$$b_2(\rho) = 2\left[\gamma_{11}(2) + w\,\rho^2\,(2\gamma_{12}(2) + \gamma_{21}(2)) + \frac{3}{2}\,w^2\rho^4 \left(\gamma_{22}(2) + \frac{1}{8}\,\gamma_{22}(4)\right)\right],$$

$$c_1(\rho) = -\lambda_0\,\rho, \qquad c_2(\rho) = 0,$$

where the notation

$$\alpha_{ij} := \frac{\omega_0^2}{8} \int_0^{+\infty} R_{ij}(x)\, dx,$$

$$\beta_{ij}(k) := \frac{\omega_0^2}{8} \int_0^{+\infty} R_{ij}(x) \cos(k\omega_0 x)\, dx, \quad k = 2,4, \tag{2.9}$$

$$\gamma_{ij}(k) := \frac{\omega_0^2}{8} \int_0^{+\infty} R_{ij}(x) \sin(k\omega_0 x)\, dx, \quad k = 2,4,$$

has been used.

Remark 2.1. Note that a_{ij}, b_i, c_i depend only on ρ (and not on φ).

Remark 2.2. Observe that in this model there is an effect due to the second harmonics, through the terms $\beta_{22}(4)$, $\gamma_{22}(4)$.

Remark 2.3. In the model studied in [23], $\nu(\cdot)$ had to be "strong enough" in order to be effective, because the nonlinearity was weak (of order $\mathcal{O}(\epsilon^2)$) in comparison to the randomness (of order $\mathcal{O}(\epsilon)$). For example, $E[\nu(\cdot)] \notin L(\mathbf{R}^+)$; $\nu(t) = \text{const.} \neq 0$ was acceptable. However, $E[\nu(\cdot)]$ had to be constant (0 or not) by the assumed stationarity.

In the present model, in some sense, $\nu(\cdot)$ has to be "not too strong," because the nonlinearity term now is considerably more important: A finite second moment, with a correlation function decaying to zero sufficiently fast, is required (by the mixing property), and $E[\nu(\cdot)] \equiv 0$ by the stationarity.

With the infinitesimal generator given by (2.6), (2.8), we can write down the Kolmogorov forward (or Fokker–Planck) equation

$$\frac{\partial p}{\partial \tau} = L^*[p], \tag{2.10}$$

$\tau := \epsilon^2 t$, satisfied by the transition probability density p, L^* denoting the adjoint operator of L (defined in (2.6)–(2.7)). Such an equation should be considered along with the initial value

$$p(\rho, \varphi; 0) = \delta(\rho - \rho_0)\delta(\varphi - \varphi_0), \tag{2.11}$$

where ρ_0 and φ_0 are the initial values of $\rho(t)$ and $\varphi(t)$, and the boundary condition

$$p(R, \varphi, \tau) = 0, \tag{2.12}$$

correspondingly to the absorbing boundary located at $\rho = R$.

We can then evaluate, in particular, the *moments* of the limiting-process $y(\tau) := \rho(\tau)\cos(\omega_0 t + \varphi(\tau))$. It is convenient to write (2.10) in the form

$$\begin{aligned}\frac{\partial p}{\partial \tau_1} = {} & \frac{\partial^2}{\partial \rho^2}\left[\left(1 + A_1 w\rho^2 + A_2 w^2\rho^4\right)\rho^2 p\right] + \frac{\partial^2}{\partial \rho \partial \varphi}\left(B_1 w \rho^3\, p\right) \\ & + \frac{\partial^2}{\partial \varphi^2}\left[\left(C_0 + C_1 w\rho^2 + C_2 w^2\rho^4\right)p\right] - \frac{\partial}{\partial \rho}\left[\left(D_0 + D_1 w\rho^2 + D_2 w^2\rho^4\right)\rho\, p\right] \\ & - \frac{\partial}{\partial \varphi}\left[\left(E_0 + E_1 w\rho^2 + E_2 w^2\rho^4\right)p\right],\end{aligned} \tag{2.13}$$

where

$$\tau_1 := \beta_{11}(2)\,\tau \equiv \beta_{11}(2)\,\epsilon^2 t, \tag{2.14}$$

$$\begin{aligned} A_1 &:= \frac{\beta_{12}(2) + \beta_{21}(2)}{2\beta_{11}(2)}, \\ A_2 &:= \frac{\beta_{22}(2) + \frac{1}{4}\beta_{22}(4)}{4\beta_{11}(2)}, \\ B_1 &:= \frac{\gamma_{12}(2) - \gamma_{21}(2)}{2\beta_{11}(2)}, \\ C_0 &:= 2\frac{\alpha_{11}}{\beta_{11}(2)} + 1, \\ C_1 &:= \frac{3(\alpha_{12} + \alpha_{21}) + 2(\beta_{12}(2) + \beta_{21}(2))}{2\beta_{11}(2)}, \\ C_2 &:= \frac{18\alpha_{22} + 16\beta_{22}(2) + \beta_{22}(4)}{16\beta_{11}(2)}, \end{aligned} \tag{2.15}$$

$$
\begin{aligned}
D_0 &:= 3 - \frac{\lambda_0}{\beta_{11}(2)},\\
D_1 &:= \frac{5(\beta_{12}(2)+\beta_{21}(2))}{2\beta_{11}(2)} \equiv 5A_1,\\
D_2 &:= \frac{7(\beta_{22}(2)+\frac{1}{4}\beta_{22}(4))}{4\beta_{11}(2)} \equiv 7A_2,\\
E_0 &:= 2\frac{\gamma_{11}(2)}{\beta_{11}(2)},\\
E_1 &:= \frac{2(2\gamma_{12}(2)+\gamma_{21}(2))}{\beta_{11}(2)},\\
E_2 &:= \frac{3(\gamma_{22}(2)+\frac{1}{8}\gamma_{22}(4))}{2\beta_{11}(2)}.
\end{aligned}
$$

Note the presence of some "second harmonics" terms (see Remark 2.2), which enter the coefficients above via the quantities $\beta_{22}(4)$, $\gamma_{22}(4)$. Their effect is to increase the value of A_2, C_2, D_2, and E_2 and thus, qualitatively, to increase the nonlinearity size, as $|w|$ would have been increased. Moreover, this happens independently of the sign of w.

Remark 2.4. It is worth noting that the solution to problem (2.10)–(2.12) is not (L^1) norm-increasing. In fact, integrating both sides of (2.13) on the domain $[0, 2\pi] \times [0, R]$ gives

$$
\begin{aligned}
&\frac{\partial}{\partial\tau_1}\int_0^{2\pi}\int_0^R p(\rho,\varphi,\tau_1)\,d\rho d\varphi\\
&= R^2(1 + A_1\,w\,R^2 + A_2\,w^2\,R^2)\int_0^{2\pi}\frac{\partial p}{\partial\rho}(R,\varphi,\tau_1)\,d\varphi, \qquad (2.16)
\end{aligned}
$$

where the absorbing boundary condition in (2.12) has been used. We now observe that $\partial p(R,\varphi,\tau_1)/\partial\rho \le 0$. In fact, such a quantity is nonpositive, given that p is positive inside the domain and zero on the boundary $\rho = R$. Therefore, $dP(\tau_1)/d\tau_1 \le 0$, where

$$
P(\tau_1) := \int_0^{2\pi}\int_0^R p(\rho,\varphi,\tau_1)\,d\rho d\varphi. \qquad (2.17)
$$

The quantity $P(\tau_1)$ represents the *survival probability* of the particle up to time τ_1, which is the probability that the particle does not hit the absorbing barrier before time τ_1. A similar result can be found in [6], where it is shown that, in the presence of an absorbing boundary condition, the norm of any initial data is not preserved.

In the special but important case that $\mu(\cdot)$ and $\nu(\cdot)$ are *uncorrelated* (possibly independent), there are some simplifications. As $R_{12}(\cdot) \equiv R_{21}(\cdot) \equiv 0$ in this case, the partial differential equation in (2.10) reduces to

$$
\begin{aligned}
\frac{\partial p}{\partial\tau_1} &= \frac{\partial^2}{\partial\rho^2}\left[(1 + A_2w^2\rho^4)\,\rho^2 p\right] + \frac{\partial^2}{\partial\varphi^2}\left[(C_0 + C_2w^2\rho^4)p\right]\\
&\quad - \frac{\partial}{\partial\rho}\left[(D_0 + D_2w^2\rho^4)\,\rho\,p\right] - \frac{\partial}{\partial\varphi}\left[(E_0 + E_2w^2\rho^4)\,p\right]. \qquad (2.18)
\end{aligned}
$$

Note that now the coefficients of the equation depend only on w^2, and therefore the results are *independent of the sign* of w. This is not true when $\mu(\cdot)$ and $\nu(\cdot)$ are

correlated. Note also that the effects of the second harmonics depend only on the autocorrelation of $\nu(\cdot)$, $R_{22}(\cdot)$ (in A_2, C_2, D_2, E_2, such an effect enters through $\beta(4)$ and $\gamma(4)$).

3. The time evolution of the moments. As the equation in (2.13) is linear and has coefficients independent of φ, we can perform a Fourier analysis. Expanding p in a Fourier series,

$$p(\rho, \varphi, \tau_1) = \sum_{m=-\infty}^{+\infty} p_m(\rho, \tau_1) e^{i\, m\varphi},$$

we obtain for the mth coefficient

$$p_m(\rho, \tau_1) = \frac{1}{2\pi} \int_0^{2\pi} p(\rho, \varphi, \tau_1)\, e^{-i\, m\, \varphi} d\varphi \tag{3.1}$$

the evolution equation

$$\begin{aligned} \frac{\partial p_m}{\partial \tau_1} &= \frac{\partial^2}{\partial \rho^2} \left[(1 + A_1 w\rho^2 + A_2 w^2 \rho^4)\, \rho^2\, p_m \right] \\ &- \frac{\partial}{\partial \rho} \left[\left(D_0 + (D_1 + i\, mB_1) w\rho^2 + D_2 w^2 \rho^4 \right) \rho\, p_m \right] \\ &+ i\, m \left[(i\, mC_0 - E_0) + (i\, mC_1 - E_1)\, w\rho^2 + (i\, mC_2 - E_2)\, w^2 \rho^4 \right] p_m, \end{aligned} \tag{3.2}$$

with the initial value

$$p_m(\rho, 0) = \frac{1}{2\pi} \delta(\rho - \rho_0)\, e^{-i\, m\, \varphi_0} \tag{3.3}$$

and the boundary condition

$$p_m(R, \tau_1) = 0. \tag{3.4}$$

The boundary point $\rho = R$ is a *regular* boundary, while $\rho = 0$ is a *natural boundary*, according to Feller's classification (see [7, 9]). On a natural boundary, such as $\rho = 0$, no boundary condition is required, while a condition is needed on a regular boundary so that the Fokker–Planck equation has a unique solution. In [7, sect. 23], the boundary value problem for the one-dimensional diffusion equation

$$u_t = \frac{\partial^2}{\partial x^2}(a(x)u) - \frac{\partial}{\partial x}(b(x)u) \tag{3.5}$$

and its adjoint has been considered on the interval $-\infty \le r_1 < x < r_2 \le +\infty$. Depending on the nature of the boundary points r_1, r_2 and the type of data imposed on them to the solution, such equations may or may not have a unique solution. Defining the function

$$W(x) := \exp\left\{ - \int_{x_0}^{x} \frac{b(s)}{a(s)}\, ds \right\}, \tag{3.6}$$

where $x_0 \in (r_1, r_2)$, the boundary points are classified as follows:

- The boundary r_j is *regular* if $W(x) \in L^1(x_0, r_j)$ and $a^{-1}(x) W^{-1}(x) \in L^1(x_0, r_j)$.

- It is an *exit* boundary if $a^{-1}(x)W^{-1}(x) \notin L^1(x_0, r_j)$ and $W(x)\int_{x_0}^{x} a^{-1}(s)W^{-1}(s)\,ds \in L^1(x_0, r_j)$.
- It is an *entrance* boundary if $a^{-1}(x)W^{-1}(x) \in L^1(x_0, r_j)$ and $a^{-1}(x)W^{-1}(x)\int_{x_0}^{x} W(s)\,ds \in L^1(x_0, r_j)$.
- It is *natural* in all other cases.

One may observe that for the heat equation $u_t = u_{xx}$ the boundaries $r_j = \pm\infty$ are both natural, while they are regular when they are finite.

It is straightforward to check that for our problem in (3.2), $r_1 = \rho = 0$ is a natural boundary, while $r_2 = \rho = R$ is a regular boundary. It follows from Feller's theory that our problem possesses a unique solution when the absorbing boundary condition in (3.4) is prescribed, while no condition is imposed on $\rho = 0$.

In the special case of $\mu(\cdot)$, $\nu(\cdot)$ uncorrelated, (2.13) reduces to (2.18) and we get the simpler problem

$$\begin{aligned}\frac{\partial p_m}{\partial \tau_1} &= \frac{\partial^2}{\partial \rho^2}\left[\left(1 + A_2 w^2 \rho^4\right)\rho^2\, p_m\right] + \frac{\partial}{\partial \rho}\left[\left(D_0 + D_2 w^2 \rho^4\right)\rho\, p_m\right] \\ &\quad + i\,m\left[\left(i\,mC_0 - E_0\right) + \left(i\,mC_2 - E_2\right) w^2 \rho^4\right] p_m,\end{aligned}$$

$$p_m(\rho, 0) = \frac{1}{2\pi}\delta(\rho - \rho_0)\, e^{-i\,m\,\varphi_0}, \qquad p_m(R, \tau_1) = 0. \tag{3.7}$$

We are primarily interested in computing the first two moments of the displacement $y_\epsilon(\cdot)$ of the oscillator by approximating it with the limiting-process $y(\cdot)$. For this we have

$$E_{\rho_0,\varphi_0}[y(\tau)] = \operatorname{Re}\left\{e^{i\omega_0 t}\, E_{\rho_0,\varphi_0}[\rho(\tau)\, e^{i\varphi(\tau)}]\right\}, \tag{3.8}$$

$$E_{\rho_0,\varphi_0}[y^2(\tau)] = \frac{1}{2}E_{\rho_0,\varphi_0}[\rho^2(\tau)] + \frac{1}{2}\operatorname{Re}\left\{e^{2i\omega_0 t}\, E_{\rho_0,\varphi_0}[\rho^2(\tau)\, e^{2i\varphi(\tau)}]\right\},$$

where the quantities $E_{\rho_0,\varphi_0}[\rho^k(\tau)\, e^{ik\varphi(\tau)}]$, $k = 1, 2$, and $E_{\rho_0,\varphi_0}[\rho^2(\tau)]$ can be computed by integrating (3.1) (or (3.7)). Recall that $\tau_1 = \beta_{11}(2)\epsilon^2 t$; see (2.14). We obtain

$$E_{\rho_0,\varphi_0}[y(\tau_1)] = 2\pi \operatorname{Re}\left[\int_0^R \rho\, p_1(\rho, \tau_1)d\rho\right]\cos\omega_0 t + 2\pi \operatorname{Im}\left[\int_0^R \rho\, p_1(\rho, \tau_1)d\rho\right]\sin\omega_0 t, \tag{3.9}$$

$$\begin{aligned}E_{\rho_0,\varphi_0}[y^2(\tau_1)] &= \pi\int_0^R \rho^2\, p_0(\rho, \tau_1)d\rho + \pi\left\{\operatorname{Re}\left[\int_0^R \rho^2\, p_2(\rho, \tau_1)d\rho\right]\cos(2\omega_0 t)\right. \\ &\quad \left. + \operatorname{Im}\left[\int_0^R \rho^2\, p_2(\rho, \tau_1)d\rho\right]\sin(2\omega_0 t)\right\}.\end{aligned} \tag{3.10}$$

The problem is clearly affected by *two time scales*, the fast (deterministic) scale, according to the "time" $\omega_0 t$, and the slow scale τ_1, on which nontrivial random phenomena occur in the diffusion limit; see [23]. In fact, it is over long times that the statistical cumulative effect of the small size noise becomes significant.

In the next section, we describe the numerical treatment carried out to solve the aforementioned problems. We give the relevant results in the form of several plots and discuss the observed features.

4. Numerical treatment. We solved numerically problem (3.7) for $m = 0, 1, 2$ on $(0, R) \times (0, T)$, with the boundary condition in (2.12), obtaining the Fourier coefficients $p_m(\rho, \tau_1)$ of the transition probability density $p(\rho, \varphi, \tau_1)$. The covariance matrix $\{R_{ij}(\cdot)\}_{i,j=1,2}$ also had to be specified. In the case above, where $\mu(\cdot)$, $\nu(\cdot)$ are supposed to be uncorrelated, we have only to assign $R_{11}(\cdot)$ and $R_{22}(\cdot)$. Let us choose

$$R_{11}(t) := e^{-\sigma_1|t|}, \qquad R_{22}(t) := e^{-\sigma_2|t|} \tag{4.1}$$

for some positive constants σ_1, σ_2. Then all quantities α_{ij}, β_{ij}, γ_{ij} and hence A_2, D_0, D_2, C_0, C_2, E_0, and E_2 can be computed.

To be more explicit, we get from (2.9)

$$\alpha_{ii} = \frac{\omega_0^2}{8}\frac{1}{\sigma_i}, \quad \beta_{ii}(q) = \frac{\omega_0^2}{8}\,\frac{\sigma_i}{\sigma_i^2 + q^2\omega_0^2}, \quad \gamma_{ii}(q) = \frac{\omega_0^2}{8}\,\frac{q\,\omega_0}{\sigma_i^2 + q^2\omega_0^2}, \quad i = 1, 2, \tag{4.2}$$

and therefore from (2.15)

$$A_2 = \frac{1}{4}\frac{\sigma_2}{\sigma_1}\left(1 + \frac{4\omega_0^2}{\sigma_1^2}\right)\left[\frac{1}{\sigma_2^2/\sigma_1^2 + 4\omega_0^2/\sigma_1^2} + \frac{1/4}{\sigma_2^2/\sigma_1^2 + 16\omega_0^2/\sigma_1^2}\right], \tag{4.3}$$

$$D_0 = 3 - 8\frac{\lambda_0/\omega_0}{\omega_0/\sigma_1}\left(1 + 4\frac{\omega_0^2}{\sigma_1^2}\right), \tag{4.4}$$

$$D_2 = 7A_2, \tag{4.5}$$

$$C_0 = 3 + 8\frac{\omega_0^2}{\sigma_1^2}, \tag{4.6}$$

$$C_2 = \frac{9}{8}\frac{1 + 4\omega_0^2/\sigma_1^2}{\sigma_2/\sigma_1} + \frac{\sigma_2}{\sigma_1}\frac{1 + 4\omega_0^2/\sigma_1^2}{\sigma_2^2/\sigma_1^2 + 4\omega_0^2/\sigma_1^2} + \frac{1}{16}\frac{\sigma_2}{\sigma_1}\frac{1 + 4\omega_0^2/\sigma_1^2}{\sigma_2^2/\sigma_1^2 + 16\omega_0^2/\sigma_1^2}, \tag{4.7}$$

$$E_0 = 4\frac{\omega_0}{\sigma_1}, \tag{4.8}$$

$$E_2 = \frac{3}{2}\left[2\frac{\omega_0}{\sigma_1}\frac{1 + 4\omega_0^2/\sigma_1^2}{\sigma_2^2/\sigma_1^2 + 4\omega_0^2/\sigma_1^2} + \frac{1}{2}\frac{\omega_0}{\sigma_1}\frac{1 + 4\omega_0^2/\sigma_1^2}{\sigma_2^2/\sigma_1^2 + 16\omega_0^2/\sigma_1^2}\right]. \tag{4.9}$$

These expressions make it clear that only the *nondimensional* quantities

$$\frac{\sigma_2}{\sigma_1}, \quad \frac{\omega_0}{\sigma_1}, \quad \text{and} \quad \frac{\lambda_0}{\omega_0} \tag{4.10}$$

play a role and therefore need to be assigned. Another parameter is the ratio of the *two time scales*, which characterize the deterministic oscillations *and* the random fluctuations. We have from (2.14) and (4.2)

$$\tau_1 \equiv \beta_{11}(2)\epsilon^2 t = \frac{\epsilon^2}{8}\,\frac{\omega_0/\sigma_1}{1 + 4\omega_0^2/\sigma_1^2}\,\omega_0 t =: \kappa\omega_0 t \tag{4.11}$$

(note that τ_1 is nondimensional) and, for $\omega_0/\sigma_1 \approx 1$, $\epsilon \approx 0.1$, we get $\kappa \approx 4000$. Note that κ is automatically determined by choosing ϵ and ω_0/σ_1. Therefore, we can assign κ or ϵ, in addition to the previous parameters.

The numerical treatment consists of implementing an *implicit* scheme of finite differences, with forward time differences and space-centered differences (the Crank–Nicholson scheme). We chose $R = 4$ and divided $[0, R]$ into sections of equal length $\Delta\rho = 10^{-3}$; the time-step size we used was $\Delta\tau_1 = 10^{-4}$, and the initial position of

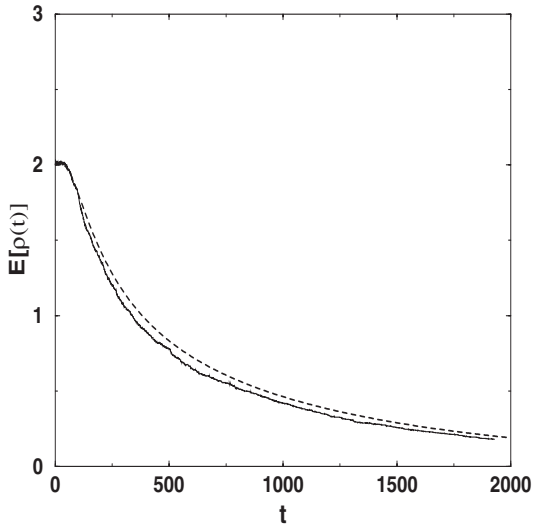


FIG. 4.1. *Time evolution of the expected value of ρ. The nonlinearity size is $w = 0.3$. The dashed line represents the solution obtained through the Fokker–Planck equation; the solid line is obtained by solving the stochastic differential equation. The other parameters are $\epsilon = 0.1$, $\sigma_2/\sigma_1 = 1$, $\omega_0/\sigma_1 = 1$, and $\lambda_0/\omega_0 = 3\beta_{11}(2)/2$.*

the oscillator was $\rho_0 = 2$. The parameters w and λ_0/ω_0 were varied to form several combinations in order to investigate the various effects, while we kept ϵ to the fixed value $\epsilon = 0.1$. The case $\epsilon = 0.5$ was also considered to test the validity of the limiting theory. The first two moments

$$E[\rho^k(\tau_1)] = 2\pi \int_0^R \rho^k p_0(\rho, \tau_1)\, d\rho, \quad k = 1, 2, \tag{4.12}$$

of the oscillator's amplitude, $\rho(\tau_1)$, defined by the van der Pool variables (in the diffusion limit), can then be evaluated. Note that $E[\rho^0] = E[1]$ coincides with the survival probability given by (2.17).

In Figure 4.1, we plotted $E[\rho]$ versus t for $w = 0.3$, $\sigma_2/\sigma_1 = 1$, $\omega_0/\sigma_1 = 1$, and $\lambda_0/\omega_0 = 3\beta_{11}(2)/2$. In Figure 4.2, the time evolution of the second moment $E[\rho^2]$ is shown for the same set of parameters. Recall that $\rho_\epsilon^2 = y_\epsilon^2 + \omega_0^{-2} y_\epsilon'^2$ represents the energy of the oscillator governed by (1.1), and hence $E[\rho^2]$ is the average energy of the oscillator in the diffusion limit. Here we also plotted $E[\rho^2]$ for $w = 0$, that is, for the corresponding linear harmonic oscillator, for the purpose of comparison.

In Figure 4.3, we plotted $E[\rho^2]$ versus τ_1 for the same values of the parameters used above, except that we considered several values of λ_0/ω_0. It is apparent that there is "a threshold" when λ_0/ω_0 goes across a certain value. Above such a value (that is, when the damping is sufficiently strong), $E[\rho^2]$ decreases monotonically in time. Below the threshold value, initially $E[\rho^2]$ grows in time, undergoing a kind of transient behavior, but then it decays in order to match the absorbing boundary condition. Such a threshold is determined by the sign of the coefficient D_0, which takes the values $3, -3, -9$ corresponding to the values $\lambda_0/\omega_0 = 0, 0.15, 0.3$, respectively.

Figure 4.4 shows the time evolution of the survival probability $P(t)$ given by (2.17). This has been computed both from the Fokker–Planck equation and by Monte Carlo simulations to provide mutual validation. The time evolution of two conditional

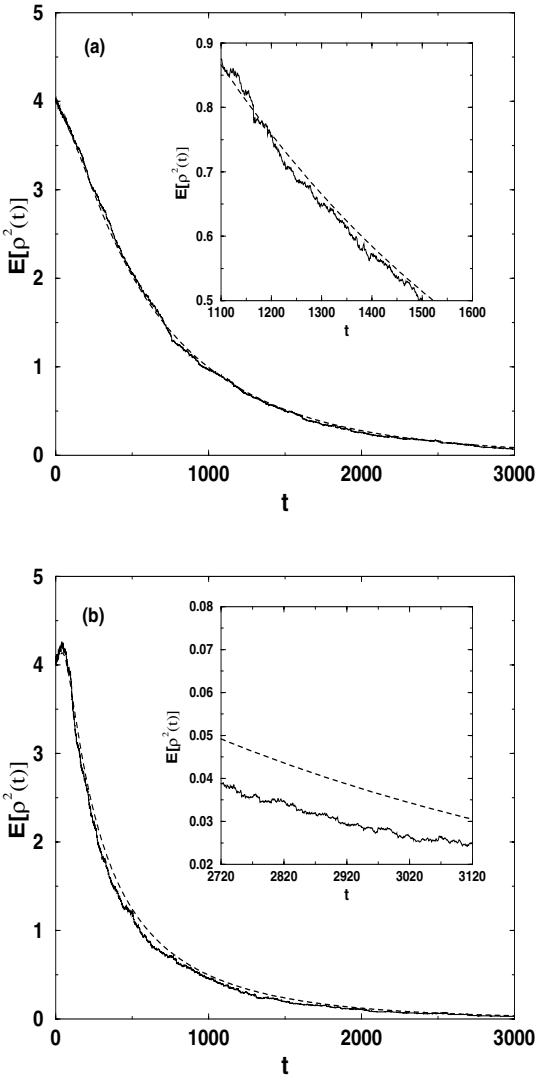


FIG. 4.2. *Time evolution of the expected value of ρ^2 for two different values of the nonlinearity parameter:* (a) $w = 0$ *and* (b) $w = 0.3$. *The other parameters are as in Figure* 4.1. *A magnification of part of the plot is shown in the inset.*

moments of ρ, assuming that the time τ_1 is less than the first hitting time, say τ_R, is plotted in Figure 4.5. Such moments can be obtained knowing the moments $E[\rho^k]$ and the survival probability $P(\tau_1)$ as follows:

$$E[\rho^k | \tau_1 < \tau_R] = \frac{E[\rho^k(\tau_1)]}{P(\tau_1)}. \tag{4.13}$$

These quantities have been computed from the Fokker–Planck equation. The same quantities have been obtained from the corresponding stochastic differential equations by Monte Carlo simulations, for the purpose of validation, and plotted in Figure 4.5 (solid line).

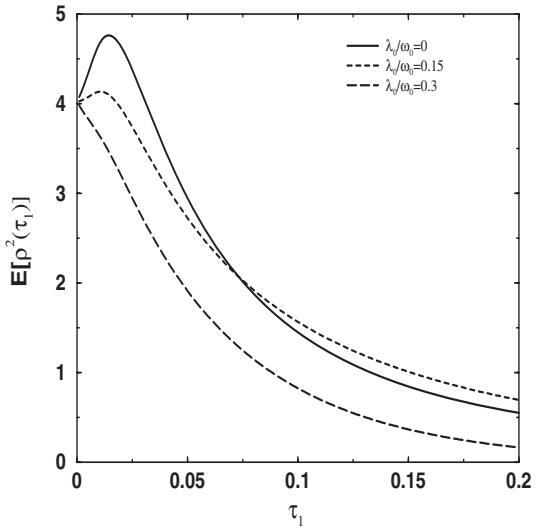


FIG. 4.3. *Time evolution of the expected value of* ρ^2 *for three different values of* λ_0/ω_0*, keeping* w *to the fixed value* 0.3*. The other parameters are as in Figure* 4.1.

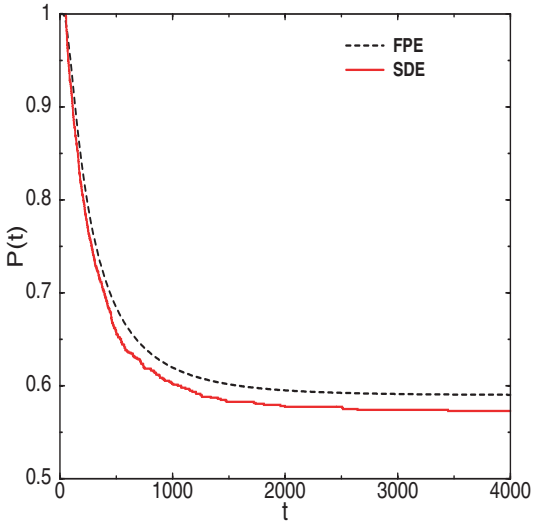


FIG. 4.4. *Time evolution of the survival probability* $P(t)$*, keeping* w *to the fixed value* 0.3 *and* $\lambda_0/\omega_0 = 0.15$*. The other parameters are as in Figure* 4.1.

It is worth noting that the survival probability seems to stabilize around a nonzero value, and the conditional moments of ρ go to zero, when time is sufficiently large. From both such features, the behavior of the surviving particles for a long time can be easily understood. The nonzero value of $P(t)$, say p_∞, would mean that a certain number of particles never exit from the boundary, while, given that the limiting value of the conditional moments when time goes to infinity is zero, it indicates that the surviving particles will be located, with high probability, around $\rho = 0$.

In Figure 4.6, the time evolution of the survival probability, $P(t)$, is plotted for three different values of R. Recall that the parameter R is part of the data of the

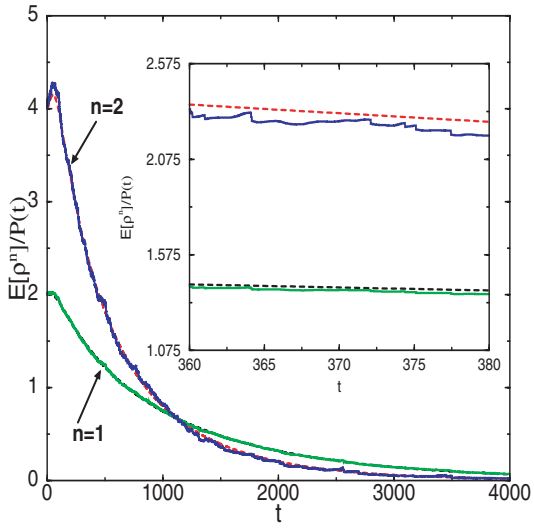


FIG. 4.5. *Time evolution of two conditional moments of ρ with $n = 1, 2$. The parameters are as in Figure* 4.4.

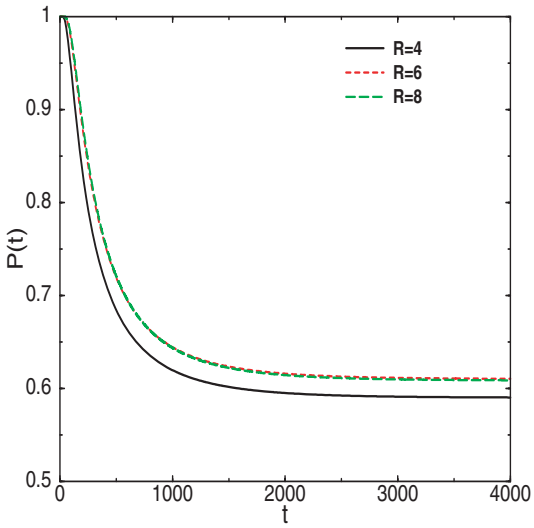


FIG. 4.6. *Time evolution of the survival probability $P(t)$ for three different values of R, obtained by solving the Fokker–Planck equation. The parameters are as in Figure* 4.4.

problem. For instance, in the example of the particle accelerator, R may be the radius of a vacuum chamber. Note that increasing the value of R increases the stationary value of $P(t)$. Clearly, when the boundary R is closer to the initial value $\rho = \rho_0$, the exiting probability of a given particle becomes larger, thus making the survival probability smaller in this case. Nevertheless, for sufficiently large values of R, such a probability seems to become independent of R. However, larger values of R cannot be used in practice to prevent computational overflow, in view of the exponential dependence of some coefficients in the Fokker–Planck equation.

In Figure 4.7, the survival probability has been plotted as a function of time

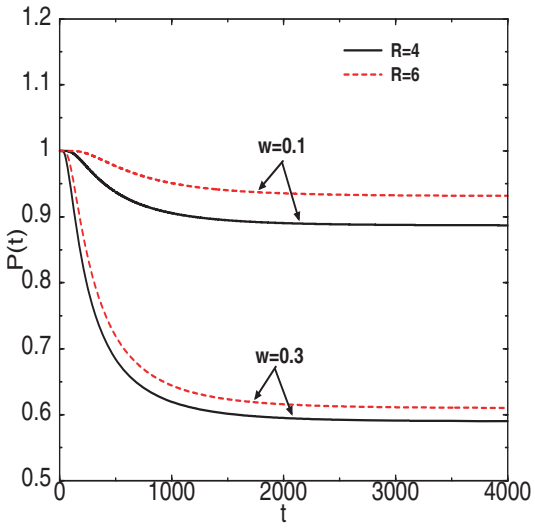


FIG. 4.7. *Time evolution of the survival probability $P(t)$ for two different values of R and the nonlinearity w. The parameters are as in Figure 4.4.*

for two pairs of values of the nonlinearity parameter w and of the location of the absorbing barrier R. These results have been obtained by integrating numerically the Fokker–Planck equation. The numerical simulations seem to show that the survival probability tends to stabilize, after a sufficiently long time, at some nonzero (positive) value. Such a value seems to decrease when the nonlinearity becomes stronger (for any given value of R), while fewer particles tend to be absorbed when the absorbing boundary is located farther from the initial position of the oscillator for any given value of w. In other words, there is numerical evidence that a stronger nonlinearity favors higher absorption of particles, while a more distant barrier makes it more difficult. No claim can be made, however, that the survival probability does indeed tend to a nonzero value when time increases. In fact, the survival probability might decay so slowly that the numerical simulations cannot fully capture its precise behavior.

In Figure 4.8, the full transition probability density, $p_0(\rho, t)$, has been plotted as a function of ρ for several times. The inset in this figure shows the corresponding values of the survival probability obtained for the times when the probability density was plotted. This picture illustrates well what happens to the ρ-profile of $p_0(\rho, t)$ at various times. Such a profile starts smooth and well spread over the full interval $0 < \rho < R = 4$, but as time goes on, it tends to become a Dirac delta function, located at the initial position of the oscillator, $\rho_0 = 0$.

In Figure 4.9, the dependence of the survival probability on two different initial positions of the oscillator, $\rho_0 = 2$ or $\rho_0 = 3$, is shown. It is clear that the closer such an initial position is to the absorbing boundary (located at $\rho = R = 4$), the higher will be the number of the particles absorbed there.

Figure 4.10 displays the dependence of p_∞ on the nonlinearity parameter w. It appears that p_∞ decreases steadily as w increases. Here $R = 4$ was kept fixed. We have also computed the first two moments of the oscillator's displacement, $y(t)$, given by (3.9), (3.10). In Figure 4.11, such moments are shown for the same values of the parameters used in Figures 4.1 and 4.2.

In order to validate the limiting theory, we conducted numerical simulations of

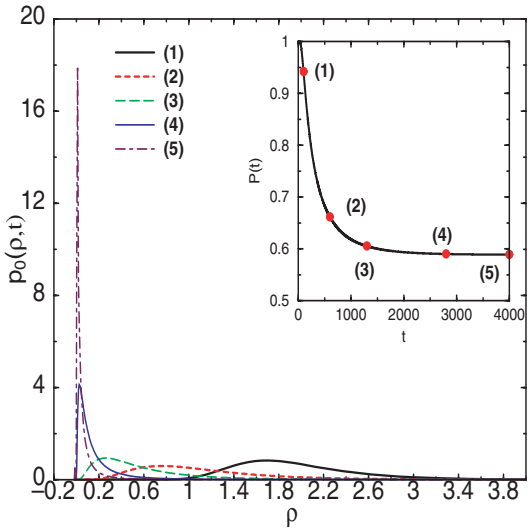


FIG. 4.8. *Transition probability density $p_0(\rho, t)$ versus ρ for several times. The inset shows the corresponding values of the survival probability for these times. The parameters are as in Figure* 4.4.

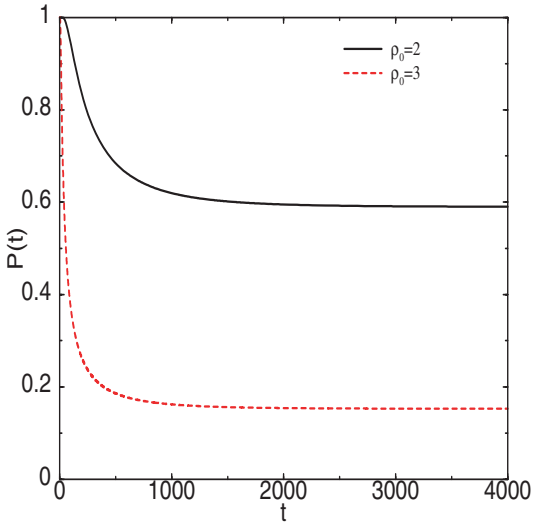


FIG. 4.9. *Time evolution of the survival probability $P(t)$ for two different initial condition data. The parameters are as in Figure* 4.4.

the Monte Carlo type on the stochastic differential equation given in (1.1)–(1.2), with the initial values $y_\epsilon(0) = \sqrt{2}$, $y'_\epsilon(0) = \sqrt{2}$. Our purpose was twofold. First, we took ϵ sufficiently small and t sufficiently large to check to what extent such a model could accurately approximate the original (ϵ-labeled process). On the other hand, the numerical results obtained by solving the stochastic differential equation have an independent interest, because their validity holds true for any size of ϵ and t. Obviously, whenever ϵ is not sufficiently small and/or t is not sufficiently large, we can expect that the resulting functionals of the limiting-process might depart even significantly from those computed on the basis of the stochastic differential equation.

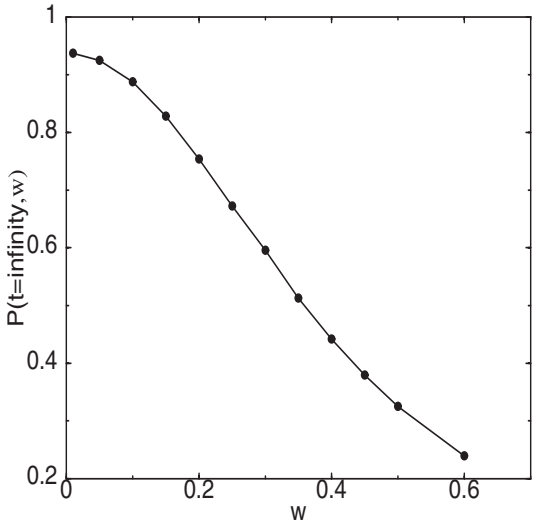


FIG. 4.10. *Long-time value of $p_0(\rho, t)$ obtained in the limit of $t \to \infty$ versus the nonlinearity parameter w. The parameters are as in Figure* 4.4.

The Ito-type system can be written as

$$\begin{cases} d\rho = \left[\epsilon F_1(t;\rho,\varphi;\mu,\nu) + \epsilon^2 G_1(t;\rho,\varphi;\mu,\nu)\right] dt, \\ d\varphi = \left[\epsilon F_2(t;\rho,\varphi;\mu,\nu) + \epsilon^2 G_2(t;\rho,\varphi;\mu,\nu)\right] dt, \\ d\mu = -\sigma_1\,\mu\,dt + \sigma_1\,dW_1, \\ d\nu = -\sigma_2\,\nu\,dt + \sigma_2\,dW_2, \end{cases} \tag{4.14}$$

where we have also displayed the dependence of the functions F_i and G_i on the stochastic processes μ and ν. Since, in the practical simulations, such processes will be colored noise processes with autocorrelations given in (4.1), the system (4.14) includes the last two equations. In fact, according to [11], the colored noise processes above can be evaluated from such Ito equations, driven by the independent standard Brownian motions W_1 and W_2. The process λ in the damping term is taken to be a constant, $\lambda(t) \equiv \lambda_0$.

While the absorbing boundary condition is imposed as a boundary condition on $\rho = R$ for the Fokker–Planck equation (2.10), here we proceed as follows. In correspondence to the first time t^*, when a given realization $\rho(t)$ attains the value R, we ignore the contribution from such a realization in the averages yielding the moments of ρ and y for $t > t^*$. This procedure yields the moments of the particle position at time t, taking into account that such a particle did not hit the absorbing barrier yet. Therefore, only the statistical properties of the particles that have survived up to time t have been included.

In principle, this numerical treatment amounts to a loss of accuracy for larger and larger times, because the sample size (i.e., the number of particles) becomes smaller, thus requiring more realizations when time increases. Numerical results, however, show that the fast decay of the moments computed (from the Fokker–Planck equation) is such that the loss of accuracy mentioned above is felt very little.

As usual, the required Monte Carlo simulations can be based on the generation of sequences of (pseudo-) random numbers, which is routine. Here we chose instead to

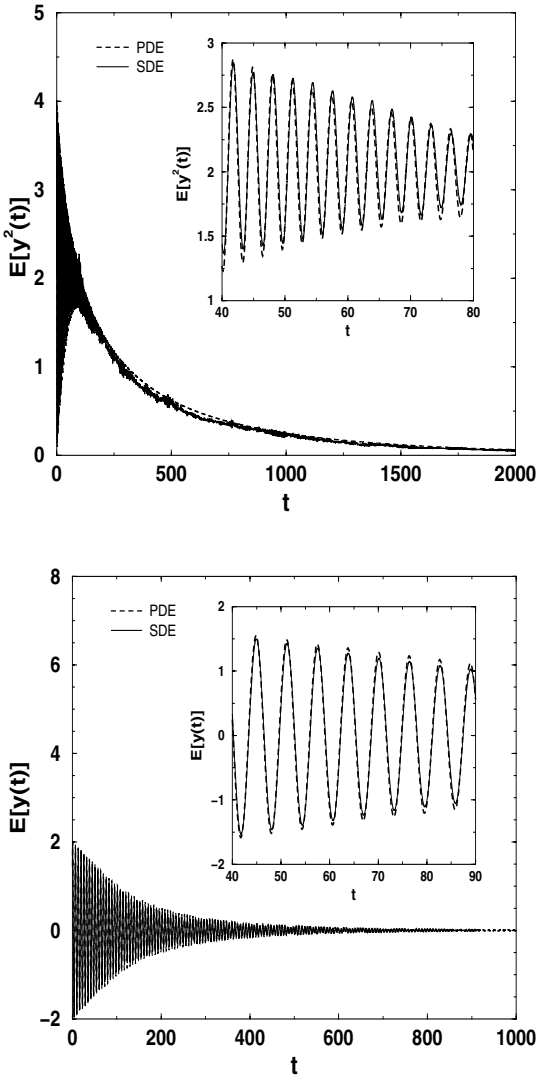


FIG. 4.11. *Time evolution of the first and second moments of the oscillator's displacement, y, keeping w to the fixed value* 0.3. *The other parameters are as in Figure* 4.1.

use the so-called sequences of *quasi-random* numbers (deterministic low discrepancy sequences); see [2, 17]. Such a choice provides a higher accuracy for a given number, N, of realizations, typically involving a deterministic error of order $\mathcal{O}(N^{-1}\log^{d^*-1}N)$ (where d^* is an "effective dimension"; see [1]) instead of the statistical error of order $\mathcal{O}(N^{-1/2})$. Quasi-random number sequences are rather delicate to exploit but have been used successfully in a number of applications; see [2, 3, 14, 15, 18]. Application of quasi-random numbers to the numerical solution of stochastic differential equations has been shown in general to be very inefficient; see [10]. In such a paper, however, no "scrambling" strategy was adopted, which action was actually shown to be important in [14, 15]. Here we have implemented a quasi-Monte Carlo algorithm with a reordering technique, as done in [15], and thus our results turned out to be highly accurate. A more general account of a successful implementation of quasi-random sequences to

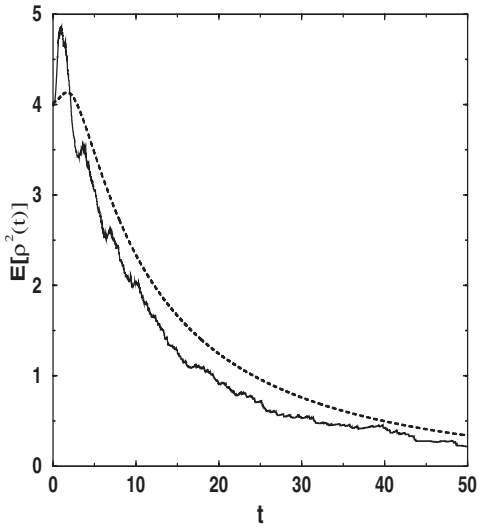


FIG. 4.12. *Time evolution of the expected value of* ρ *for* $\epsilon = 0.5$ *and* $w = 0.3$. *The other parameters are as in Figure* 4.1.

solve stochastic differential equations has been presented in [1].

To solve the system of stochastic differential equations in (4.14), we have implemented a "weak scheme" of order 2 (which reduces to the Heun scheme in the deterministic case); see [13, Chap. 15, sect. 15.1] and [8, 22]. The time-step size used was $\Delta t = 0.01$, with $N = 1000$ realizations. We used Halton sequences of quasi-random numbers [2], being aware that a different choice of numbers, such as Sobol' or Faure numbers, was shown earlier to be essentially irrelevant [16]. In order to simulate random paths to solve the system in (4.14) it is required to generate as many random variables as time steps, which amounts to generating *high-dimensional* quasi-random sequences. This is a sensible issue, in that high-dimensional quasi-random sequences may fail to work appropriately because unwanted correlations can be easily introduced. Such a difficulty was removed in our algorithm by introducing a suitable reordering, as was done in [15, 14]; see also [1]. Using reordering, the effective dimension of the system reduces to the number of dependent variables, ρ and φ. Note that the truncation error will be of order $\mathcal{O}((\Delta t)^2) \approx 10^{-4}$, which is negligible compared to the quasi-Monte Carlo error, which is of order $\mathcal{O}(N^{-1} \log N) \approx 0.9 \times 10^{-3}$.

In Figures 4.1 and 4.2, the continuous line shows the time evolution of the mean value of the oscillator amplitude ρ and of its square ρ^2, and the survival probability, with $\epsilon = 0.1$ and $w = 0.3$. Therefore, a comparison is made with the solution obtained in the diffusion limit. The agreement is very good for all times and also improves when time increases, while instead it gets a little worse in Figure 4.4.

In Figure 4.12, the comparison is made with $\epsilon = 0.5$. All these plots show to what extent the diffusion limit provides a good approximation of the original problem in (1.1)–(1.2). In any case, even when ϵ is not very small, the qualitative agreement gets better for time sufficiently large. In Figure 4.11, the mean values of the oscillator's displacement, y, and of y^2 are depicted versus t (solid line). Again, they are compared with the corresponding results obtained through the Fokker–Planck equation (dashed line).

5. Summary and concluding remarks. In closing, we stress the high points of the paper. A model for a Duffing oscillator with random parameters and subject to an absorbing boundary condition has been analyzed in the diffusion limit, on a scale where the nonlinear term is much stronger than in the model analyzed in [23]. A higher-order nonlinear effect, such as the appearance of second harmonics effects, can be observed here.

A numerical solution of the stochastic differential equation has been obtained, for small fixed ϵ, both for the purpose of validating the diffusion limit theory (when ϵ is sufficiently small and t sufficiently large) and because the problem for finite ϵ and t is meaningful in itself. The dependence on the nonlinearity strength w and on the location R of the absorbing boundary has also been investigated, mostly numerically. We have also shown that quasi-random sequences of numbers can be efficiently used in the Monte Carlo simulation, provided that a reordering strategy is adopted.

It is natural to ask whether the problem in (1.1)–(1.2) is meaningful when no absorbing boundary condition is imposed. The (formal) diffusion problem, say $\mathcal{P}_\infty$, is described by (2.10)–(2.11) on the unbounded space domain $(\rho, \varphi) \in (0, +\infty) \times (0, 2\pi)$. Suppose that we are interested in computing the moments of ρ, among them the mean energy $E[\rho^2]$. According to (4.12), this requires computing $p_0(\rho, \tau_1)$, the solution to (3.7) with $m = 0$, subject to the initial value $p_0(\rho, 0) = \delta(\rho - \rho_0)/2\pi$ on the domain $0 < \rho < +\infty$. It turns out that, according to Feller's classification sketched in section 3, the boundary point $\rho = +\infty$ is an entrance point. Note that in the model studied in [23] the point $\rho = +\infty$ was of a different type, namely, a natural boundary (as for the heat equation). In [6, 7], it was established that when one of the boundaries is natural while the other is an entrance point, no boundary condition should be imposed (on the latter point) and uniqueness is lost. More precisely, several solutions to the problem $\mathcal{P}_\infty$ exist, but only one is characterized by being positive and norm-preserving. Therefore, such a solution is a candidate to be a probability density function. Moreover, such a solution enjoys the property of having its value as well as its flux equal to zero at the same time, at $\rho = +\infty$, so that this boundary can be simultaneously considered an absorbing and a reflecting boundary. All the other solutions, characterized by an arbitrary value of the flux, are instead either negative or norm-increasing.

In order to compute numerically such a solution, $p_0(\rho, \tau_1)$, the problem $\mathcal{P}_\infty$ has to be approximated by the problem $\mathcal{P}_{\rho_{\max}}$, obtained by cutting the unbounded domain to the bounded domain $(0, \rho_{\max})$. Since the boundary $\rho = \rho_{\max}$ now becomes regular, a boundary condition can be imposed on it. Recalling that the unique solution to $\mathcal{P}_\infty$, i.e., that one which is positive and norm-preserving, vanishes along with its space derivative at $\rho = +\infty$, either of these two conditions at $\rho = \rho_{\max}$ can be imposed, provided that $\rho_{\max}$ is sufficiently large. Approximating such a solution to $\mathcal{P}_\infty$ (which has an entrance boundary) by the solutions to problems like $\mathcal{P}_{\rho_{\max}}$ (which has a regular boundary) is a quite delicate task. In fact, on the one hand, a homogeneous Neumann condition is known to preserve the norm of the solution (see (2.16)). On the other hand, as was mentioned before, the solutions to $\mathcal{P}_\infty$ are very sensitive with respect to small departures from the zero value of the space derivative. Indeed one would obtain either negative or norm-increasing solutions. Imposing instead the Dirichlet condition $p_0(\rho_{\max}, \tau_1) = 0$, a much more stable behavior would be observed, but the norm of the solution would not be preserved in this case.

REFERENCES

[1] J. A. Acebrón and R. Spigler, *Fast simulations of stochastic dynamical systems*, J. Comput. Phys., 208 (2005), pp. 106–115.

[2] R. E. Caflisch, *Monte Carlo and quasi-Monte Carlo methods*, in Acta Numerica, 1998, Acta Numer. 7, Cambridge University Press, Cambridge, UK, 1998, pp. 1–49.

[3] R. E. Caflisch, W. Morokoff, and A. B. Owen, *Valuation of mortgage-backed securities using Brownian bridges to reduce effective dimension*, J. Comput. Finance, 1 (1997), pp. 27–46.

[4] H. Cramér and M. R. Leadbetter, *Stationary and Related Stochastic Processes. Sample Function Properties and Their Applications*, Wiley, New York, 1967.

[5] M. Di Paola, ed., special issue, Meccanica, 37 (2002).

[6] W. Feller, *Two singular diffusion problems*, Ann. of Math. (2), 54 (1951), pp. 173–182.

[7] W. Feller, *The parabolic differential equations and the associated semi-groups of transformations*, Ann. of Math. (2), 55 (1952), pp. 468–519.

[8] D. J. Higham, *An algorithmic introduction to numerical simulation of stochastic differential equations*, SIAM Rev., 43 (2001), pp. 525–546.

[9] E. Hille, *Les probabilités continues en chaîne*, C. R. Acad. Sci. Paris, 230 (1950), pp. 34–35.

[10] N. Hofmann and P. Mathé, *On quasi-Monte Carlo simulation of stochastic differential equations*, Math. Comp., 66 (1997), pp. 573–589.

[11] R. L. Honeycutt, *Stochastic Runge-Kutta algorithms.* II. *Colored noise*, Phys. Rev. A (3), 45 (1992), pp. 604–610.

[12] R. Z. Khas'minskii, *A limit theorem for the solutions of differential equations with random right-hand sides*, Theory Probab. Appl., 11 (1966), pp. 390–406.

[13] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer, Berlin, 1999.

[14] C. Lecot and F. El Khettabi, *Quasi-Monte Carlo simulation of diffusion*, J. Complexity, 15 (1999), pp. 342–359.

[15] W. J. Morokoff and R. E. Caflisch, *A quasi-Monte Carlo approach to particle simulation of the heat equation*, SIAM J. Numer. Anal., 30 (1993), pp. 1558–1573.

[16] W. J. Morokoff and R. E. Caflisch, *Quasi-random sequences and their discrepancies*, SIAM J. Sci. Comput., 15 (1994), pp. 1251–1279.

[17] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 63, SIAM, Philadelphia, PA, 1992.

[18] S. Ogawa and C. Lécot, *A quasi-random walk method for one-dimensional reaction-diffusion equations*, Math. Comput. Simulation, 62 (2003), pp. 487–494.

[19] G. C. Papanicolaou, *Wave propagation in a one-dimensional random medium*, SIAM J. Appl. Math., 21 (1971), pp. 13–18.

[20] G. Papanicolaou and J. B. Keller, *Stochastic differential equations with applications to random harmonic oscillators and wave propagation in random media*, SIAM J. Appl. Math., 21 (1971), pp. 287–305.

[21] G. Papanicolaou and W. Kohler, *Asymptotic theory of mixing stochastic ordinary differential equations*, Comm. Pure Appl. Math., 27 (1974), pp. 641–668.

[22] E. Platen, *An introduction to numerical methods for stochastic differential equations*, in Acta Numerica, 1999, Acta Numer. 8, Cambridge University Press, Cambridge, UK, 1999, pp. 197–246.

[23] R. Spigler, *A stochastic model for nonlinear oscillators of Duffing type*, SIAM J. Appl. Math., 45 (1985), pp. 990–1005.

[24] R. L. Stratonovich, *Topics in the Theory of Random Noise*, Vols. 1 and 2, Gordon and Breach, New York, 1963.

[25] J. J. Stoker, *Nonlinear Vibrations in Mechanical and Electrical Systems*, Wiley-Interscience, New York, 1950.

[26] A. M. Yaglom, *An Introduction to the Theory of Stationary Random Functions*, Prentice–Hall, Englewood Cliffs, NJ, 1962.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 286–308

THE KELLER–SEGEL MODEL WITH LOGISTIC SENSITIVITY FUNCTION AND SMALL DIFFUSIVITY*

YASMIN DOLAK† AND CHRISTIAN SCHMEISER‡

Abstract. The Keller–Segel model is the classical model for chemotaxis of cell populations. It consists of a drift-diffusion equation for the cell density coupled to an equation for the chemoattractant. Here a variant of this model is studied in one-dimensional position space, where the chemotactic drift is turned off for a limiting cell density by a logistic term and where the chemoattractant density solves an elliptic equation modeling a quasi-stationary balance of reaction and diffusion with production of the chemoattractant by the cells. The case of small cell diffusivity is studied by asymptotic and numerical methods. On a time scale characteristic for the convective effects, convergence of solutions to weak entropy solutions of the limiting nonlinear hyperbolic conservation law is proven. Numerical and analytic evidence indicates that solutions of this problem converge to irregular patterns of cell aggregates separated by entropic shocks from vacuum regions as time tends to infinity. Close to each of these patterns an "almost" stationary solution of the full parabolic problem can be constructed up to an exponentially small (in terms of the cell diffusivity) residual. Based on a metastability hypothesis, the methods of exponential asymptotics are used to derive systems of ordinary differential equations approximating the long-time behavior of the parabolic problem on exponentially large time scales. The observed behavior is a coarsening process reminiscent of phase change models. A hybrid asymptotic-numerical approach for the simulation of the system is introduced, and the accuracy of this new approach is shown by comparison to numerical simulations of the full problem.

Key words. chemotaxis, hyperbolic limit, entropy solution, exponential asymptotics

AMS subject classifications. 35K55, 35K65, 35B40

DOI. 10.1137/040612841

1. Introduction.

Chemotaxis, the active motion of organisms influenced by chemical gradients, has been studied both experimentally and theoretically by a large number of authors. The first mathematical model for chemotaxis was derived by Patlak [11] and by Keller and Segel [6]. In its most widely used formulation, the cell density $\varrho(x,t)$ at position $x \in \mathbb{R}^n$ and time $t > 0$ solves the convection diffusion equation

$$\partial_t \varrho + \nabla \cdot (\chi(\varrho, S)\varrho \nabla S - D(\varrho, S)\nabla \varrho) = 0. \tag{1.1}$$

This equation is coupled to an equation for the chemical concentration $S(x,t)$, typically a parabolic or elliptic equation with a reaction term describing production and degradation of the chemoattractant.

The Keller–Segel model has been applied to many different problems, ranging from bacterial chemotaxis to cancer growth or the immune response. For some applications, it turns out that the diffusivity of cells plays only a minor role. In Dolak and Schmeiser [3], a convection equation with a small diffusion term as higher order

*Received by the editors August 4, 2004; accepted for publication (in revised form) May 10, 2005; published electronically November 15, 2005. This research was supported financially by the Austrian Science Foundation, grants W008 and P16174-N05, and by the European HYKE network.
http://www.siam.org/journals/siap/66-1/61284.html
†Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Altenbergerstr. 69, 4040 Linz, Austria (yasmin.dolak@oeaw.ac.at).
‡Faculty of Mathematics, University of Vienna, Nordbergstr. 15, 1090 Vienna, Austria (christian.schmeiser@univie.ac.at).

correction is derived from a kinetic model for chemotaxis. Taking this case as a motivation, we will study a Keller–Segel model with a small diffusion constant and its limit of vanishing diffusivity. More precisely, we investigate

$$\partial_t \varrho + \partial_x(\chi(\varrho)\varrho\partial_x S) = D\partial_x^2\varrho, \tag{1.2}$$

with $x \in (0, L)$ and $t > 0$. We assume the diffusion D to be constant and the chemotactic sensitivity $\chi(\varrho)$ to be of the form

$$\chi(\varrho) = \chi_0\left(1 - \frac{\varrho}{\varrho_{max}}\right), \tag{1.3}$$

the maximal cell density ϱ_{max} and χ_0 being positive constants. Thus, the chemotactic response of the cells is shut off when a maximal density is reached. Models of this type were first investigated by Hillen and Painter in [5]. In [10], the authors derive a chemotaxis model comprising a chemotactic sensitivity of the above form from a master equation describing a random walk on a one-dimensional lattice, by taking into account the finite size of cells.

The evolution of the chemoattractant S is described by

$$\partial_x^2 S = \beta S - \alpha\varrho. \tag{1.4}$$

This elliptic equation, instead of the more frequently used parabolic equation, is appropriate if we assume that the diffusion rate of the chemoattractant is large in relation to the characteristic time and length scales of the problem.

We nondimensionalize (1.2) and (1.4) by choosing reference values for time, length, cell density, and the chemical concentration, respectively:

$$x_0 = \frac{1}{\sqrt{\beta}}, \quad t_0 = \frac{1}{\alpha\chi_0\varrho_{max}}, \quad \varrho_0 = \varrho_{max}, \quad S_0 = \frac{\alpha\varrho_{max}}{\beta}.$$

By introducing the dimensionless quantities

$$\bar{x} = \frac{x}{x_0}, \quad \bar{t} = \frac{t}{t_0}, \quad \bar{\varrho} = \frac{\varrho}{\varrho_0}, \quad \text{and} \quad \bar{S} = \frac{S}{S_0}$$

and immediately dropping the bars, we obtain the nondimensionalized system

$$\partial_t \varrho + \partial_x(\varrho(1-\varrho)\partial_x S) = \varepsilon\partial_x^2\varrho, \tag{1.5}$$

$$\partial_x^2 S = S - \varrho. \tag{1.6}$$

The only remaining dimensionless parameter is now

$$\varepsilon = \frac{D\beta}{\alpha\chi_0\varrho_{max}},$$

and in the following, we will shall assume $\varepsilon \ll 1$. This corresponds to a situation where the cells react to chemotactic signals strongly enough such that their velocity distribution is significantly biased towards the chemoattractant gradient, as opposed to the case where unbiased reorientation dominates the behavior of individual cells. In the relevant situation, a small diffusion term can be derived as a correction to the purely convective macroscopic limit of a kinetic transport model (see, e.g., [3], as mentioned above).

The initial condition is given by

$$\varrho(x,0) = \varrho_I^\varepsilon. \tag{1.7}$$

We choose homogeneous Neumann boundary conditions, i.e.,

$$\partial_x \varrho(0,t) = \partial_x \varrho(L,t) = 0, \quad \partial_x S(0,t) = \partial_x S(L,t) = 0. \tag{1.8}$$

In the next section, we will analyze the limit $\varepsilon \to 0$ of system (1.5), (1.6). By deriving estimates which are uniformly valid for $\varepsilon > 0$, we will, by a compactness argument, show convergence of ϱ and S to entropy solutions of the corresponding hyperbolic system,

$$\partial_t \bar{\varrho} + \partial_x(\bar{\varrho}(1-\bar{\varrho})\partial_x \bar{S}) = 0, \tag{1.9}$$

$$\partial_x^2 \bar{S} = \bar{S} - \bar{\varrho}, \tag{1.10}$$

with

$$\partial_x \bar{S}(0,t) = \partial_x \bar{S}(L,t) = 0 \tag{1.11}$$

and subject to the initial condition

$$\bar{\varrho}(x,0) = \bar{\varrho}_I. \tag{1.12}$$

As a consequence of (1.11), the characteristics of (1.9) are parallel to the boundary, and no boundary conditions for $\bar{\varrho}$ are needed.

In sections 3 and 4, we study the long-time behavior of solutions of the hyperbolic and the full parabolic system, respectively. In the latter, the formation of so-called pseudostationary or metastable states can be observed. We will use formal asymptotic methods to derive a system of ODEs describing the exponentially slow movement of these patterns. Finally, in section 5, we will investigate the long-time behavior of solutions numerically. The metastability analysis is strongly related to recent work by Potapov and Hillen [13]. However, different scaling assumptions are used there, and consequently, a direct comparison of results is not straightforward.

2. Convergence of solutions. In this section, we investigate the limit $\varepsilon \to 0$ in (1.5), (1.6), (1.7), (1.8). A similar problem from semiconductor physics is considered in Markowich and Szmolyan [8]. There, however, the nonlinearity of the flux is only due to a coupling with an electric field (the equivalent to the chemical concentration here), and the formation of shocks in the hyperbolic problem is not observed.

Equations (1.5) and (1.6) differ from the system analyzed in Hillen and Painter [5] by the fact that an elliptic instead of a parabolic equation for S is considered here. In [5] existence of an invariant region for (ϱ, S) in $\mathbb{R}^2$ and, consequently, global existence of smooth solutions is shown. In our case, the proof (based on a straightforward maximum principle) is much simpler and presented below for completeness.

We make the following assumption on the initial data:

$$0 \le \varrho_I^\varepsilon \le 1, \quad \varrho_I^\varepsilon \in W^{1,1}(0,L), \quad \text{uniformly in } \varepsilon. \tag{A1}$$

THEOREM 2.1. *Let assumption* (A1) *hold. Then there exists a unique, global, smooth solution of* (1.5)–(1.8) *satisfying*

$$0 \le \varrho(x,t),\ S(x,t) \le 1 \quad \textit{and} \quad \int_0^L \varrho(x,t)\,dx = \int_0^L \varrho_I(x)\,dx \tag{2.1}$$

and

$$S \in L^\infty((0,\infty); W^{2,\infty}(0,L)), \tag{2.2}$$

uniformly in ε *as* $\varepsilon \to 0$.

Proof. With Green's function

$$G(x,y) = \frac{1}{2}e^{-|x-y|} + \frac{e^{x+y} + e^{2L-x-y} + e^{x-y} + e^{y-x}}{2(e^{2L}-1)}, \tag{2.3}$$

the chemoattractant density can be computed from (1.6), (1.8) in terms of the cell density:

$$S(x,t) = \mathcal{S}[\varrho](x,t) := \int_0^L G(x,y)\varrho(y,t)dy. \tag{2.4}$$

Using this in (1.5), the resulting equation

$$\partial_t \varrho = \varepsilon \partial_x^2 \varrho - \partial_x(\varrho(1-\varrho)\partial_x \mathcal{S}[\varrho])$$

falls into the class of abstract semilinear parabolic equations. Local existence of unique smooth solutions can be shown by semigroup techniques [12]. Then global existence follows from a comparison principle: writing (1.5) as

$$\partial_t \varrho + (2\varrho - 1)\partial_x S \partial_x \varrho + \varrho(1-\varrho)(S-\varrho) = \varepsilon \partial_x^2 \varrho,$$

it follows immediately that $\varrho = 0$ and $\varrho = 1$ are lower and upper solutions, respectively. A uniform (in time) bound for S is an obvious consequence of (2.4). ∎

We continue with estimates for the derivatives of ϱ.

LEMMA 2.2. *Let assumption* (A1) *hold. Then the solution of* (1.5)–(1.8) *satisfies*

$$\varrho \in L^\infty_{loc}((0,\infty); W^{1,1}(0,L)), \text{ uniformly in } \varepsilon.$$

Proof. Differentiation of (1.5) with respect to x yields

$$\partial_t \partial_x \varrho + \partial_x((1-2\varrho)\partial_x \varrho \partial_x S + \varrho(1-\varrho)\partial_x^2 S) = \varepsilon \partial_x^3 \varrho. \tag{2.5}$$

We define an approximation of the sign function by $\sigma_\delta(z) = \sigma(z/\delta)$, $0 < \delta \ll 1$, with σ smooth and increasing, $\sigma(0) = 0$, and $\sigma(z) = \operatorname{sign} z$ for $|z| > z_0$. Then, with $\operatorname{abs}_\delta(z) := \int_0^z \sigma_\delta(\xi)d\xi$, the convergence of $\operatorname{abs}_\delta(z)$ to $|z|$ as $\delta \to 0$ is uniform in $z \in \mathbb{R}$. Multiplying (2.5) with $\sigma_\delta(\partial_x \varrho)$ and integrating with respect to x yields

$$\begin{aligned} &\int_0^L \sigma_\delta(\partial_x\varrho)\partial_t\partial_x\varrho\,dx + \int_0^L \sigma_\delta(\partial_x\varrho)\partial_x(\partial_x\varrho\partial_x S(1-2\varrho))\,dx \\ &\qquad + \int_0^L \sigma_\delta(\partial_x\varrho)\partial_x(\varrho(1-\varrho)(S-\varrho))\,dx = \varepsilon\int_0^L \sigma_\delta(\partial_x\varrho)\partial_x^3\varrho\,dx. \end{aligned} \tag{2.6}$$

We integrate (2.6) by parts. The boundary terms vanish and we obtain

$$\begin{aligned} &\frac{d}{dt}\int_0^L \operatorname{abs}_\delta(\partial_x\varrho)\,dx - \int_0^L \sigma_\delta'(\partial_x\varrho)\partial_x\varrho\partial_x^2\varrho\partial_x S(1-2\varrho)\,dx \\ &\qquad + \int_0^L \sigma_\delta(\partial_x\varrho)\partial_x(\varrho(1-\varrho)(S-\varrho))\,dx = -\varepsilon\int_0^L \sigma_\delta'(\partial_x\varrho)(\partial_x^2\varrho)^2\,dx \le 0. \end{aligned} \tag{2.7}$$

The function $f_\delta(z) = \sigma_\delta(z)z - \operatorname{abs}_\delta(z)$ satisfies $f_\delta'(z) = \sigma_\delta'(z)z$ and converges to 0 uniformly in $z \in \mathbb{R}$. We integrate the second term in (2.7) by parts, which gives

$$\begin{aligned}\frac{d}{dt}\int_0^L \operatorname{abs}_\delta(\partial_x\varrho)\,dx \le &-\int_0^L f_\delta(\partial_x\varrho)\partial_x(\partial_x S(1-2\varrho))\,dx\\ &-\int_0^L \sigma_\delta(\partial_x\varrho)\partial_x(\varrho(1-\varrho)(S-\varrho))\,dx.\end{aligned} \tag{2.8}$$

The last term can be estimated by

$$\begin{aligned}-\int_0^L \sigma_\delta(\partial_x\varrho)\partial_x(\varrho(1-\varrho)(S-\varrho))\,dx &= -\int_0^L \sigma_\delta(\partial_x\varrho)\varrho(1-\varrho)\partial_x S\,dx\\ -\int_0^L \sigma_\delta(\partial_x\varrho)\partial_x\varrho(3\varrho^2-2\varrho(S+1)+S)\,dx &\le c_1 + c_2\int_0^L |\partial_x\varrho|\,dx.\end{aligned}$$

In the limit $\delta \to 0$, the first term of the right-hand side of (2.8) vanishes, and we obtain

$$\frac{d}{dt}\int_0^L |\partial_x\varrho|\,dx \le c_1 + c_2\int_0^L |\partial_x\varrho|dx. \tag{2.9}$$

The assertion of Lemma 2.2 now follows from the Gronwall inequality. ∎

LEMMA 2.3. *Let* (A1) *hold. Then the solution of* (1.5)–(1.8) *satisfies*

$$\sqrt{\varepsilon}\partial_x\varrho, \partial_t\partial_x S \in L^2_{loc}((0,\infty)\times[0,L]), \quad \textit{uniformly in } \varepsilon.$$

Proof. We write (1.5) as

$$\partial_t\varrho = \partial_x(\varepsilon\partial_x\varrho - \varrho(1-\varrho)\partial_x S). \tag{2.10}$$

Multiplication by ϱ and integration with respect to x leads to

$$\frac{1}{2}\frac{d}{dt}\int_0^L \varrho^2\,dx + \varepsilon\int_0^L (\partial_x\varrho)^2\,dx = \int_0^L \varrho(1-\varrho)\partial_x S\partial_x\varrho\,dx. \tag{2.11}$$

Since the integrand in the last term is in $L^1((0,L))$ uniformly in t and ε by the previous result, we obtain the boundedness of $\sqrt{\varepsilon}\partial_x\varrho$ by integration with respect to t. As a consequence, the flux density $J = \varrho(1-\varrho)\partial_x S - \varepsilon\partial_x\varrho$ is also uniformly bounded in $L^2_{loc}((0,\infty)\times[0,L])$. Differentiating (1.6) with respect to x and t and using $\partial_t\varrho + \partial_x J = 0$, we obtain

$$\partial_t\partial_x^3 S - \partial_t\partial_x S = \partial_x^2 J.$$

Thus, $\partial_t\partial_x S = -\mathcal{S}[\partial_x^2 J]$. Since the expression on the right-hand side is a bounded operator applied to $J \in L^2_{loc}((0,\infty)\times[0,L])$, the proof is complete. ∎

THEOREM 2.4. *Let the assumption* (A1) *hold,* (ϱ, S) *be a solution of* (1.5)–(1.8)*, and* $T > 0$*. Then, as* $\varepsilon \to 0$ *(restricting to subsequences),*

$$\varrho \to \bar\varrho \text{ in } C([0,T];L^1((0,L))) \quad \textit{and} \quad S \to \bar S \text{ in } C([0,T];C^1([0,L])). \tag{2.12}$$

The limit $(\bar\varrho, \bar S) \in L^\infty((0,T);BV((0,L))\times W^{2,\infty}((0,L)))$ *solves* (1.9), (1.10), (1.11)*, where* $\bar\varrho_I \in BV((0,L))$ *is an accumulation point of* ϱ_I^ε*. Moreover,* $\bar\varrho$ *is an entropy solution of* (1.9)*; i.e.,*

$$\partial_t\eta(\bar\varrho) + \partial_x(\psi(\bar\varrho)\partial_x\bar S) + (\bar\varrho(1-\bar\varrho)\eta'(\bar\varrho) - \psi(\bar\varrho))(\bar S - \bar\varrho) \le 0 \tag{2.13}$$

holds in the weak sense for every smooth, convex η *and with* $\psi'(\bar\varrho) = (1-2\bar\varrho)\eta'(\bar\varrho)$.

Remark. Note that the entropy inequality does not give rise to a decaying entropy functional.

Proof. The boundedness of the flux density (proof of Lemma 2.3) gives $\partial_t \varrho \in L^2((0,T);H^{-1}((0,L)))$. Together with Lemma 2.2 this implies that ϱ is in a compact set in $C([0,T];L^1((0,L)))$ (see Simon [15]). From Theorem 2.1, Lemma 2.3, and an anisotropic generalization of the Sobolev embedding of $W^{1,p}$ in $C^{0,1-n/p}$, $p>n$ (see Haskovec and Schmeiser [4]), it follows that $\partial_x S$ is uniformly bounded in $C^{0,1/3}([0,T]\times\bar{\Omega})$, $T>0$. An application of the Arzela–Ascoli theorem concludes the proof of (2.12). The strong convergence of ϱ and $\partial_x S$ allows us to pass to the limit in the weak formulation of (1.5)–(1.8), giving the weak formulation of (1.9)–(1.11) for $\bar{\varrho}$ and $\bar{S}$. The entropy inequality (2.13) follows analogously. ∎

3. Long-time behavior of the hyperbolic system. In this section, we investigate the stability and the asymptotic behavior of entropy solutions of the hyperbolic system. Stationary solutions of (1.9)–(1.11) satisfy

$$\bar{\varrho}(1-\bar{\varrho})\partial_x \bar{S} = 0, \tag{3.1}$$

$$\partial_x^2 \bar{S} = \bar{S} - \bar{\varrho}. \tag{3.2}$$

It can be immediately seen that $\bar{\varrho} = \bar{S} = const$ is a solution.

LEMMA 3.1. *The constant solution, $\bar{\varrho} = \bar{S} = \frac{m}{L}$, where $0 < m < L$ is the total mass, of system* (1.9)–(1.11) *is unstable.*

Proof. We multiply (1.9) by $\bar{S}$ and differentiate (1.10) with respect to t to obtain

$$\frac{1}{2}\frac{d}{dt}\int_0^L (\bar{S}^2 + (\partial_x \bar{S})^2)dx = \int_0^L \bar{\varrho}(1-\bar{\varrho})(\partial_x \bar{S})^2 dx. \tag{3.3}$$

For small nonconstant perturbations, the right-hand side of this equation is positive $\forall t$, and hence $\int_0^L (\bar{S}^2 + (\partial_x \bar{S})^2)dx$ is increasing in time. We rearrange this integral by writing

$$\begin{aligned}\int_0^L (\bar{S}^2 + (\partial_x \bar{S})^2)dx &= \int_0^L \left[\Big(\frac{m}{L} + \bar{S} - \frac{m}{L}\Big)^2 + (\partial_x \bar{S})^2\right] dx \\ &= \frac{m^2}{L} + 2\int_0^L \frac{m}{L}\Big(\bar{S} - \frac{m}{L}\Big)dx + \int_0^L \left[\Big(\bar{S} - \frac{m}{L}\Big)^2 + (\partial_x \bar{S})^2\right] dx. \end{aligned} \tag{3.4}$$

Since the total mass is conserved, we consider only perturbations with mass 0. Thus, we have $\int_0^L \bar{S}\,dx = \int_0^L \bar{\varrho}\,dx = m\ \forall t$, and the second term on the right-hand side vanishes. Hence,

$$\min_{\int \bar{S}dx = m} \int_0^L \Big(\bar{S}^2 + (\partial_x \bar{S})^2\Big)dx = \frac{m^2}{L},$$

which is only achieved for $\bar{S} = \frac{m}{L}$. As the integral on the left-hand side is increasing in time, Lemma 3.1 follows. ∎

LEMMA 3.2. *As $t\to\infty$, $\bar{\varrho}(1-\bar{\varrho})(\partial_x \bar{S})^2 \to 0$ in the following sense:*

$$\int_\tau^\infty \int_0^L \bar{\varrho}(1-\bar{\varrho})(\partial_x \bar{S})^2 dx\,dt \xrightarrow{\tau\to\infty} 0. \tag{3.5}$$

Proof. Integration of (3.3) from $t = 0$ to ∞ shows that

$$\int_0^\infty \int_0^L \bar{\varrho}(1-\bar{\varrho})(\partial_x \bar{S})^2 dx dt < \infty, \tag{3.6}$$

which implies the assertion. ∎

From this, and the steady state equations (3.1), (3.2), we expect convergence to piecewise constant steady states, with $\bar{\varrho} = 0$, $\bar{\varrho} = 1$, or $\bar{S}_x = 0$. Going back to the time-dependent problem (1.9), (1.10) and applying the method of characteristics, we find that along characteristics given by $\dot{x} = (1 - 2\bar{\varrho})\partial_x \bar{S}$, $\bar{\varrho}$ evolves according to $\dot{\bar{\varrho}} = (\bar{\varrho} - \bar{S})\bar{\varrho}(1 - \bar{\varrho})$. It immediately follows that $\bar{\varrho} = \bar{S} = const$, with $0 < const < 1$, is unstable. If $\bar{\varrho}$ gets sufficiently small such that $\bar{S} > \bar{\varrho}$, then $\bar{\varrho} = 0$ is attracting, and a similar argument holds for $\bar{\varrho} = 1$. Hence, we expect solutions to approach (as $t \to \infty$) functions of the form

$$\bar{\varrho}_\infty(x) = \frac{1 - (-1)^{k_i}}{2} \qquad \text{for } x_i < x < x_{i+1}, \tag{3.7}$$

with $0 = x_0 < x_1 < \cdots < x_{M+1} = L$, $k_i = k_0 + i$, and

$$\bar{S}_\infty = \mathcal{S}[\bar{\varrho}_\infty]. \tag{3.8}$$

Plateaus, where $\bar{\varrho}_\infty = 1$, alternate with vacuum regions ($\bar{\varrho}_\infty = 0$). At the left, it starts with a plateau for $k_0 = 1$, or with a vacuum region for $k_0 = 0$. The union of all plateau regions is denoted by

$$P = \bigcup_{k_i \text{ odd}} (x_i, x_{i+1}).$$

It follows from mass conservation that

$$l(P) = \sum_{k_i \text{ odd}} (x_{i+1} - x_i) = \int_0^L \bar{\varrho}_I \, dx.$$

Not all possible stationary solutions $\bar{\varrho}_\infty$ are indeed entropy solutions. For scalar conservation laws the sign of the jump of the density $\bar{\varrho}_\infty$ at an entropic shock is related to the convexity behavior of the flux $\bar{\varrho}_\infty(1 - \bar{\varrho}_\infty)\partial_x \bar{S}_\infty$. This leads to a condition on the sign of $\partial_x \bar{S}_\infty$ at the plateau edges:

$$(-1)^{k_i} \bar{S}_{\infty,x}(x_i) < 0, \quad 1 \le i \le M. \tag{3.9}$$

Formally, $\partial_x \bar{S}_\infty$ would also be allowed to be zero. Such a solution would, however, be unstable, since a small perturbation would lead to a violation of the entropy condition. A possible derivation of (3.9) is given in the following section by the construction of shock profiles for the full parabolic problem, i.e., boundary layer solutions smoothing the jumps of $\bar{\varrho}_\infty$.

Next, we investigate the stability of the stationary solution $(\bar{\varrho}_\infty, \bar{S}_\infty)$ with respect to a particular class of perturbations. We introduce the initial data

$$\bar{\varrho}_I(x) = \frac{1 - (-1)^{k_i}}{2} + \varepsilon u_I(x) \qquad \text{for } x \in (x_i + \varepsilon\xi_i(0), x_{i+1} + \varepsilon\xi_{i+1}(0)) = I_i(0), \tag{3.10}$$

where u_I is a piecewise smooth function and $|\varepsilon| \ll 1$. Then, solutions of (1.9), (1.10) have jumps at $x_i + \varepsilon\xi_i(t)$, and

(3.11)
$$\bar{\varrho}(x,t) = \frac{1-(-1)^{k_i}}{2} + \varepsilon u(x,t) \qquad \text{for } x \in (x_i + \varepsilon\xi_i(t), x_{i+1} + \varepsilon\xi_{i+1}(t)) = I_i(t).$$

The Rankine–Hugoniot jump condition reads

$$\varepsilon\dot{\xi}_i[\bar{\varrho}] = [\bar{\varrho}(1-\bar{\varrho})]\partial_x\bar{S}|_{x=x_i+\varepsilon\xi_i},$$

which, at leading order, yields

$$\dot{\xi}_i(t) = -\left(u(x_i+,t) + u(x_i-,t)\right)\bar{S}_{\infty,x}(x_i).$$

Using (3.11) in (1.9), it follows that $u(x,t)$ approximately satisfies

$$\begin{aligned} &\partial_t u - \partial_x(u\partial_x\bar{S}_\infty) = 0 \qquad \text{in } P,\\ &\partial_t u + \partial_x(u\partial_x\bar{S}_\infty) = 0 \qquad \text{in } Z := (0,L)\setminus P. \end{aligned}$$

By the method of characteristics, we derive

$$\dot{x} = -\partial_x\bar{S}_\infty, \quad \dot{u} = u\partial_x^2\bar{S}_\infty = u(\bar{S}_\infty - 1) \qquad \text{in } P,$$

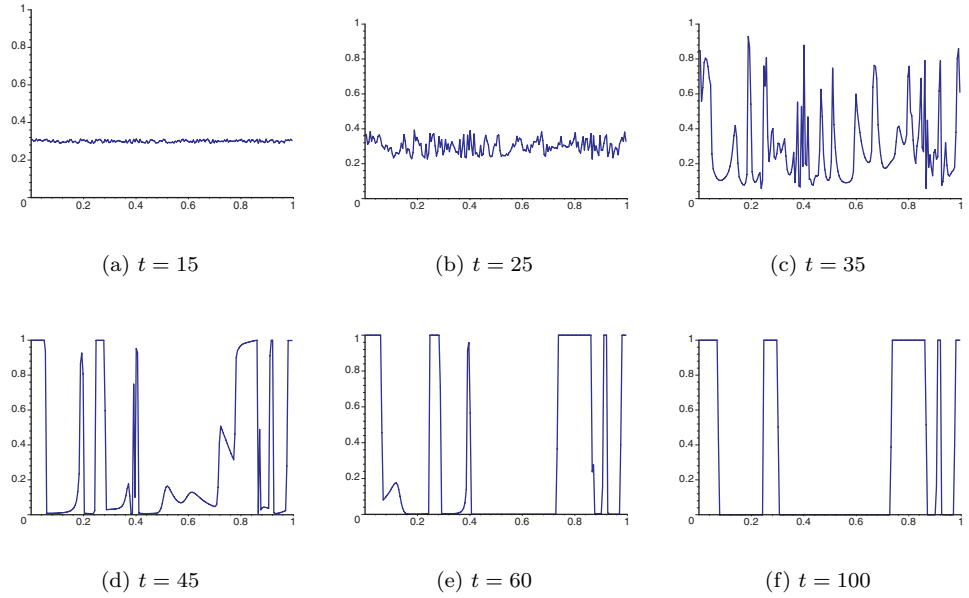
$$\dot{x} = \partial_x\bar{S}_\infty, \quad \dot{u} = -u\partial_x^2\bar{S}_\infty = -u\bar{S}_\infty \qquad \text{in } Z.$$

Since $\bar{S}_\infty$ is concave in P and convex in Z, exactly one extremum $\bar{x}_{i+1/2}$ exists between x_i and x_{i+1} for $1 \le i \le M-1$. Since the derivative of $\bar{S}_\infty$ also vanishes at the boundary points, we introduce the notation $\bar{x}_{1/2} := 0$, $\bar{x}_{M+1/2} = L$. All of the characteristics except those starting at $\bar{x}_{i+1/2}$, $0 \le i \le M$, go into one of the x_i, and u decays along characteristics. The limit of the length of the plateau $I_i(t)$ (k_i odd) as $t \to \infty$ is given by

$$\begin{aligned} l(I_i(\infty)) &= l(I_i(0)) + \varepsilon\int_0^\infty \left(\dot{\xi}_{i+1} - \dot{\xi}_i\right)dt\\ &= l(I_i(0)) + \varepsilon\int_0^\infty [-(u\partial_x\bar{S}_\infty)(x_{i+1}+) - (u\partial_x\bar{S}_\infty)(x_{i+1}-)\\ &\qquad + (u\partial_x\bar{S}_\infty)(x_i+) + (u\partial_x\bar{S}_\infty)(x_i-)]dt\\ &= l(I_i(0)) - \varepsilon\int_0^\infty\int_{x_i}^{x_{i+1}} \partial_x(u\partial_x\bar{S}_\infty)dxdt + \varepsilon\int_0^\infty\int_{x_{i+1}}^{\bar{x}_{i+3/2}} \partial_x(u\partial_x\bar{S}_\infty)dxdt\\ &\quad + \varepsilon\int_0^\infty\int_{\bar{x}_{i-1/2}}^{x_i} \partial_x(u\partial_x\bar{S}_\infty)dxdt\\ &= l(I_i(0)) - \varepsilon\left(\int_{x_i}^{x_{i+1}} u\,dx + \int_{x_{i+1}}^{\bar{x}_{i+3/2}} u\,dx + \int_{\bar{x}_{i-1/2}}^{x_i} u\,dx\right)_{t=0}^{\infty}. \end{aligned}$$

Since $u \xrightarrow{t\to\infty} 0$, we obtain

$$l(I_i(\infty)) = l(I_i(0)) + \varepsilon\int_{\bar{x}_{i-1/2}}^{\bar{x}_{i+3/2}} u_I dx. \tag{3.12}$$



(a) $t = 15$ (b) $t = 25$ (c) $t = 35$

(d) $t = 45$ (e) $t = 60$ (f) $t = 100$

FIG. 3.1. *Temporal evolution of the cell density* $\bar{\varrho}$*, starting from random initial data* $\bar{\varrho}_I \in [0.3, 0.31]$ *and with* $L = 1$.

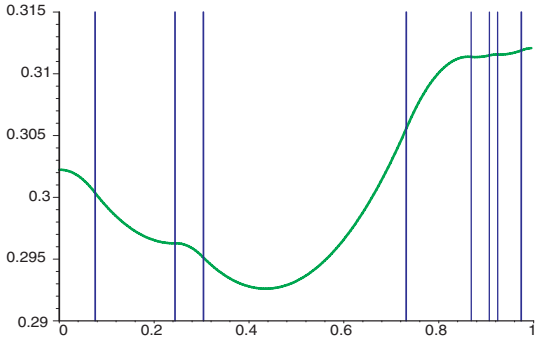


FIG. 3.2. *Cell density* $\bar{\varrho}$ *(dark) and chemical concentration* $\bar{S}$ *(light) at* $t = 100$.

Thus, as $t \to \infty$, each plateau attracts all the mass initially distributed between the neighboring minima of $\bar{S}_\infty$. We can interpret this as a neutral stability of steady states with alternating plateau and vacuum regions with respect to perturbations of the type (3.10).

In Figures 3.1 and 3.2, we solved the problem (1.9)–(1.11) numerically. At each time step, first the new chemical concentration is calculated from the old cell density, and then the cell density is updated using an upwind method. In Figure 3.1, we can observe the formation of shocks and rarefaction waves, until, in the last picture, the stationary state is reached and no further movement of the plateaus can be observed. In Figure 3.2, the corresponding chemical concentration $\bar{S}$ is shown. Note that, as discussed above, the chemical follows the course of the cell density $\bar{\varrho}$, even in the case of the slim plateau on the right-hand side of the domain.

4. Long-time behavior of the parabolic system. In this section, we will be concerned with the stability and the asymptotic behavior of solutions of the full parabolic problem (1.5)–(1.8). All stationary solutions have been characterized by Potapov and Hillen in [13] by a phase plane analysis. They are the restrictions of periodic solutions of the stationary differential equations to the interval $(0, L)$.

The dynamic problem studied in [13] differs from (1.5)–(1.8) by the fact that S is given by a parabolic equation. The authors show that all stationary solutions lie on branches bifurcating from the spatially uniform stationary solution, dependent on a bifurcation parameter inversely proportional to ε. It turns out that the constant solution is linearly stable for large enough diffusivity. After a first bifurcation its stability is transferred to a bifurcating solution. For the model (1.5)–(1.8), the linear stability result can be extended to global nonlinear stability.

LEMMA 4.1. *Let assumption* (A1) *hold, and let* $\varepsilon > \frac{1}{4}$. *Then the solution of* (1.5)–(1.8) *converges to the constant stationary solution as* $t \to \infty$.

Proof. As in the proof of Lemma 3.1, we multiply (1.5) by S and differentiate (1.6) with respect to t to obtain, after integration by parts of the last term on the right-hand side,

$$\frac{1}{2}\frac{d}{dt}\int_0^L (S^2 + (\partial_x S)^2)dx = \int_0^L \varrho(1-\varrho)(\partial_x S)^2 dx - \varepsilon \int_0^L ((\partial_x S)^2 + (\partial_x^2 S)^2)dx. \tag{4.1}$$

We can estimate the left-hand side by

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_0^L (S^2 + (\partial_x S)^2)dx &= \int_0^L \varrho(1-\varrho)(\partial_x S)^2 dx - \varepsilon \int_0^L ((\partial_x S)^2 + (\partial_x^2 S)^2)dx \\ &\le \Big(\frac{1}{4} - \varepsilon\Big) \int_0^L ((\partial_x S)^2 + (\partial_x^2 S)^2)dx. \end{aligned} \tag{4.2}$$

For $\varepsilon > \frac{1}{4}$, the right-hand side of (4.2) is negative. Integration from 0 to t yields

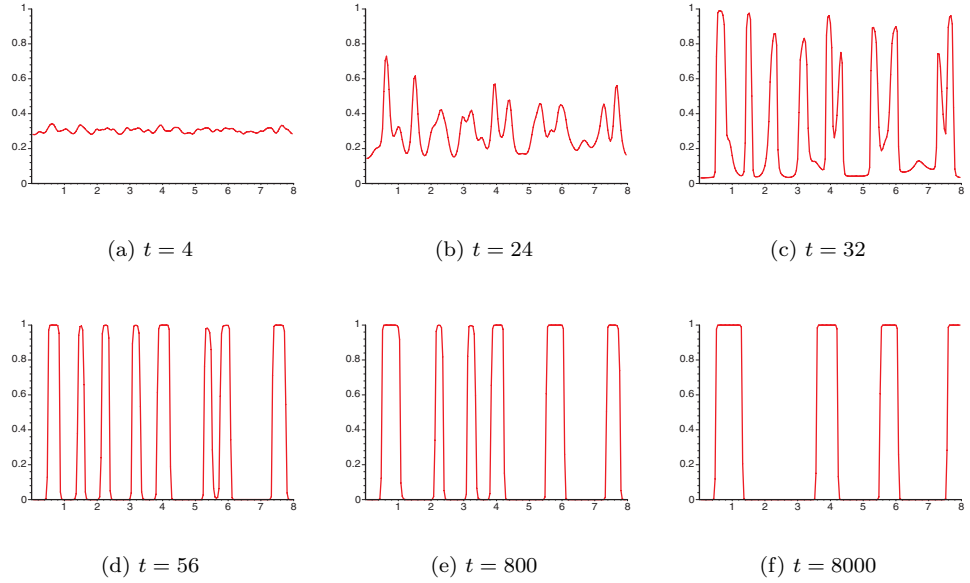
$$\int_0^L (\partial_x S)^2(x,t)\,dx \le \Big(\frac{1}{2} - 2\varepsilon\Big) \int_0^t \int_0^L (\partial_x S)^2 dx + \int_0^L (S^2 + (\partial_x S)^2)dx\Big|_{t=0}.$$

Applying Gronwall's lemma, it follows that $\|\partial_x S\|_{L^2(0,L)} \to 0$ and as a consequence, $S \to const$ as $t \to \infty$. ∎

The result is sharp in the sense that a linear stability analysis yields $\varepsilon = \frac{1}{4}$ as the first bifurcation point, where the constant steady state loses its stability.

We motivate our study of the dynamics for small values of ε by presenting the result of a numerical experiment. Figure 4.1 shows a numerical solution of (1.5)–(1.8) with $\varepsilon = 2 \times 10^{-4}$. We used the same numerical scheme as in the previous section with an explicit discretization of the diffusion term. Starting from homogeneous initial data with small perturbations, a pattern with several plateaus is formed as for the hyperbolic problem. Once this pattern has formed, it remains structurally stable for a long time, with the plateaus moving very slowly. Eventually, neighboring plateaus merge with each other. This merging process occurs on a comparatively fast time scale. The new pattern, now with one peak less, undergoes the same coarsening process.

Experimentally, this so-called metastable behavior is a well-known phenomenon in many fields, for instance solid-state physics. Mathematically, it has been studied in



(a) $t = 4$ (b) $t = 24$ (c) $t = 32$

(d) $t = 56$ (e) $t = 800$ (f) $t = 8000$

FIG. 4.1. *Numerical solution of the parabolic system* (1.5), (1.6) *with random initial data* $\varrho_I \in [0.3, 0.31]$, $L = 10$, *and* $\varepsilon = 2 \times 10^{-4}$.

various contexts such as the movement of viscous shocks [7], [14] or the Cahn–Hillard equation (for instance, [1] and [2]). A chemotaxis model featuring the formation of spike solutions is considered in [16].

The peculiar long-time behavior of system (1.5), (1.6) can be interpreted as follows. Each pseudostationary state of the parabolic system is exponentially close to a stationary entropy solution of the hyperbolic system. In contrast to the latter, however, the small diffusion allows plateaus to communicate with each other, and smaller plateaus are attracted by neighboring larger ones producing more chemoattractant. The whole phenomenon depends on a nonzero diffusion coefficient ε. Eventually, plateaus will get so close to each other that in general, the corresponding stationary solutions of the hyperbolic system cannot satisfy the entropy condition any more. Then a fast transition takes place, and the smaller plateau merges with the larger one. On this fast time scale, solutions behave practically like in the hyperbolic case, and a smoothened version of a rarefaction wave can be observed.

After the two peaks have merged, it is again diffusion that dominates the behavior. The whole process repeats itself, until only one single plateau is left, which will typically move to one of the domain boundaries. Thus, the only stable stationary state seems to be one plateau at the left or right boundary of the domain.

By construction of approximate solutions, it is shown numerically and analytically in [13] that the eigenvalues describing the slow movement of the peaks exponentially approach zero as the length of the domain increases. These exponentially small eigenvalues are typical features of metastable systems. The authors also derive an ODE describing the dynamics of a structure with two plateaus at the domain boundaries. Here, we will use exponential asymptotics to formally derive a system of ODEs describing the slow movement of the plateaus. This method has been successfully used in various applications; see, for instance, Ward [18] and references therein.

Metastable dynamics of the parabolic system. In the following analysis of the long-time behavior of system (1.5), (1.6), now rewritten as

$$\partial_t \varrho + \partial_x J = 0, \qquad J = \varrho(1-\varrho)\partial_x \mathcal{S}[\varrho] - \varepsilon \partial_x \varrho, \qquad J = 0 \quad \text{at } x = 0, L, \tag{4.3}$$

we will assume that the formation of patterns from the initial data has already taken place, and that a quasi-stationary pattern with plateaus (close to (3.7)) has been formed. Our aim is to derive a system of equations describing the evolution of the positions of the plateau boundaries $x_1(t), \dots, x_M(t)$.

A first approximation to a solution of the parabolic problem with plateaus is a stationary entropy solution of the hyperbolic problem $(\bar\varrho_\infty, \bar S_\infty)$ solving (3.7) and (3.8). However, we need a much better approximation $(\tilde\varrho, \tilde S)$ with boundary layer corrections close to the jumps of $\bar\varrho_\infty$. Actually we shall try to solve the parabolic steady state problem

$$\varepsilon \partial_x \varrho - \varrho(1-\varrho)\partial_x \mathcal{S}[\varrho] = 0 \tag{4.4}$$

as precisely as possible. The approximation $\tilde S$ for the chemoattractant density will be constructed such that it is close to $\bar S_\infty$ with the same qualitative behavior. In particular, it has the same monotonicity behavior at the plateau boundaries and extrema $\tilde x_{i+1/2}$ with the ordering $0 = \tilde x_{1/2} < x_1 < \tilde x_{3/2} < \cdots < \tilde x_{M-1/2} < x_M < \tilde x_{M+1/2} = L$ as the extrema $\bar x_{i+1/2}$ of $\bar S_\infty$ introduced above. For the construction of the approximating cell density, we consider the boundary layer problem

$$\varepsilon \partial_x \hat\varrho_i = \hat\varrho_i(1-\hat\varrho_i)\partial_x \tilde S, \qquad \hat\varrho_i(x_i) = \frac{1}{2}, \tag{4.5}$$

where the auxiliary condition fixes the position of the boundary layer. The solution

$$\hat\varrho_i[\tilde S](x) = \left[1 + \exp\left(\frac{\tilde S(x_i) - \tilde S(x)}{\varepsilon}\right)\right]^{-1} \tag{4.6}$$

will be used for $\tilde x_{i-1/2} < x < \tilde x_{i+1/2}$. The shape of $\hat\varrho_i$ depends on the monotonicity of $\tilde S$ in this interval: for increasing $\tilde S$, $\hat\varrho_i(\tilde x_{i-1/2}) \approx 0$ and $\hat\varrho_i(\tilde x_{i+1/2}) \approx 1$, and vice versa for decreasing $\tilde S$. Thus, the boundary layer solution has the appropriate behavior for jumps satisfying the entropy condition (3.9).

The construction of the boundary layer solution is nonstandard from the point of view of singular perturbation theory, where the standard procedure would lead to evaluation of $\partial_x \tilde S$ at $x = x_i$ in (4.5) and, consequently, $\hat\varrho_i[\tilde S](x) = [1 + \exp(\partial_x \tilde S(x_i)(x_i - x)/\varepsilon)]^{-1}$. This is, of course, an approximation of (4.6), obtained by Taylor expansion of $\tilde S$. The better accuracy of (4.6) is needed in the exponential asymptotics here.

By patching together the boundary layer solutions at the points $\tilde x_{i+1/2}$, exponentially small jump discontinuities would be created. By shifting the contributions appropriately, a continuous (actually continuously differentiable) approximate cell density is constructed:

$$\tilde\varrho[\tilde S](x) = \hat\varrho_i[\tilde S](x) - \Delta\varrho_i[\tilde S] \qquad \text{for } \tilde x_{i-1/2} < x < \tilde x_{i+1/2} \tag{4.7}$$

with

$$\Delta\varrho_i[\tilde S] = \sum_{j=1}^{i-1} \Big[\hat\varrho_{j+1}[\tilde S](\tilde x_{j+1/2}) - \hat\varrho_j[\tilde S](\tilde x_{j+1/2})\Big].$$

Note again that these corrections are exponentially small as $\varepsilon \to 0$. There is a certain arbitrariness in their choice. They would also serve their purpose if they would all be shifted by the same constant, which could be fixed by prescribing the total mass, i.e., the total number of cells. However, since our final result will only contain differences of the $\Delta\varrho_i[\tilde{S}]$, this issue is not important for us.

Finally, we require the chemoattractant density to satisfy

$$\tilde{S} = \mathcal{S}[\tilde{\varrho}[\tilde{S}]]. \tag{4.8}$$

This is a highly nonlinear problem whose solvability is not trivial at all. In the appendix we prove, for small ε, existence of a unique solution close to $\bar{S}_\infty$ and satisfying the qualitative assumptions mentioned above. The fact that $(\tilde{\varrho}, \tilde{S})$ is completely determined by the positions $x_1, \dots, x_M$ of the plateau boundaries motivates the notation $\tilde{\varrho} = \tilde{\varrho}(x; x_1, \dots, x_M)$, $\tilde{S} = \tilde{S}(x; x_1, \dots, x_M)$. If these coincide with the points where a stationary cell density takes the value $1/2$, then all the $\Delta\varrho_i$ vanish and $(\tilde{\varrho}, \tilde{S})$ is an exact steady state, since it also satisfies the Neumann boundary conditions $\partial_x\tilde{\varrho} = 0$, $x = 0, L$, by its construction and by the boundary conditions for $\tilde{S}$. In general, however, we obtain the residual

$$R := \varepsilon\partial_x\tilde{\varrho} - \tilde{\varrho}(1-\tilde{\varrho})\partial_x\tilde{S} = \Delta\varrho_i(1 - 2\hat{\varrho}_i + \Delta\varrho_i)\partial_x\tilde{S} \quad \text{in } (\tilde{x}_{i-1/2}, \tilde{x}_{i+1/2}). \tag{4.9}$$

The following procedure is an adaption of the methodology developed by Ward and coworkers for an asymptotic approximation of the metastable dynamics of the Ginzburg–Landau equation, viscous shocks, and the viscous Cahn–Hilliard equation (for an overview see, for instance, [18], [19] and references therein). The term "metastable" can be made more precise by considering the linearization of the problem around the M-parameter family of approximate steady states. The exponential smallness of the residuals leads to expecting M exponentially small eigenvalues. Metastability means that all the other eigenvalues are nonnegative. We do not have any results on the spectral problem, but the assumption of metastability is strongly supported by our numerical experiments and, even more strongly, by the numerical studies of the eigenvalue problem in [13].

We start by introducing a correction term for our approximate solution:

$$\varrho(x,t) = \tilde{\varrho}(x; x_1(t), \dots, x_M(t)) + r(x,t). \tag{4.10}$$

Since the approximate solution satisfies the boundary conditions, we have $\partial_x r = 0$ at $x = 0, L$. Just as in [19], we now consider an approximate version of (4.3), dropping nonlinear terms in r and assuming $|\partial_t r| \ll |\partial_t\tilde{\varrho}|$:

$$\partial_t\tilde{\varrho} + \partial_x J = 0, \qquad \mathcal{L}r + J + R = 0, \qquad J = 0 \quad \text{at } x = 0, L, \tag{4.11}$$

with the linearization of (4.4),

$$\mathcal{L}r = \varepsilon\partial_x r - (1 - 2\tilde{\varrho})\partial_x\tilde{S}r - \tilde{\varrho}(1-\tilde{\varrho})\partial_x\mathcal{S}[r]. \tag{4.12}$$

The nonlocal term $\mathcal{S}[r]$ is one of the major differences between our approach and the work in the above-mentioned references. As a first step, J will be computed by integrating the first equation in (4.11). From the definition of $\tilde{\varrho}$ in $(\tilde{x}_{i-1/2}, \tilde{x}_{i+1/2})$ we have

$$\partial_t\tilde{\varrho} = \frac{\hat{\varrho}_i(1-\hat{\varrho}_i)}{\varepsilon}\left[-\partial_x\tilde{S}(x_i)\dot{x}_i + \sum_{j=1}^{M}\left(\partial_{x_j}\tilde{S}(x) - \partial_{x_j}\tilde{S}(x_i)\right)\dot{x}_j\right] - \sum_{j=1}^{M}\partial_{x_j}\Delta\varrho_i\dot{x}_j.$$

From the differential equation (4.5) for the boundary layer term $\hat{\varrho}_i$ we see that $\hat{\varrho}_i(1-\hat{\varrho}_i)/\varepsilon$ is an approximate Delta-distribution concentrated at $x = x_i$ and with weight $|\partial_x \tilde{S}(x_i)|^{-1}$. The derivatives of the corrections $\Delta\varrho_i$ with respect to the x_j are expected to be exponentially small just as the $\Delta\varrho_i$ themselves. With these observations and (3.9), integration of the first equation in (4.11) gives

$$J(x) \approx \sum_{j=1}^{i} (-1)^{k_j} \dot{x}_j \qquad \text{for } x_i < x < x_{i+1}.$$

Here the boundary condition $J(0) = 0$ has been used. The other boundary condition $J(L) = 0$ leads to the equation

$$\sum_{j=1}^{M} (-1)^{k_j} \dot{x}_j = 0, \tag{4.13}$$

representing conservation of mass.

Now the second equation in (4.11) is multiplied by a test function $\psi(x,t)$ and integrated,

$$\varepsilon \psi r|_{x=0}^{L} + \int_0^L (r\mathcal{L}^*\psi + \psi(J+R))dx = 0, \tag{4.14}$$

with the formally adjoint operator

$$\mathcal{L}^*\psi = -\varepsilon\partial_x\psi - (1-2\tilde{\varrho})\partial_x\tilde{S}\psi + \partial_x\mathcal{S}[\tilde{\varrho}(1-\tilde{\varrho})\psi]. \tag{4.15}$$

In the computation of $\mathcal{L}^*$, the symmetry of Green's function G (see (2.3)) has been used.

The further procedure is motivated by the following observations: With $(\tilde{\varrho}, \tilde{S})$ we have constructed an M-parameter family of approximate solutions of the steady state problem (4.4) producing exponentially small residuals R (see (4.9)). Therefore we expect the linearized operator $\mathcal{L}$ to possess M exponentially small eigenvalues with eigenfunctions approximately given by the derivatives of $(\tilde{\varrho}, \tilde{S})$ with respect to the parameters $x_1, \ldots, x_M$. As a consequence, the inverse of $\mathcal{L}$ will act as a bounded operator on the inhomogeneity $J+R$ in the equation for r only if this inhomogeneity satisfies M solvability conditions characterized by the eigenfunctions of $\mathcal{L}^*$ corresponding to the exponentially small eigenvalues. Since $\mathcal{L}$ is not self-adjoint (the other major difference compared to earlier work for, e.g., the Cahn–Hilliard equation), the computation of approximations for these eigenfunctions is not immediate. We proceed pragmatically by trying to determine candidates for ψ such that the terms involving the unknown r in (4.14) become negligibly small.

The first two terms in (4.15) constitute a singularly perturbed differential operator with turning points (see [9]) close to $x = x_i$ (where $\tilde{\varrho} = 1/2$) and at $x = \tilde{x}_{i+1/2}$ (the extrema of $\tilde{S}$). These turning points are of different character since the sign of the coefficient $(1-2\tilde{\varrho})\partial_x\tilde{S}$ changes from positive to negative close to the x_i and vice versa at the $\tilde{x}_{i+1/2}$. The second group of turning points is interesting for us, since their character allows for spike layer solutions of the differential equation

$$\varepsilon\partial_x\psi + (1-2\tilde{\varrho})\partial_x\tilde{S}\psi = 0.$$

Such spike layer solutions will be the basis for our construction of appropriate ψ's. More precisely, we choose (for $i = 1, \dots, M-1$)

$$\psi_{i+1/2}(x) = c_{i+1/2} \exp\left(-\frac{1}{\varepsilon} \int_{\tilde{x}_{i+1/2}}^{x} (1 - 2\tilde{\varrho}(z)) \partial_x \tilde{S}(z) dz \right)$$

for $x_i \le x \le x_{i+1}$. Outside of this interval we extend $\psi_{i+1/2}$ as a smooth function satisfying

$$\begin{aligned}
&\psi_{i+1/2}(x) = 0 \quad \text{for } 0 \le x \le \tilde{x}_{i-1/2},\\
&|\psi_{i+1/2}(x)| \le \psi_{i+1/2}(x_i) \quad \text{for } \tilde{x}_{i-1/2} \le x \le x_i,\\
&|\psi_{i+1/2}(x)| \le \psi_{i+1/2}(x_{i+1}) \quad \text{for } x_{i+1} \le x \le \tilde{x}_{i+3/2},\\
&\psi_{i+1/2}(x) = 0 \quad \text{for } \tilde{x}_{i+3/2} \le x \le L,
\end{aligned}$$

which is possible under the basic assumption of the whole asymptotic procedure that all the points x_j and $\tilde{x}_{j+1/2}$ are well separated from each other. The constant $c_{i+1/2}$ is chosen such that $\psi_{i+1/2}$ approximately becomes a Delta-family for $\varepsilon \to 0$. This involves the computation of the integral

$$\begin{aligned}
&\int_0^L \exp\left(-\frac{1}{\varepsilon} \int_{\tilde{x}_{i+1/2}}^{x} (1 - 2\tilde{\varrho}(z)) \partial_x \tilde{S}(z) dz \right) dx\\
&\approx \int_{x_i}^{x_{i+1}} \exp\left(-\frac{|\partial_x^2 \tilde{S}(\tilde{x}_{i+1/2})|(x - \tilde{x}_{i+1/2})^2}{2\varepsilon} \right) dx \approx \sqrt{\frac{2\pi\varepsilon}{|\partial_x^2 \tilde{S}(\tilde{x}_{i+1/2})|}},
\end{aligned}$$

leading to

$$c_{i+1/2} = \sqrt{\frac{|\partial_x^2 \tilde{S}(\tilde{x}_{i+1/2})|}{2\pi\varepsilon}}.$$

In this construction we have neglected the last term in the adjoint operator (4.15) so far. Since $\tilde{\varrho}(1 - \tilde{\varrho})\psi_{i+1/2}$ is uniformly exponentially small, the same holds for $\partial_x \mathcal{S}[\tilde{\varrho}(1 - \tilde{\varrho})\psi_{i+1/2}]$, and, thus, also for $\mathcal{L}^* \psi_{i+1/2}$.

Using the functions $\psi_{i+1/2}$ in (4.14), we have to compute

$$\int_0^L \psi_{i+1/2} J \, dx \approx J(\tilde{x}_{i+1/2}) \approx \sum_{j=1}^{i} (-1)^{k_j} \dot{x}_j$$

as well as

$$\int_0^L \psi_{i+1/2} R \, dx.$$

The approximation of the second integral is less straightforward since, by (4.9), the integrand vanishes at $\tilde{x}_{i+1/2}$, where the mass of $\psi_{i+1/2}$ concentrates. We therefore split the integral into four parts A, B, C, and D, corresponding to the subintervals $(0, x_i)$, $(x_i, \tilde{x}_{i+1/2})$, $(\tilde{x}_{i+1/2}, x_{i+1})$, and (x_{i+1}, L), respectively. Using (4.9) and the properties of $\psi_{i+1/2}$, we easily estimate the first and the last terms:

$$\begin{aligned}
&|A| \le c\psi_{i+1/2}(x_i)|\Delta\varrho_i|,\\
&|D| \le c\psi_{i+1/2}(x_{i+1})|\Delta\varrho_{i+1}|,
\end{aligned}$$

with an ε-independent constant c. In a neighborhood of $\tilde{x}_{i+1/2}$, the residual can be approximated up to an exponentially small relative error by $R \approx \Delta\varrho_i(1-2\tilde{\varrho})\partial_x\tilde{S}$ for $x < \tilde{x}_{i+1/2}$, and by $R \approx \Delta\varrho_{i+1}(1-2\tilde{\varrho})\partial_x\tilde{S}$ for $x > \tilde{x}_{i+1/2}$. For the other two subintegrals we therefore obtain

$$\begin{aligned} B &\approx c_{i+1/2}\Delta\varrho_i \int_{x_i}^{\tilde{x}_{i+1/2}} \exp\left(-\frac{1}{\varepsilon}\int_{\tilde{x}_{i+1/2}}^{x}(1-2\tilde{\varrho}(z))\partial_x\tilde{S}(z)dz\right)(1-2\tilde{\varrho}(x))\partial_x\tilde{S}(x)dx \\ &\approx -\varepsilon c_{i+1/2}\Delta\varrho_i \end{aligned}$$

and similarly

$$C \approx \varepsilon c_{i+1/2}\Delta\varrho_{i+1}.$$

Since $\Delta\varrho_i$ is multiplied by the $O(\sqrt{\varepsilon})$-constant $\varepsilon c_{i+1/2}$ in B and by the exponentially small $\psi_{i+1/2}(x_i)$ in A, A is negligible compared to B and, analogously, D is negligible compared to C.

Collecting our results and using that $\psi_{i+1/2}$ vanishes on the boundary, (4.14) with $\psi = \psi_{i+1/2}$ leads to

$$\sum_{j=1}^{i}(-1)^{k_j}\dot{x}_j = \varepsilon c_{i+1/2}(\Delta\varrho_{i+1} - \Delta\varrho_i) \tag{4.16}$$

for $1 \le i \le M-1$. As previously mentioned, a common additive constant in the $\Delta\varrho_i$ would not change this result. From (4.16) with $i = 1$ we obtain an ODE for x_1:

$$\dot{x}_1 = (-1)^{k_1}\varepsilon c_{3/2}(\Delta\varrho_2 - \Delta\varrho_1). \tag{4.17}$$

Equations for x_i, $2 \le i \le M-1$, are derived by taking differences of consecutive versions of (4.16):

$$\dot{x}_i = (-1)^{k_i}\varepsilon\left[c_{i+1/2}(\Delta\varrho_{i+1} - \Delta\varrho_i) - c_{i-1/2}(\Delta\varrho_i - \Delta\varrho_{i-1})\right]. \tag{4.18}$$

Finally, the difference between the mass conservation equation (4.13) and (4.16) with $i = M-1$ gives

$$\dot{x}_M = (-1)^{k_M}\varepsilon c_{M-1/2}(\Delta\varrho_{M-1} - \Delta\varrho_M). \tag{4.19}$$

In principal, this completes the asymptotic procedure, and the dynamics of $x_1(t), \dots, x_M(t)$ is completely determined by (4.17)–(4.19). However, the right-hand sides in (4.17)–(4.19) are given in terms of $\tilde{S}$, the solution of (4.8), which is not known explicitly. On the other hand, it will be shown in the appendix that the explicitly computable $\bar{S}_\infty$ is a good enough approximation (essentially up to $O(\varepsilon^2)$) for $\tilde{S}$ to maintain the accuracy of the leading terms in (4.17)–(4.19). Therefore, the final result of our asymptotics is the system (4.17)–(4.19) with

$$\begin{aligned} \Delta\varrho_{i+1} - \Delta\varrho_i &= \left[1+\exp\left(\frac{\bar{S}_\infty(x_{i+1}) - \bar{S}_\infty(\bar{x}_{i+1/2})}{\varepsilon}\right)\right]^{-1} \\ &\quad - \left[1+\exp\left(\frac{\bar{S}_\infty(x_i) - \bar{S}_\infty(\bar{x}_{i+1/2})}{\varepsilon}\right)\right]^{-1}, \\ c_{i+1/2} &= \sqrt{\frac{|\bar{S}_\infty(\bar{x}_{i+1/2}) - \bar{\varrho}_\infty(\bar{x}_{i+1/2})|}{2\pi\varepsilon}}, \\ \bar{S}_\infty(x) &= \sum_{k_i \text{ odd}} \int_{x_i}^{x_{i+1}} \left[\frac{1}{2}e^{-|x-y|} + \frac{e^{x+y} + e^{2L-x-y} + e^{x-y} + e^{y-x}}{2(e^{2L}-1)}\right] dy. \end{aligned}$$

The whole asymptotic approach is based on the fact that the movement of the boundary layers is exponentially slow. It is valid only as long as x_1 and x_M stay away from the boundaries and an extremal point $\bar{x}_{i+1/2} \in (x_i, x_{i+1})$ of $\bar{S}_\infty$ exists for every pair $x_i < x_{i+1}$. These conditions are equivalent to the requirement that all plateau boundaries satisfy the entropy conditions (3.9). As soon as they are violated, the hyperbolic dynamics starts to dominate.

With the above approximations, the right-hand sides of (4.17)–(4.19) can be evaluated explicitly in terms of $x_1, \dots, x_M$. However, in the general case the formulas are very long and not very instructive. They involve not only the evaluation of the integrals in the last equation above, but also the computation of all the extremal points $\bar{x}_{3/2}, \dots, \bar{x}_{M-1/2}$ of $\bar{S}_\infty$. As an example, we discuss the simplest situation $M = 2$ with $k_0 = 1$, i.e., two plateaus adjacent to the boundaries with one vacuum region in the middle. In this case, the system (4.17)–(4.19) reduces to

$$\dot{x}_1 = \dot{x}_2 = \varepsilon c_{3/2}(\Delta\varrho_2 - \Delta\varrho_1). \tag{4.20}$$

Conservation of the initial mass m implies $x_1 + (L - x_2) = m$, and the system can be reduced to a scalar equation for x_1, substituting $x_2 = L - m + x_1$. The chemoattractant density is given by

$$\bar{S}_\infty(x) = \frac{\sinh(x_1)\cosh(L - x) + \sinh(m - x_1)\cosh(x)}{\sinh(L)}$$

for $x_1 \le x \le x_2$ with one interior minimum at

$$\bar{x}_{3/2} = \frac{1}{2}\log\frac{e^L \sinh(x_1) + \sinh(m - x_1)}{e^{-L}\sinh(x_1) + \sinh(m - x_1)}.$$

Steady states of (4.20) have to satisfy $\bar{S}_\infty(x_1) = \bar{S}_\infty(x_2)$. It is easily seen that $x_1 = m/2$ is the only solution; i.e., the only steady state is the symmetric one, where the mass is distributed equally between the two plateaus. In general, $\operatorname{sign}(x_1 - m/2) = \operatorname{sign}(\bar{S}_\infty(x_1) - \bar{S}_\infty(x_2)) = \operatorname{sign}(\dot{x}_1)$, showing the instability of the steady state. If initially one of the plateaus is bigger, then it will grow at the expense of the smaller one until all the mass is concentrated adjacent to one of the boundaries. This is expected to be the stable stationary state. In the general case $M > 2$, our understanding of the qualitative behavior of (4.17)–(4.19) is less complete. It is easy to see that steady states are characterized by the requirement that $\bar{S}_\infty$ take the same value at the plateau edges $x_1, \dots, x_M$. This implies that all plateaus have the same length and that the same is true for the vacuum regions separating them, with plateaus and/or vacuum regions adjacent to the boundary having half the interior length. This shows that all stationary solutions of the full problem as characterized in [13] are represented. We conjecture that all these solutions are unstable, but a proof is lacking.

About the dynamics we observe that, generically, one of the exponentially small terms $\Delta\varrho_{i+1} - \Delta\varrho_i$ will dominate all the others. As a consequence, effectively only two neighboring plateau edges x_i and x_{i+1} will move (with the same velocity), while the others are approximately stationary.

In Figure 4.2, we compare the numerical solution of the full system (1.5), (1.6) with the solution of system (4.17)–(4.19) by plotting the position of the boundaries of a single plateau situated at $x_1 = 0.6$ and $x_2 = 0.8$ for different values of ε. Light lines represent the solution obtained by solving the full system with an upwind scheme with grid size $\Delta x = \Delta t = 3 \times 10^{-4}$; dark lines were obtained by solving system

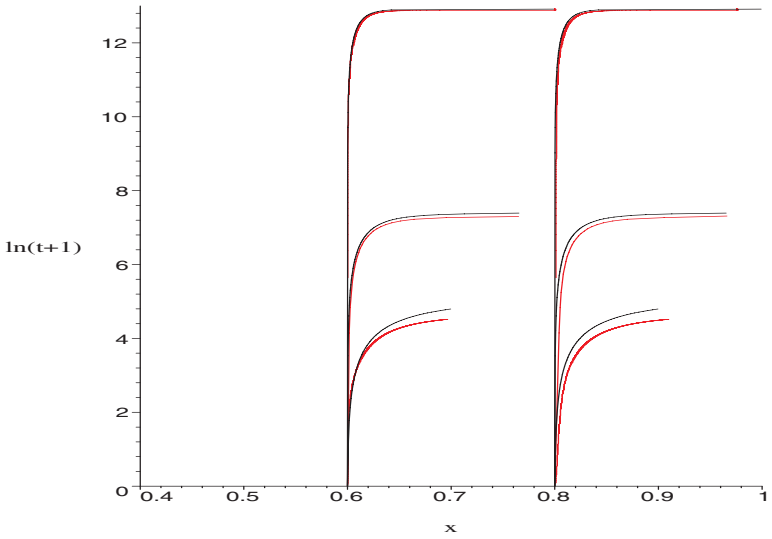
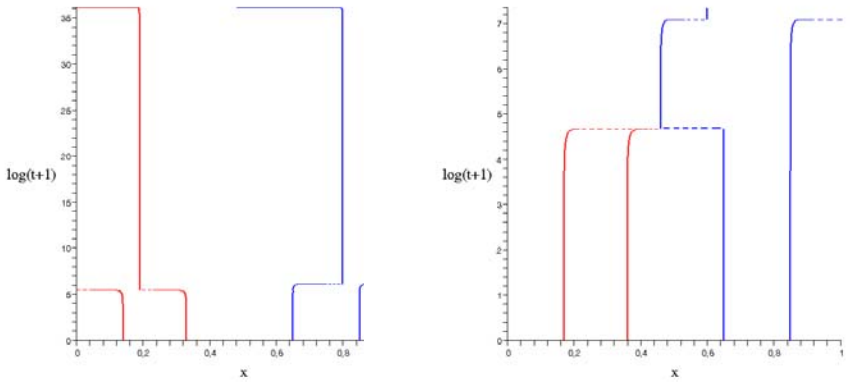


FIG. 4.2. *Comparison of the numerical solution of system* (1.5), (1.6) *with* $\varepsilon = 4 \times 10^{-4}$, 2×10^{-4}, *and* 1×10^{-4} *obtained with an upwind scheme (light) and the numerical solution of the corresponding ODE system (dark). The different lengths of the branches is due to the fact that as* ε *becomes smaller, solutions keep their plateau-like shape even when they get quite close to the boundary, whereas for larger* ε, *the hyperbolic dynamics start to take over much faster.*

(4.17)–(4.19) for $M = 2$ using the MAPLE routine `lsode` (a Livermore stiff ODE solver). The two lowest branches in the figure correspond to $\varepsilon = 4 \times 10^{-4}$, and it can be observed that the time it takes the plateau to advance towards the boundary calculated by the two different approaches differs slightly. However, as we decrease ε ($\varepsilon = 2 \times 10^{-4}$ for the middle branches, $\varepsilon = 1 \times 10^{-4}$ for the top branches) and thus the error introduced by the approximating assumptions we had to take in order to obtain the ODEs (4.17)–(4.19), the distance between the lines decreases until, for $\varepsilon = 1 \times 10^{-4}$, the trajectories of x_1 and x_2 obtained by the upwind scheme and the ODE system are practically identical.

5. A hybrid numerical-asymptotic approach. Developing a numerical scheme that correctly captures both the short- and the long-time behavior of the parabolic system is a nontrivial task. If we use a standard discretization of (1.5), (1.6) with a grid size that is too large, the long-time behavior of the system will be driven by numerical errors dominating the exponentially small terms responsible for the exact dynamics. Choosing a grid that is fine enough to reduce numerical errors in a sufficient way leads to very large computation times. Another approach is to solve the equations for the positions of the plateau edges (4.17)–(4.19), and then to specify an approximate solution $\tilde{\varrho}$ at each time step according to (4.7). However, this solution is valid only as long as the conditions (3.9) are satisfied.

These observations motivate a combined approach for the numerical solution of (1.5), (1.6): As long as (3.9) holds, we use the asymptotic approximations (4.17)–(4.19). We solve the equations with MAPLE and calculate the corresponding $\tilde{\varrho}$ at each time step. When the velocities of the plateau edges become $\mathcal{O}(\varepsilon)$, we switch to a full numerical solution with the finite difference scheme described above. A similar numerical-asymptotic approach was developed in [17] to solve the viscous Cahn–Hillard equation in one space dimension.



(a) $x_1 = 0.13$, $x_2 = 0.32$ (b) $x_1 = 0.15$, $x_2 = 0.36$

FIG. 5.1. *Solutions of the ODE system* (4.17)–(4.19) *with different initial conditions and* $\varepsilon = 2 \times 10^{-4}$. *The position of the right plateau is* $x_3 = 0.65$, $x_4 = 0.85$ *in all pictures. An animation (obtained with the combined numerical-asymptotic approach described in the text) corresponding to Figure* 5.1(b) *can be found at http://www.ricam.oeaw.ac.at/people/page/dolak/animation.html.*

Example: Behavior near a stationary state. As an example, we investigate the dynamics of the parabolic system when solutions are initially close to an unstable stationary state.

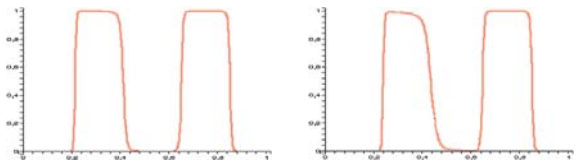
For a given initial mass m, this stationary solution consists of two plateaus of equal mass, with the outer edges being exactly half the distance between the plateaus away from the boundaries. The stationary solution is given by

$$x_1 = \frac{L-m}{4}, \quad x_2 = \frac{L+m}{4}, \quad x_3 = \frac{3L-m}{4}, \quad x_4 = \frac{3L+m}{4}. \tag{5.1}$$

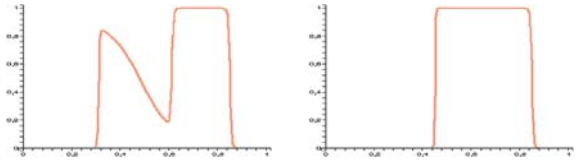
In our experiments, we set $m = 0.4$ and $L = 1$ to obtain the boundary layer positions $x_1 = 0.15$, $x_2 = 0.35$, $x_3 = 0.65$, and $x_4 = 0.85$ from (5.1). Then we choose two different sets of initial conditions close to this stationary point and calculate the temporal evolution of the boundary layers according to (4.17)–(4.19) with $M = 4$. After a plateau has moved to the domain boundary or merged with another plateau, we solve the system for $M = 3$ or $M = 2$, respectively.

In Figure 5.1(a), the left plateau has initially been made smaller and moved towards the left boundary. The evolution proceeds in three steps, as follows. 1. The left plateau moves to the left, until it reaches the boundary. 2. The right plateau moves to the right boundary. 3. The two remaining plateau edges move to the left, meaning that the bigger right plateau attracts cells from the left. The evolution stops after the left plateau has disappeared and one plateau adjacent to the right boundary is left as a stable steady state. As mentioned in the previous section, during each step only one pair of plateau edges moves in parallel.

Figure 5.1(b) features an initial condition, where the left plateau has again been made smaller but now moved towards the center compared to the unstable steady state. We observe a two-step evolution, as follows. 1. The left plateau moves to the right until it loses stability and is absorbed by the bigger right plateau. 2. The remaining plateau moves to the right until it reaches the boundary.



(a) $t = 48000$ (b) $t = 48130$



(c) $t = 48158$ (d) $t = 48165$

FIG. 5.2. *Fast dynamics of the parabolic system, corresponding to the dashed lines in Figure* 5.1(b). (a) *The cell density calculated according to the asymptotic approximations* (4.17)–(4.19) *and* (4.7) *shortly before the hyperbolic dynamics start to dominate.* (b), (c) *A rarefaction wave obtained by solving* (1.5), (1.6) *with an upwind scheme until, in* (d), *only one plateau is left.*

Details of the fast hyperbolic dynamics at the end of step 1 are shown in Figure 5.2. As the left plateau moves towards the right one, a rarefaction wave starts to form when the entropy condition for the corresponding hyperbolic system becomes violated. The outer plateau edge of the right plateau is not affected by this merging process and does not move, since locally the entropy condition is still satisfied. In general, however, it is an open problem to predict the outcome of the hyperbolic evolution because of the nonlocal coupling. After the loss of stability of a plateau edge, the hyperbolic evolution could induce stability losses of other plateau edges. Therefore it cannot be ruled out that several plateaus disappear within one "hyperbolic transition layer."

Appendix. We shall prove solvability of the approximate steady state problem (4.8) and an approximation result for its solution. Recalling the definition (3.7), (3.8) of the plateau state $(\bar\varrho_\infty, \bar S_\infty)$ and of the extremal points $\bar x_{i+1/2} \in (x_i, x_{i+1})$ of $\bar S_\infty$, we define

$$x_{i\pm1/4} := \frac{x_i + \bar x_{i\pm1/2}}{2} \quad \text{for } 1 \le i \le M, \qquad x_{1/4} := 0, \qquad x_{M+3/4} := L,$$

and

$$A_1 := \bigcup_{i=1}^{M} (x_{i-1/4}, x_{i+1/4}), \qquad A_2 := [0, L] \setminus A_1 = \bigcup_{i=0}^{M} [x_{i+1/4}, x_{i+3/4}].$$

Since A_1 stays away from the extremal points of $\bar S_\infty$, and A_2 stays away from its turning points x_i, there exists a $\delta > 0$ such that

$$|\partial_x \bar S_\infty| \ge 2\delta \quad \text{in } A_1, \qquad |\partial_x^2 \bar S_\infty| \ge 2\delta \quad \text{in } A_2.$$

For chemoattractant densities we shall use the Banach space $\mathcal{B}_1 := C^1([0,L]) \cap C^2(A_2)$ with its natural norm

$$\|S\|_1 := \|S\|_{L^\infty((0,L))} + \|\partial_x S\|_{L^\infty((0,L))} + \|\partial_x^2 S\|_{L^\infty(A_2)}$$

and the ball

$$B_\delta := \{S:\ \|S - \bar{S}_\infty\|_1 < \delta\}.$$

Then $S \in B_\delta$ implies

$$|\partial_x S| \geq \delta \quad \text{in } A_1, \qquad |\partial_x^2 S| \geq \delta \quad \text{in } A_2.$$

Therefore, in every subinterval $[x_{i+1/4}, x_{i+3/4}]$ of A_2, S has a unique local extremum $x_{i+1/2}[S]$ of the same character as $\bar{x}_{i+1/2} = x_{i+1/2}[\bar{S}_\infty]$. Consequently $\tilde{\varrho}[S]$ is well defined by (4.7). A fixed point operator for solving (4.8) is now defined on B_δ by

$$F[S](x) := \mathcal{S}[\tilde{\varrho}[S]]. \tag{A.1}$$

Our Banach space for cell densities will be $\mathcal{B}_2 := L^1((0,L)) \cap C(A_2)$ with the norm

$$\|\varrho\|_2 := \|\varrho\|_{L^1((0,L))} + \|\varrho\|_{L^\infty(A_2)}.$$

First we prove that the map $\mathcal{S}$ from ϱ to S given by solving the elliptic S-problem is continuous.

LEMMA A.1. *There exists a $c > 0$ such that, for every $\varrho \in \mathcal{B}_2$, $\mathcal{S}[\varrho] \in \mathcal{B}_1$ holds and $\|\mathcal{S}[\varrho]\|_1 \leq c\|\varrho\|_2$.*

Proof. The result is a consequence of the facts that G and $\partial_x G$ are uniformly bounded and that $S = \mathcal{S}[\varrho]$ solves the differential equation $\partial_x^2 S = S - \varrho$. ∎

Now we are ready to prove the main result of this section.

THEOREM A.2. *For ε small enough, problem (4.8) has a unique solution in B_δ.*

Proof. For every $S \in B_\delta$, $\tilde{\varrho}[S]$ deviates from $\bar{\varrho}_\infty$ only by boundary layer corrections close to the discontinuities $x_1, \ldots, x_M \in A_1$. The thickness of the boundary layers is $O(\varepsilon)$. In A_2, $\tilde{\varrho}[S]$ and $\bar{\varrho}_\infty$ are exponentially close as $\varepsilon \to 0$. This immediately implies $\|\tilde{\varrho}[S] - \bar{\varrho}_\infty\|_2 = O(\varepsilon)$ and, thus, from Lemma A.1, $\|F[S] - \bar{S}_\infty\|_1 = O(\varepsilon)$. This proves that for ε small enough, F maps B_δ into itself.

Now let $S_1, S_2 \in B_\delta$ and set $x^l_{i+1/2} := x_{i+1/2}[S_l]$, $\hat{\varrho}^l_i := \hat{\varrho}_i[S_l]$, $\Delta\varrho^l_i := \Delta\varrho_i[S_l]$, $l = 1, 2$. Assume $\partial_x S_1(x_i), \partial_x S_2(x_i) > 0$ and $x_i < x < \min\{x^1_{i+1/2}, x^2_{i+1/2}\}$. Then we have

$$\begin{aligned}
|\hat{\varrho}^1_i(x) - \hat{\varrho}^2_i(x)| &\leq \left|\exp\left(\frac{S_1(x_i) - S_1(x)}{\varepsilon}\right) - \exp\left(\frac{S_2(x_i) - S_2(x)}{\varepsilon}\right)\right| \\
&\leq \exp\left(\frac{\delta(x_i - x)}{\varepsilon}\right) \frac{1}{\varepsilon} \left|S_1(x_i) - S_1(x) - S_2(x_i) + S_2(x)\right| \\
&\leq \exp\left(\frac{\delta(x_i - x)}{\varepsilon}\right) \frac{x - x_i}{\varepsilon} \|S_1 - S_2\|_1.
\end{aligned}$$

Analogous estimates for $\max\{x^1_{i-1/2}, x^2_{i-1/2}\} < x < x_i$ and for $\partial_x S_l(x_i) < 0$ lead to

$$|\hat{\varrho}^1_i(x) - \hat{\varrho}^2_i(x)| \leq \exp\left(-\frac{\delta|x_i - x|}{\varepsilon}\right) \frac{|x - x_i|}{\varepsilon} \|S_1 - S_2\|_1 \tag{A.2}$$

for $1 \leq i \leq M$ and $\max\{x^1_{i-1/2}, x^2_{i-1/2}\} < x < \min\{x^1_{i+1/2}, x^2_{i+1/2}\}$.

The mean value theorem gives

$$\partial_x S_1(x^2_{i+1/2}) - \partial_x S_2(x^2_{i+1/2}) = \partial_x S_1(x^2_{i+1/2}) = \partial_x^2 S_1(\xi_{i+1/2})(x^2_{i+1/2} - x^1_{i+1/2}),$$

with $\xi_{i+1/2} \in A_2$. Since $S_1 \in B_\delta$,

$$|x^1_{i+1/2} - x^2_{i+1/2}| \leq \frac{1}{\delta} \|S_1 - S_2\|_1 \tag{A.3}$$

follows. Now assume $x^1_{i+1/2} < x^2_{i+1/2}$. Then we can estimate

$$|\hat{\varrho}_i^1(x_{i+1/2}^1) - \hat{\varrho}_i^2(x_{i+1/2}^2)| \le |\hat{\varrho}_i^1(x_{i+1/2}^1) - \hat{\varrho}_i^2(x_{i+1/2}^1)| + |\hat{\varrho}_i^2(x_{i+1/2}^1) - \hat{\varrho}_i^2(x_{i+1/2}^2)|.$$

For the first term on the right-hand side, (A.2) can be used to give a bound of the form $\mathcal{ESST}\|S_1 - S_2\|_1$, where the abbreviation $\mathcal{ESST}$ means "exponentially small term," i.e., a term of the form $\exp(-\kappa/\varepsilon)$ with $\kappa > 0$. For the second term we use (A.3) and the fact that $\partial_x \hat{\varrho}_i^2$ is exponentially small in A_2 to obtain an estimate of the same type. Interchanging the roles of S_1 and S_2, the same can be done for $x_{i+1/2}^2 < x_{i+1/2}^1$. A consequence of these results is

$$|\Delta\varrho_i^1 - \Delta\varrho_i^2| \le \mathcal{ESST}\|S_1 - S_2\|_1 \tag{A.4}$$

for $1 \le i \le M$. Combining (A.2) and (A.4), we have

$$|\tilde{\varrho}[S_1](x) - \tilde{\varrho}[S_2](x)| \le \left[\mathcal{EST} + \exp\left(-\frac{\delta|x_i - x|}{\varepsilon}\right)\frac{|x - x_i|}{\varepsilon}\right]\|S_1 - S_2\|_1, \tag{A.5}$$

for $1 \le i \le M$ and $\max\{x_{i-1/2}^1, x_{i-1/2}^2\} < x < \min\{x_{i+1/2}^1, x_{i+1/2}^2\}$. It remains to consider $\eta_{i+1/2} := \min\{x_{i+1/2}^1, x_{i+1/2}^2\} < x < \max\{x_{i+1/2}^1, x_{i+1/2}^2\}$:

$$|\tilde{\varrho}[S_1](x) - \tilde{\varrho}[S_2](x)| \le |\tilde{\varrho}[S_1](\eta_{i+1/2}) - \tilde{\varrho}[S_2](\eta_{i+1/2})| + \mathcal{ESST}|x - \eta_{i+1/2}|,$$

since $\partial_x \tilde{\varrho}[S_l]$ is exponentially small in A_2. For the first term on the right-hand side we use (A.5) and for the second (A.3), to obtain

$$|\tilde{\varrho}[S_1](x) - \tilde{\varrho}[S_2](x)| \le \mathcal{ESST}\|S_1 - S_2\|_1, \tag{A.6}$$

for $0 \le i \le M$ and $\min\{x_{i+1/2}^1, x_{i+1/2}^2\} < x < \max\{x_{i+1/2}^1, x_{i+1/2}^2\}$. Since the integral of the second term in the bracket in (A.5) is $O(\varepsilon)$, a combination of (A.5) and (A.6) leads to

$$\|\tilde{\varrho}[S_1] - \tilde{\varrho}[S_2]\|_2 \le c\,\varepsilon\|S_1 - S_2\|_1,$$

and, with Lemma A.1,

$$\|F[S_1] - F[S_2]\|_1 \le c\,\varepsilon\|S_1 - S_2\|_1,$$

showing that F is a contraction for ε small enough, and thus completing the proof of the theorem. □

Finally, it will be shown by formal asymptotic arguments that it is asymptotically correct to approximate $\tilde{S}$ by $\bar{S}_\infty$ in the right-hand sides of the ODEs (4.17)–(4.19). It is easily seen that the exponentially small terms in the $\Delta\varrho_i$ are approximated with a $O(\varepsilon)$ relative error if the chemoattractant density $\tilde{S}$ is approximated up to $O(\varepsilon^2)$. This holds for $\bar{S}_\infty$ since

$$\begin{aligned}
\tilde{S}(x) - \bar{S}_\infty(x) &= \int_0^L G(x,y)(\tilde{\varrho}[\tilde{S}](y) - \bar{\varrho}_\infty(y))dy \\
&= \sum_{i=1}^M \int_{\tilde{x}_{i-1/2}}^{\tilde{x}_{i+1/2}} G(x,y)(\hat{\varrho}_i[\tilde{S}](y) - \bar{\varrho}_\infty(y) - \Delta\varrho_i)dy \\
&= \varepsilon\sum_{i=1}^M (-1)^{k_i} G(x,x_i)\left(\int_0^\infty \frac{d\xi}{1+\exp(|\partial_x\tilde{S}(x_i)|\xi)}\right. \\
&\qquad\qquad \left. - \int_{-\infty}^0 \frac{d\xi}{1+\exp(-|\partial_x\tilde{S}(x_i)|\xi)}\right) + O(\varepsilon^2) \\
&= O(\varepsilon^2).
\end{aligned}$$

The third equality is due to the substitution $y = x_i + \varepsilon\xi$ and straightforward Taylor expansion.

Acknowledgments. Y. D. wants to thank T. Hillen for valuable discussions. The authors also owe thanks to two anonymous referees, whose comments motivated major improvements in this work.

REFERENCES

[1] P. Bates and J. Xun, *Metastable patterns for the Cahn-Hilliard equation* I, J. Differential Equations, 111 (1994), pp. 421–457.
[2] P. Bates and J. Xun, *Metastable patterns for the Cahn-Hilliard equation* II. *Layer dynamics and slow invariant manifold*, J. Differential Equations, 117 (1995), pp. 165–216.
[3] Y. Dolak and C. Schmeiser, *Kinetic models for chemotaxis: Hydrodynamic limits and the back-of-the-wave problem*, J. Math. Biol., to appear.
[4] J. Haskovec and C. Schmeiser, *Transport in semiconductors at saturated velocities*, Comm. Math. Sci., 3 (2005), pp. 219–233.
[5] T. Hillen and K. Painter, *Global existence for a parabolic chemotaxis model with prevention of overcrowding*, Adv. Appl. Math., 26 (2001), pp. 280–301.
[6] E. F. Keller and L. A. Segel, *Initiation of slime mold aggregation viewed as an instability*, J. Theoret. Biol., 26 (1970), pp. 399–415.
[7] J. G. L. Laforgue and R. E. O'Malley, Jr., *Shock layer movement for Burgers' equation*, SIAM J. Appl. Math., 55 (1995), pp. 332–347.
[8] P. A. Markowich and P. Szmolyan, *A system of convection-diffusion equations with small diffusion coefficient arising in semiconductor physics*, J. Differential Equations, 81 (1989), pp. 234–254.
[9] R. E. O'Malley, *Introduction to Singular Perturbations*, Academic Press, New York, 1974.
[10] K. Painter and T. Hillen, *Volume-filling and quorum sensing in models for chemosensitive movement*, Canad. Appl. Math. Quart., 10 (2003), pp. 280–301.
[11] C. S. Patlak, *Random walk with persistence and external bias*, Bull. Math. Biophys., 15 (1953), pp. 311–338.
[12] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1983.
[13] A. B. Potapov and T. Hillen, *Metastability in chemotaxis models*, J. Dynam. Differential Equations, 17 (2005), pp. 293–330.
[14] L. Reyna and M. Ward, *On the exponentially slow motion of a viscous shock*, Comm. Pure Appl. Math., 48 (1995), pp. 79–120.
[15] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Anal. Math. Pura Appl., 146 (1987), pp. 65–96.
[16] B. D. Sleeman, M. J. Ward, and J. C. Wei, *The existence and stability of spike patterns in a chemotaxis model*, SIAM J. Appl. Math., 65 (2005), pp. 790–817.
[17] X. Sun and M. Ward, *The dynamics and coarsening of interfaces for the viscous Cahn-Hilliard equation in one spatial dimension*, Stud. Appl. Math., 105 (2000), pp. 203–234.
[18] M. Ward, *Exponential asymptotics and convection-diffusion-reaction models*, in Analyzing Multiscale Phenomena Using Singular Perturbation Methods, Proc. Sympos. Appl. Math. 56, AMS, Providence, RI, 1998, pp. 151–184.
[19] M. Ward, *Dynamic metastability and singular perturbations*, in Boundaries, Interfaces, and Transitions, CRM Proc. Lecture Notes 13, M. C. Delfour, ed., AMS, Providence, RI, 1998, pp. 237–263.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 309–338

TOUCHDOWN AND PULL-IN VOLTAGE BEHAVIOR OF A MEMS DEVICE WITH VARYING DIELECTRIC PROPERTIES*

YUJIN GUO†, ZHENGUO PAN†, AND M. J. WARD†

Abstract. The pull-in voltage instability associated with a simple MEMS device, consisting of a thin dielectric elastic membrane supported above a rigid conducting ground plate, is analyzed. The upper surface of the membrane is coated with a thin conducting film. In a certain asymptotic limit representing a thin device, the mathematical model consists of a nonlinear partial differential equation for the deflection of the thin dielectric membrane. When a voltage V is applied to the conducting film, the dielectric membrane deflects towards the bottom plate. For a slab, a circular cylindrical, and a square domain, numerical results are given for the saddle-node bifurcation value V_*, also referred to as the pull-in voltage, for which there is no steady-state membrane deflection for $V > V_*$. For $V > V_*$ it is shown numerically that the membrane dynamics are such that the thin dielectric membrane touches the lower plate in finite time. Results are given for both spatially uniform and nonuniform dielectric permittivity profiles in the thin dielectric membrane. By allowing for a spatially nonuniform permittivity profile, it is shown that the pull-in voltage instability can be delayed until larger values of V and that greater pull-in distances can be achieved. Analytical bounds are given for the pull-in voltage V_* for two classes of spatially variable permittivity profiles. In particular, a rigorous analytical bound V_1, which depends on the class of permittivity profile, is derived that guarantees for the range $V > V_1 > V_*$ that there is no steady-state solution for the membrane deflection and that finite-time touchdown occurs. Numerical results for touchdown behavior, both for $V > V_1$ and for $V_* < V < V_1$, together with an asymptotic construction of the touchdown profile, are given for both a spatially uniform and a spatially nonuniform permittivity profile.

Key words. quenching, pull-in voltage, saddle-node, MEMS, dielectric permittivity

AMS subject classifications. 34K55, 74H10, 74K15

DOI. 10.1137/040613391

1. Introduction.

Microelectromechanical systems (MEMS) combine electronics with micro-size mechanical devices to design various types of microscopic machinery. MEMS devices are key components of many commercial systems, including accelerometers for airbag deployment in automobiles, ink jet printer heads, and chemical sensors. Mathematical models of physical phenomena associated with the rapidly developing field of MEMS technology are discussed in [13].

A key component of many MEMS systems is the simple device shown in Figure 1. The upper part of this device consists of a thin deformable elastic membrane that is held fixed along its boundary. This membrane is modeled as a dielectric of a thin, but finite, thickness. The upper surface of this membrane is coated with a negligibly thin metallic conducting film. The thin dielectric membrane lies above a rigid inelastic conducting ground plate. When a voltage V is applied to the conducting film, the thin dielectric membrane deflects towards the ground plate. A similar deflection phenomenon, but on a macroscopic length scale, occurs in the field of electrohydrodynamics. In this context, Taylor [17] studied the electrostatic deflection of two oppositely charged soap films, and he predicted a critical voltage for which the

*Received by the editors August 13, 2004; accepted for publication (in revised form) May 10, 2005; published electronically November 15, 2005.
http://www.siam.org/journals/siap/66-1/61339.html
†Department of Mathematics, University of British Columbia, Vancouver, V6T 1Z2 Canada (yjgguo@math.ubc.ca, panzg@math.ubc.ca, ward@math.ubc.ca). The third author gratefully acknowledges support from NSERC under grant 81541.

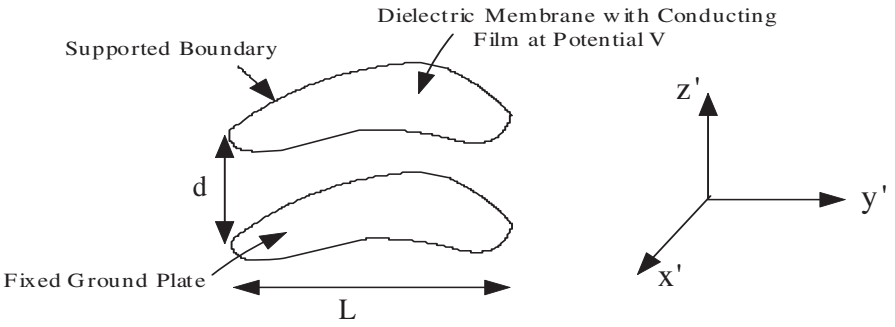


FIG. 1. *The MEMS capacitor. The upper surface of the elastic membrane is coated with an ultrathin conducting film.*

two soap films would touch together.

A similar physical limitation on the applied voltage occurs for the MEMS device of Figure 1 in that there is a maximum voltage, called the pull-in voltage V_*, that can be safely applied to the system. More specifically, if the applied voltage V is increased past V_*, there is no longer a steady-state solution for the membrane deflection (cf. [11], [14]). The existence of such a pull-in voltage was first demonstrated for a lumped mass-spring model of electrostatic actuation in the pioneering study of [10], where the restoring force of the deflected membrane is modeled by a Hookean spring. In this lumped model the attractive inverse square law electrostatic force between the membrane and the ground plate dominates the restoring force of the spring for small gap sizes and large applied voltages. This leads to a touchdown or snap-through behavior whereby the membrane hits the ground plate when the applied voltage is sufficiently large. Although the lumped model qualitatively predicts the existence of a pull-in voltage and the snap-through phenomenon, it cannot quantitatively account for details such as membrane geometry.

A more detailed mathematical model of this phenomena, leading to a partial differential equation (PDE) for the dimensionless deflection u of the membrane, was derived and analyzed in [6], [11], [12], [14], and [15] (see also the references therein). In the damping-dominated limit, and by modeling the thin dielectric as a membrane with zero rigidity, a narrow-gap asymptotic analysis was used in [6] and [14] to derive that u satisfies

$$\frac{\partial u}{\partial t} = \Delta u - \frac{\lambda f(x,y)}{(1+u)^2}, \quad x \in \Omega; \qquad u = 0, \quad (x,y) \in \partial\Omega; \qquad u(x,y,0) = 0. \tag{1.1}$$

An outline of the derivation of (1.1) following that detailed in [14] and [6] is given in the appendix. In (1.1), λ characterizes the relative strength of electrostatic and mechanical forces in the system, and is given by

$$\lambda = \frac{\varepsilon_0 V^2 L^2}{2Td^3}. \tag{1.2}$$

Here V is the applied voltage, d is the undeflected gap size (see Figure 1), L and T are the length scale and tension of the membrane, respectively, and ε_0 is the permittivity of free space. In (1.1), Ω is a bounded domain in $\mathbb{R}^2$, and $f(x,y)$ is the *permittivity*

profile, defined in terms of the dielectric permittivity $\varepsilon_2(x,y)$ of the membrane, by

$$f(x,y) = \frac{\varepsilon_0}{\varepsilon_2(x,y)}. \tag{1.3}$$

The initial condition in (1.1) assumes that the membrane is initially undeflected and that the voltage is applied at time $t = 0$. Mathematically, the pull-in voltage is obtained from (1.2) in terms of the largest possible saddle-node bifurcation value λ_* of λ for which (1.1) has a steady-state solution.

In the actual design of a MEMS device there are several issues that must be considered. Typically, one of the primary device design goals is to achieve the maximum possible stable steady-state deflection, referred to as the *pull-in distance*, with a relatively small applied voltage V. Another consideration may be to increase the stable operating range of the device by increasing the pull-in voltage V_* subject to the constraint that the range of the applied voltage is limited by the available power supply. This increase in the stable operating range may be important for the design of microresonators. For other devices such as microvalves, where touchdown behavior is explicitly exploited, it is of interest to decrease the time for touchdown, thereby increasing the switching speed. One way of achieving larger values of λ_*, and hence larger values of V_*, while simultaneously increasing the pull-in distance, is to use a voltage control scheme imposed by an external circuit in which the device is placed (cf. [12]). This approach leads to a nonlocal problem for the deflection of the membrane (cf. [12]). A different approach, studied theoretically in [14], is to introduce a spatial variation in the dielectric permittivity $\varepsilon_2(x,y)$ of the membrane so that $\varepsilon_2(x,y)$ is largest, and consequently $f(x,y)$ smallest, in the region where the membrane deflection would normally be largest under a spatially uniform permittivity. For a power-law permittivity profile in a slab domain, this approach was shown in [14] to allow for an increase in both the pull-in voltage and the pull-in distance.

The first main goal of this paper is to extend the steady-state analysis in [14] by giving analytical and numerical results for the saddle-node value λ_* and the pull-in distance for (1.1) for some general classes of permittivity profiles $f(x,y)$. For the first class, we assume that $f(x,y)$ is bounded away from zero, so that

$$0 < C_0 \le f(x,y) \le 1, \qquad x \in \Omega. \tag{1.4}$$

For the second class of profile, we allow for part of the membrane to be perfectly conducting, so that

$$0 \le f(x,y) \le 1, \qquad x \in \Omega. \tag{1.5}$$

In Theorem 3.1 of [14], restated below in Theorem 2.1 of section 2, an upper bound for λ_* is obtained for permittivity profiles satisfying (1.4). This bound, however, does not apply to profiles satisfying (1.5). In particular, it does not apply to the power-law permittivity profiles considered in section 4 of [14], which vanish at one point in Ω. To treat this case, in Theorem 2.2 we use a different approach to obtain an upper bound for λ_* for permittivity profiles satisfying (1.5). In section 2.1 we give numerical results for λ_* for a power-law permittivity profile and an exponential permittivity profile, which satisfy (1.5) and (1.4), respectively. The precise forms for these profiles, which each depend on a parameter α, are given below in (2.16). Numerical results for λ_* and the pull-in distance as a function of α are given for a slab domain, a unit disk, and a square domain. For large values of α, these profiles

are such that $f(x,y) \ll 1$ except in a boundary layer near $\partial\Omega$. In this limit, we derive a scaling law for $\lambda*$.

The second main goal of this paper is to analyze and compute time-dependent touchdown behavior for (1.1) for permittivity profiles satisfying either (1.4) or (1.5). The solution u of (1.1) is said to touchdown at finite time if the minimum value of u reaches -1 at some $t = T_* < \infty$. At such a time, the membrane touches the bottom fixed plate. In section 3.1 we determine bounds on λ for which touchdown occurs in finite time. This approach also yields bounds on the touchdown time T_*. The first bound, which is obtained from the method of [6] (see also [16]), applies to a permittivity profile satisfying (1.4). The second bound applies to a permittivity profile satisfying (1.5).

In section 3.2 we analytically construct the local touchdown profile for the constant permittivity profile $f(x,y) \equiv 1$. To do so, we introduce a nonlinear change of variables in a manner similar to that used in [8] to determine the local behavior of the solution to a semilinear heat equation near the blow-up time and blow-up location. This approach leads to a PDE that has smooth solutions near the touchdown point. By constructing a formal power series solution for this PDE, the local form of the touchdown profile is obtained. As discussed in [8], this transformed PDE is also readily amenable to numerical computations. In this way, touchdown behavior is computed numerically. In section 3.3, we briefly construct the local touchdown profile for the constant permittivity profile $f(x,y) \equiv 1$ by using a formal center manifold analysis of a PDE that results from a near-similarity group transformation of (1.1). Such a dynamical systems approach has been used previously in [5] to study quenching behavior in one space dimension and in [4] to study blow-up behavior for a semilinear heat equation in N space dimensions. Another approach for studying quenching behavior is given in [7].

In section 4 we give some asymptotic results for the touchdown profile for spatially variable permittivity profiles. Numerical results of touchdown behavior are also shown. Finally, in section 5, we list a few open mathematical problems.

2. The pull-in voltage: Location of a saddle-node value. In this section we study the steady-state deflection u, which satisfies

$$\Delta u = \frac{\lambda f(x)}{(1+u)^2}, \quad x \in \Omega; \qquad u = 0, \quad x \in \partial\Omega; \qquad u > -1. \tag{2.1}$$

Here we let $x = (x,y)$, and $\Omega \in \mathbb{R}^2$ is a bounded domain. For several domain shapes Ω and permittivity profiles f, we compute the maximum value of λ, labeled by λ_*, for which (2.1) has a solution. This then determines the pull-in voltage from (1.2). Bounds for λ_* are also obtained. The bounds on λ_* derived below are characterized in terms of the smallest eigenvalue $\mu_0 > 0$, with corresponding eigenfunction ϕ_0, of the Dirichlet eigenvalue problem

$$\Delta\phi + \mu\phi = 0, \quad x \in \Omega; \qquad \phi = 0, \quad x \in \partial\Omega. \tag{2.2}$$

The following result for λ_* was proved in [14].

THEOREM 2.1. *Suppose that $f(x)$ satisfies*

$$0 < C_0 \le f(x) \le 1, \quad x \in \Omega. \tag{2.3}$$

Then, there exists a $\lambda_* < \infty$ *such that there is no solution to* (2.1) *for* $\lambda > \lambda_*$. *Moreover, we have the bound*

$$\lambda_* \le \bar{\lambda}_1 \equiv \frac{4\mu_0}{27C_0}. \tag{2.4}$$

Proof. This is Theorem 3.1 of [14]. We only briefly sketch the proof here. We fix the sign $\phi_0 > 0$ in Ω. We multiply (2.1) by ϕ_0, integrate the resulting equation over Ω, and use Green's identity to get

$$\int_\Omega \left(\mu_0 u + \frac{\lambda f(x)}{(1+u)^2}\right)\phi_0\,dx = 0. \tag{2.5}$$

Since $f(x) \ge C_0 > 0$ and $\phi_0 > 0$, the equality in (2.5) is impossible when

$$\mu_0 u + \frac{\lambda C_0}{(1+u)^2} > 0 \quad \text{for all } u > -1. \tag{2.6}$$

Clearly (2.6) holds for λ sufficiently large, which proves that λ_* is finite. A simple calculation using (2.6) shows that (2.6) holds when $\lambda > \bar{\lambda}_1$, where $\bar{\lambda}_1$ is given in (2.4). ∎

As shown below, the bound (2.4) on λ_* is rather good for the constant permittivity profile $f(x) \equiv 1$. However, since this bound relies on the minimum of $f(x)$ on Ω, it cannot be used to estimate λ_* for the power-law permittivity profile $f(x) = |x|^\alpha$ with $\alpha > 0$ considered in [14]. For such a profile, $C_0 = 0$ in (2.3). Therefore, it is desirable to obtain a bound on λ_* that depends on more global properties of $f(x)$. Such a bound is given in the next result.

THEOREM 2.2. *Suppose that* $f(x)$ *satisfies*

$$0 \le f(x) \le 1, \quad x \in \Omega, \tag{2.7}$$

where $f > 0$ *on a set of positive measure. Then, for some* $\lambda_* < \infty$, *there is no solution to* (2.1) *for* $\lambda > \lambda_*$. *Moreover, in terms of the eigenfunction* ϕ_0 *of* (2.2) *normalized by* $\int_\Omega \phi_0\,dx = 1$, *we have the bound*

$$\lambda_* \le \bar{\lambda}_2 \equiv \frac{\mu_0}{3}\left(\int_\Omega f\phi_0\,dx\right)^{-1}. \tag{2.8}$$

Proof. The proof that λ_* is finite follows from (2.5). To obtain the bound (2.8), we take $\phi_0 > 0$ and we normalize ϕ_0 so that $\int_\Omega \phi_0\,dx = 1$. We then multiply (2.1) by $\phi_0(1+u)^2$ and integrate the resulting equation over Ω to get

$$\int_\Omega \lambda f\phi_0\,dx = \int_\Omega \phi_0(1+u)^2\Delta u\,dx. \tag{2.9}$$

Using the identity $\nabla\cdot(Hg) = g\nabla\cdot H + H\cdot\nabla g$ for any smooth scalar field g and vector field H, together with the divergence theorem, we calculate

$$\int_\Omega \lambda f\phi_0\,dx = \int_{\partial\Omega}(1+u)^2\phi_0\nabla u\cdot\hat{n}\,dS - \int_\Omega \nabla u\cdot\nabla\left[\phi_0(1+u)^2\right]\,dx, \tag{2.10}$$

where $\hat{n}$ is the unit outward normal to $\partial\Omega$. Since $\phi_0 = 0$ on $\partial\Omega$, the first term on the right-hand side of (2.10) vanishes. By calculating the second term on the right-hand

side of (2.10), and noting that $u > -1$, we estimate

$$\int_\Omega \lambda f \phi_0 \, dx = -\int_\Omega 2(1+u)\phi_0 |\nabla u|^2 \, dx - \int_\Omega (1+u)^2 \nabla u \cdot \nabla \phi_0 \, dx \tag{2.11a}$$

$$\leq -\int_\Omega \frac{1}{3} \nabla \phi_0 \cdot \nabla \left[(1+u)^3 \right] \, dx. \tag{2.11b}$$

The right-hand side of (2.11b) is evaluated explicitly, with the result

$$\int_\Omega \lambda f \phi_0 \, dx \leq -\frac{1}{3} \int_{\partial\Omega} (1+u)^3 \nabla \phi_0 \cdot \hat{n} \, dS - \frac{\mu_0}{3} \int_\Omega (1+u)^3 \phi_0 \, dx. \tag{2.12}$$

For $u > -1$, the last term on the right-hand side of (2.12) is positive. Moreover, $u = 0$ on $\partial\Omega$, and from (2.2) we get that $\int_{\partial\Omega} \nabla \phi_0 \cdot \hat{n} \, dS = -\mu_0$ since $\int_\Omega \phi_0 \, dx = 1$. Therefore, if (2.1) has a solution, then, from (2.12), we must have that

$$\lambda \int_\Omega f \phi_0 \, dx \leq \frac{\mu_0}{3}. \tag{2.13}$$

This proves that there is no solution to (2.1) for $\lambda > \bar{\lambda}_2$, where $\bar{\lambda}_2$ is given in (2.8). ∎

2.1. Some explicit examples. We now compute λ_* numerically for several choices of the domain Ω and the permittivity profile $f(x)$. In the computations below we consider three choices for Ω,

$$\begin{aligned} &\Omega : [-1/2, 1/2] \quad \text{(slab)}, \quad \Omega : x^2 + y^2 \leq 1 \quad \text{(unit disk)}, \\ &\Omega : [0, \sqrt{\pi}] \times [0, \sqrt{\pi}] \quad \text{(square)}. \end{aligned} \tag{2.14}$$

The unit disk and the square are chosen to have the same area. To compute the bounds $\bar{\lambda}_1$ and $\bar{\lambda}_2$, we must calculate the first eigenpair μ_0 and ϕ_0 of (2.2), normalized by $\int_\Omega \phi_0 \, dx = 1$. A simple calculation yields that

$$\mu_0 = \pi^2, \qquad \phi_0 = \frac{\pi}{2} \sin\left[\pi \left(x + \frac{1}{2} \right) \right] \quad \text{(slab)}, \tag{2.15a}$$

$$\mu_0 = z_0^2 \approx 5.783, \qquad \phi_0 = \frac{z_0}{J_1(z_0)} J_0(z_0 |x|) \quad \text{(unit disk)}, \tag{2.15b}$$

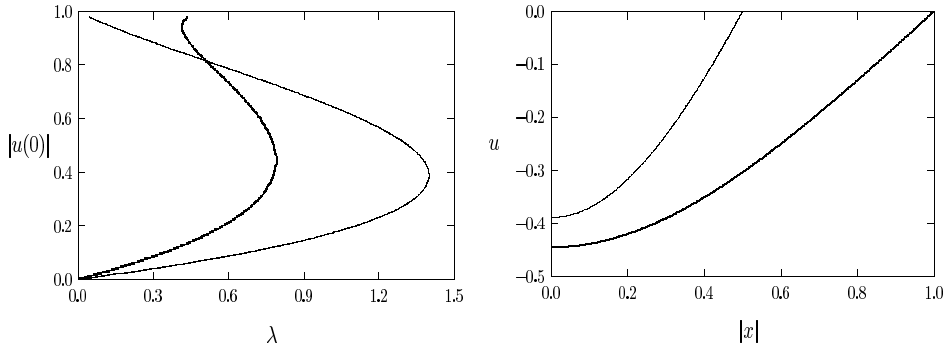
$$\mu_0 = 2\pi, \qquad \phi_0 = \frac{\pi}{4} \sin\left(\sqrt{\pi} x \right) \sin\left(\sqrt{\pi} y \right) \quad \text{(square)}. \tag{2.15c}$$

Here J_0 and J_1 are Bessel functions, and $z_0 \approx 2.4048$ is the first zero of $J_0(z)$. The bounds $\bar{\lambda}_1$ and $\bar{\lambda}_2$ are obtained by substituting (2.15) into (2.4) and (2.8). However, $\bar{\lambda}_2$ must typically be evaluated by a numerical quadrature.

We first consider the constant permittivity profile $f(x) \equiv 1$. For the slab domain, the solution to (2.1) can be reduced to quadrature, and λ_* can be computed from a transcendental equation. To compute λ_* for the unit disk, the scale invariance property of (2.1) can be used as in [11] to reduce the boundary value problem (BVP) (2.1) to an initial value problem, which is then readily solved. Our method for determining λ_* for the disk and the slab uses the BVP solver COLSYS (cf. [1]) with a Newton iteration step to locate λ_*. This approach is similar to that employed in [18]

TABLE 1
Numerical results for the maximum value λ_* of λ for which (2.1) has a solution for the three domains of (2.14). The upper bounds $\bar{\lambda}_1$ and $\bar{\lambda}_2$ on λ_* given in (2.4) and (2.8) are also shown.

Ω	λ_*	$\bar{\lambda}_1$	$\bar{\lambda}_2$
(slab)	1.401	1.462	3.290
(unit disk)	0.789	0.857	1.928
(square)	0.857	0.931	2.094



(a) $|u(0)|$ versus λ (b) u versus $|x|$

FIG. 2. (a) *Plot of $|u(0)|$ versus λ for the unit disk (heavy solid curve) and for the slab (lighter curve).* (b) *Plot of u versus $|x|$ at $\lambda = \lambda_*$ for the unit disk (heavy solid curve) and the slab (lighter curve). For both figures we have taken the constant permittivity profile $f \equiv 1$.*

for Arrhenius nonlinearities and is useful for computing λ_* below for spatially varying permittivity profiles. For the square domain, we compute λ_* using the nonlinear elliptic solver PLTMG (cf. [2]), which uses a finite-element discretization of (2.1) together with path-following methods to compute the solution as λ is varied. This software package allows for the accurate computation of saddle-node bifurcation points. In Table 1 we give numerical results for λ_* for the three domains of (2.14) together with numerical values for the bounds $\bar{\lambda}_1$ and $\bar{\lambda}_2$. Notice that the bound for $\bar{\lambda}_1$ is rather close to λ_*, and is better than that of $\bar{\lambda}_2$. In Figure 2(a) we plot the bifurcation diagram $|u(0)|$ versus λ for the slab domain and for the unit disk. For $\lambda = \lambda_*$, in Figure 2(b) we plot u versus $|x|$ for the slab domain and for the unit disk. In Figure 3(a) we plot the bifurcation diagram for the square domain. For this domain, in Figure 3(b) we show a surface plot of u versus (x, y) when $\lambda = \lambda_*$. The computations were done with 1152 finite elements.

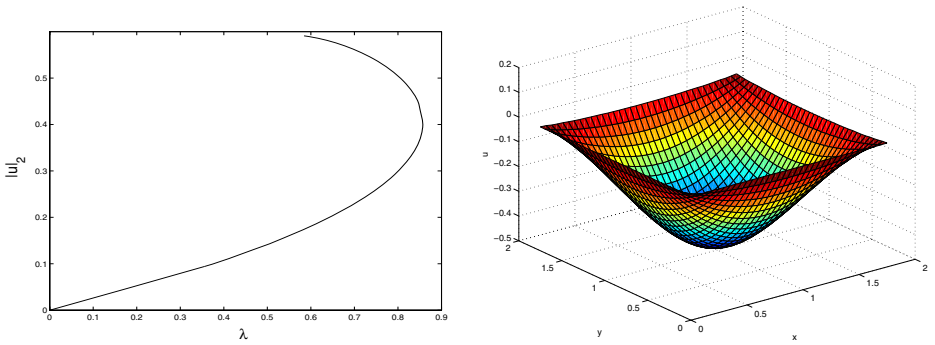
For each of the domains of (2.14), we now calculate λ_* for the following two forms of the permittivity profile $f(x)$:

$$\text{(2.16a)} \qquad \text{(slab)}: \quad f(x) = |2x|^\alpha \quad \text{(power-law)}, \quad f(x) = e^{\alpha(x^2-1/4)} \quad \text{(exponential)},$$

$$\text{(2.16b)} \quad \text{(unit disk)}: \quad f(x) = |x|^\alpha \quad \text{(power-law)}, \quad f(x) = e^{\alpha(|x|^2-1)} \quad \text{(exponential)},$$

$$\text{(2.16c)} \quad \text{(square)}: \quad f(x) = \left(\frac{2}{\pi}\right)^{\alpha/2} |x - x_0|^\alpha \quad \text{(power-law)},$$

$$f(x) = \exp\left(\alpha\left(\frac{2|x - x_0|^2}{\pi} - 1\right)\right),$$



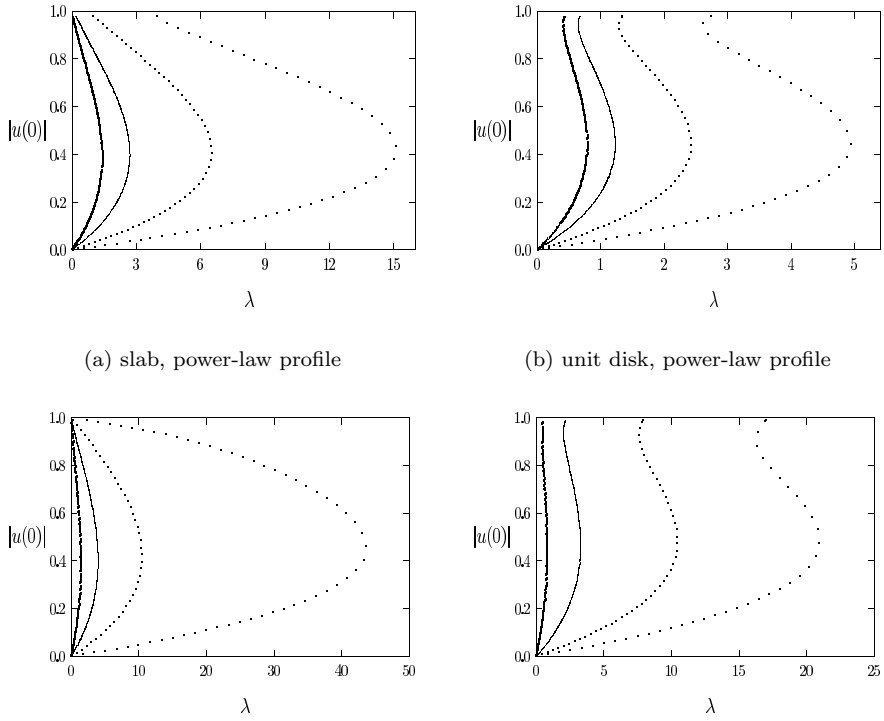
(a) $|u|_2$ versus λ (b) u versus (x, y) at $\lambda = \lambda_*$

FIG. 3. (a) *Plot of the bifurcation diagram of the* L_2 *norm* $|u|_2$ *versus* λ *for a square domain. We do not show any secondary bifurcations.* (b) *Surface plot of* u *versus* $x = (x, y)$ *when* $\lambda = \lambda_* \approx .857$. *For these figures the permittivity profile is* $f(x) \equiv 1$.

where $\alpha > 0$. In (2.16b) and (2.16c), $x \equiv (x, y)$ and $x_0 = (\sqrt{\pi}/2, \sqrt{\pi}/2)$ is the center of the square. For the domains of (2.14), we note that $0 \le f(x) \le 1$ for $x \in \Omega$. In addition, both the power-law and exponential profiles satisfy the property that the minimum of $f(x)$ occurs at the point where the deflection u of the upper membrane in Figure 1 is the largest. This effect leads to larger values of λ_* and, from (1.2), it increases the pull-in voltage. Physically, from (1.3), this corresponds to tailoring the dielectric permittivity ε_2 of the upper membrane so that ε_2 is significantly larger than the free-space permittivity ε_0 in regions where the membrane deflection will be largest. This idea of modifying the dielectric permittivity ε_2 to increase both λ_* and the pull-in distance was first introduced and studied in [14] for the slab and disk domains. For these domains, it was shown in [14] that (2.1) has a scaling invariance property under a power-law profile for $f(x)$. This property, which reduces (2.1) to the study of an ordinary differential equation (ODE), was used in [14] to give a detailed analysis of the bifurcation diagram of (2.1) for the slab and disk domains. Although the power-law profile for $f(x)$ is mathematically very convenient as a result of the scale invariance property, it is not so realistic from a modeling perspective, in that it predicts an infinite dielectric permittivity ε_2 at the center of the membrane. The exponential profile in (2.16) does not have this artifact of an infinite membrane permittivity.

For four values of α, in Figure 4 we plot the bifurcation diagram $|u(0)|$ versus λ for both the power-law and the exponential profiles. The plots are shown for both the slab and the unit disk. The bifurcation diagram of the steady-state problem shown in this figure is typical, in that the transition from existence to nonexistence is due to the first fold. A more detailed study of the bifurcation diagram for a slab geometry under a power-law profile was made in [14]. In [14] it was shown that for $0 \le \alpha < \alpha_c$, where $\alpha_c \equiv -\frac{1}{2} + \frac{1}{2}\sqrt{\frac{27}{2}}$, there is exactly one saddle-node point, and so at most two solutions to (2.1). Alternatively, for $\alpha > \alpha_c$, the bifurcation diagram has an infinite number of fold points, which tend to a common limiting value λ_{*c} as $u(0) \to -1^+$. Although the details of the solution multiplicity obtained in [14] are very interesting, they are not germane to the determination of λ_*.

In Figure 5(a) we plot the saddle-node value λ_* versus α for the slab domain. A



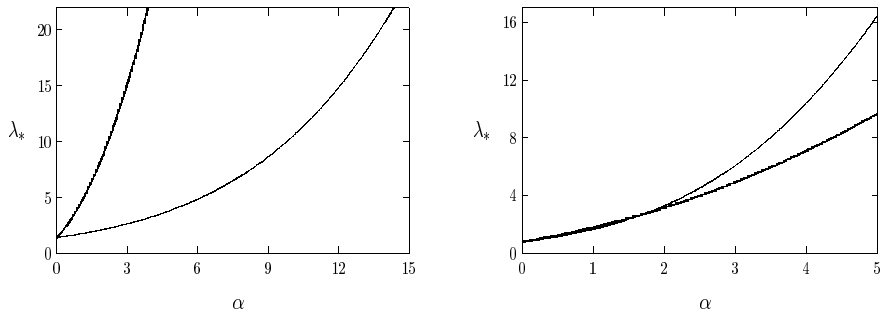
(a) slab, power-law profile

(b) unit disk, power-law profile

(c) slab, exponential profile

(d) unit disk, exponential profile

FIG. 4. *The bifurcation diagram $|u(0)|$ versus λ for the slab and the unit disk, and for both the power-law and exponential profiles.* (a) $\alpha = 0$, $\alpha = 0.5$, $\alpha = 1.5$, $\alpha = 3.0$. (b) $\alpha = 0$, $\alpha = 0.5$, $\alpha = 1.5$, $\alpha = 3.0$. (c) $\alpha = 0$, $\alpha = 5$, $\alpha = 10$, $\alpha = 19$. (d) $\alpha = 0$, $\alpha = 2$, $\alpha = 4$, $\alpha = 5.6$. *In each figure the first saddle-node value increases with α.*



(a) λ_* versus α (slab)

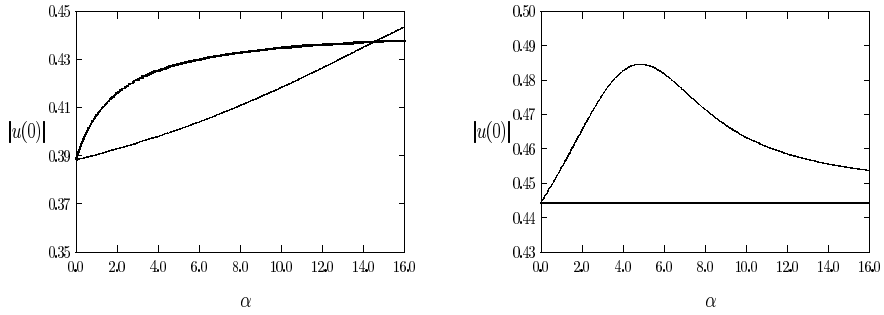
(b) λ_* versus α (unit disk)

FIG. 5. *Plots of λ_* versus α for a power-law profile (heavy solid curve) and the exponential profile (lighter curve).* (a) *corresponds to the slab domain, while* (b) *corresponds to the unit disk.*

similar plot is shown in Figure 5(b) for the unit disk. The numerical computations were done using COLSYS [1] to solve the BVP (2.1) and Newton's method to determine the saddle-node point. Although Theorem 2.2 guarantees a pull-in voltage for any $\alpha > 0$, λ_* is seen to increase rapidly with α. Therefore, by increasing α, or equivalently by increasing the spatial extent where $f(x) \ll 1$, one can increase the

TABLE 2
Comparison of numerical values for λ_* with the bounds $\bar{\lambda}_1$ and $\bar{\lambda}_2$ given in (2.4) and (2.8) for the exponential permittivity profile.

Ω	α	λ_*	$\bar{\lambda}_1$	$\bar{\lambda}_2$
(slab)	1.0	1.733	1.878	4.023
(slab)	3.0	2.637	3.095	5.965
(slab)	6.0	4.848	6.553	10.50
(slab)	10.0	10.40	17.81	21.14
(unit disk)	0.5	1.153	1.413	2.706
(unit disk)	1.0	1.661	2.329	3.746
(unit disk)	2.0	3.296	6.331	6.864
(unit disk)	3.0	6.091	17.21	11.86



(a) slab (b) unit disk

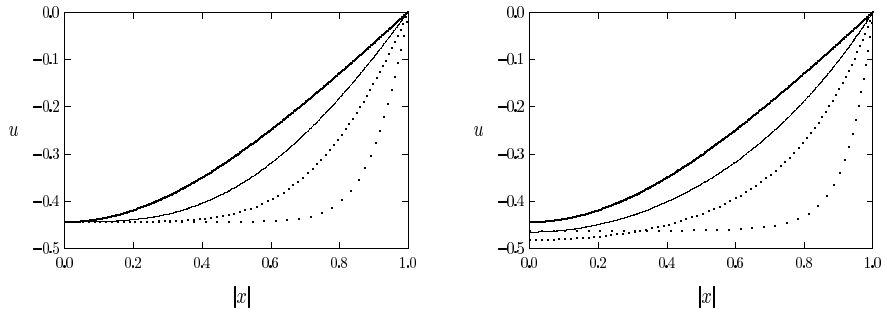
FIG. 6. *Plots of the pull-in distance $|u(0)|$ versus α for the power-law profile (heavy solid curve) and the exponential profile (lighter curve).* (a) *The slab domain.* (b) *The unit disk.*

stable operating range of the MEMS capacitor. In Table 2 we give numerical results for λ_* together with the bounds $\bar{\lambda}_1$ and $\bar{\lambda}_2$ for the exponential permittivity profile computed from (2.4), (2.8), (2.15), and (2.16). A numerical quadrature is used to evaluate the integral defining $\bar{\lambda}_2$. From this table, we observe that the bound $\bar{\lambda}_1$ for λ_* is better than $\bar{\lambda}_2$ for small values of α. However, for $\alpha \gg 1$, we can use Laplace's method on the integral defining $\bar{\lambda}_2$ to obtain for the exponential permittivity profile that

$$\bar{\lambda}_1 = \frac{4b_1^2}{27} e^{c_1\alpha}, \qquad \bar{\lambda}_2 \sim c_2\alpha^2. \tag{2.17}$$

Here $b_1 = \pi^2$, $c_1 = 1/4$, $c_2 = 1/3$ for the slab domain, and $b_1 = z_0^2$, $c_1 = 1$, $c_2 = 4/3$ for the unit disk, where z_0 is the first zero of $J_0(z) = 0$. Therefore, for $\alpha \gg 1$, the bound $\bar{\lambda}_2$ is better than $\bar{\lambda}_1$. A similar calculation can be done for the power-law profile. Recall for the power-law profile that $\bar{\lambda}_1$ is undefined. However, by using Laplace's method, we readily obtain for $\alpha \gg 1$ that $\bar{\lambda}_2 \sim \alpha^2/3$ for the unit disk and $\bar{\lambda}_2 \sim 4\alpha^2/3$ for the slab domain.

Next, we compute the pull-in distance for a slab domain for both the power-law and the exponential permittivity profiles. The pull-in distance, defined as the value of $|u(0)|$ at the fold point $\lambda = \lambda_*$, gives the maximum stable steady-state membrane deflection that can be achieved. For the slab domain, in Figure 6(a) we plot $|u(0)|$ versus α for both the power-law and the exponential conductivity profile. For the power-law profile, the plot of $|u(0)|$ versus α is equivalent to that in Figure 5.1 of



(a) power-law profile (b) exponential profile

FIG. 7. (a) *Plots of u versus $|x|$ at $\lambda = \lambda_*$ for $\alpha = 0$, $\alpha = 1$, $\alpha = 3$, and $\alpha = 10$, in the unit disk for the power-law profile.* (b) *Plots of u versus $|x|$ at $\lambda = \lambda_*$ for $\alpha = 0$, $\alpha = 2$, $\alpha = 4$, and $\alpha = 10$, in the unit disk for the exponential profile. In both figures the solution develops a boundary-layer structure near $|x| = 1$ as α is increased.*

[14]. A similar plot of $|u(0)|$ versus α is shown in Figure 6(b) for the unit disk. For the power-law profile in the unit disk we observe that $|u(0)| \approx 0.444$ for any $\alpha > 0$. Therefore, rather curiously, the power-law profile does not increase the pull-in distance for the unit disk. For the exponential profile we observe from Figure 6(b) that the pull-in distance is not a monotonic function of α. The maximum value occurs at $\alpha \approx 4.8$, where $\lambda_* \approx 15.11$ (see Figure 5(b)) and $|u(0)| = 0.485$. For $\alpha = 0$, we have $\lambda_* \approx 0.789$ and $|u(0)| = 0.444$. Therefore, since λ_* is proportional to V^2 from (1.2), we conclude that the exponential permittivity profile for the unit disk can increase the pull-in distance by roughly 9% if the voltage is increased by roughly a factor of four.

For device design purposes one of the primary goals is to maximize the pull-in distance over a certain allowable voltage range that is set by the power supply. To address this problem it would be interesting to formulate an optimization problem that computes a dielectric permittivity $f(x)$ that maximizes the pull-in distance for a prescribed range of the saddle-node threshold λ_*. However, such an optimization problem is beyond the scope of this study.

For the unit disk, in Figure 7(a) we plot u versus $|x|$ at $\lambda = \lambda_*$ for four values of α for the power-law profile. Notice that $u(0)$ is the same for each of these values of α. A similar plot is shown in Figure 7(b) for the exponential permittivity profile. From these figures, we observe that u has a boundary-layer structure when $\alpha \gg 1$. In this limit, $f(x) \ll 1$ except in a narrow zone near the boundary of the domain. For $\alpha \gg 1$ the pull-in distance $|u(0)|$ also reaches some limiting value (see Figures 6(a), 6(b), and 7). For the slab domain with an exponential permittivity profile, we remark that the limiting asymptotic behavior of $|u(0)|$ for $\alpha \gg 1$ is beyond the range shown in Figure 6(a).

For $\alpha \gg 1$, we now use a boundary-layer analysis to determine a scaling law for λ_* for both types of permittivity profiles and for either a slab domain or the unit disk. We illustrate the analysis for a power-law permittivity profile in the unit disk. For $\alpha \gg 1$, there is an outer region defined by $0 \le r \ll 1 - O(\alpha^{-1})$, and an inner region where $r - 1 = O(1/\alpha)$. In the outer region, where $\lambda r^\alpha \ll 1$, (2.1) reduces asymptotically to $\Delta u = 0$. Therefore, the leading-order outer solution is a constant

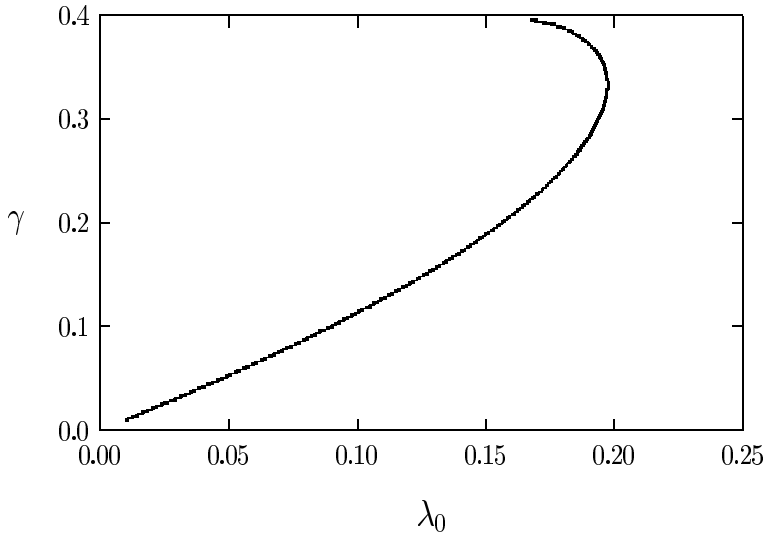


FIG. 8. *Bifurcation diagram of* $w'(0) = -\gamma$ *versus* λ_0 *from the numerical solution of* (2.19).

$u = A$. In the inner region, we introduce new variables w and ρ by

$$w(\rho) = u\left(1 - \frac{\rho}{\alpha}\right), \qquad \rho = \alpha(1 - r). \tag{2.18}$$

Substituting (2.18) into (2.1) with $f(r) = r^\alpha$, using the limiting behavior $(1 - \rho/\alpha)^\alpha \to e^{-\rho}$ as $\alpha \to \infty$, and defining $\lambda = \alpha^2\lambda_0$, we obtain the leading-order boundary-layer problem

$$w'' = \frac{\lambda_0 e^{-\rho}}{(1+w)^2}, \quad 0 \le \rho < \infty; \quad w(0) = 0, \quad w'(\infty) = 0, \qquad \lambda = \alpha^2\lambda_0. \tag{2.19}$$

In terms of the solution to (2.19), the leading-order outer solution is $u = A = w(\infty)$.

We define γ by $w'(0) = -\gamma$, for $\gamma > 0$, and we solve (2.19) numerically using COLSYS [1] to determine $\lambda_0 = \lambda_0(\gamma)$. In Figure 8 we plot $\lambda_0(\gamma)$ and show that this curve has a saddle-node point at $\lambda_0 = \lambda_{0*} \equiv 0.1973$. At this value, we compute $w(\infty) \approx 0.445$, which sets the limiting membrane deflection for $\alpha \gg 1$. Therefore, for $\alpha \gg 1$, the saddle-node value, from (2.19), has the scaling law behavior $\lambda_* \sim 0.1973\alpha^2$ for a power-law profile in the unit disk. A similar boundary-layer analysis can be done to determine the scaling law for λ_* when $\alpha \gg 1$ for the other cases. In each case we can relate λ_* to the saddle-node value of the boundary-layer problem (2.19). In this way, for $\alpha \gg 1$, we obtain

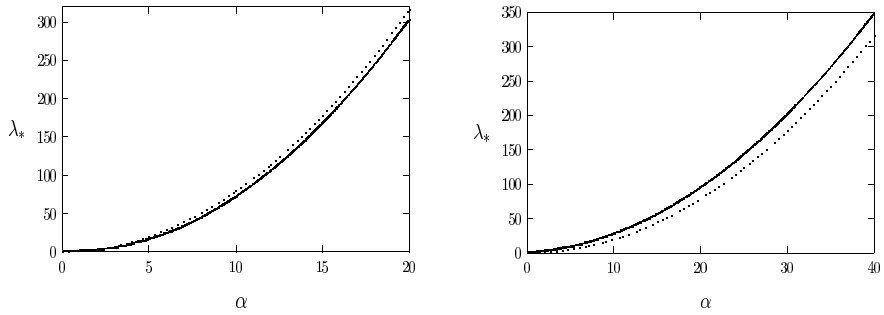
(2.20a)

$$\lambda_* \sim 4(0.1973)\alpha^2, \quad \bar{\lambda}_2 \sim \frac{4\alpha^2}{3} \quad \text{(power-law, slab) or (exponential, unit disk)},$$

(2.20b)

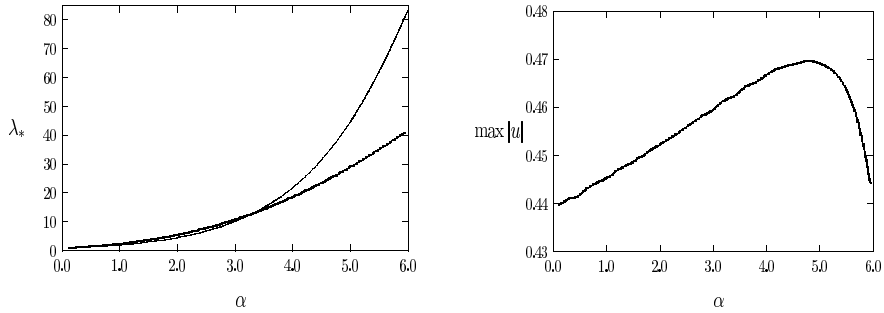
$$\lambda_* \sim (0.1973)\alpha^2, \quad \bar{\lambda}_2 \sim \frac{\alpha^2}{3} \quad \text{(power-law, unit disk) or (exponential, slab)}.$$

Notice that $\bar{\lambda}_2 = O(\alpha^2)$, with a factor that is about 5/3 times as large as the multiplier of α^2 in the asymptotic formula for λ_*. For an exponential profile and a power-law



(a) exponential profile (b) power law profile

FIG. 9. *Comparison of numerically computed λ_* (heavy solid curve) with the asymptotic result (dotted curve) from* (2.20) *for the unit disk.* (a) *The exponential profile.* (b) *The power-law profile.*



(a) λ_* (b) $\max|u|$

FIG. 10. (a) λ_* *versus* α *for the square domain. The power-law profile is the heavy solid curve, and the exponential profile is the lighter curve.* (b) *The pull-in distance* $\max|u|$ *at* $\lambda = \lambda_*$ *versus* α *for the exponential profile in the square domain.*

TABLE 3

Numerical results for λ_* versus α for the exponential and power-law permittivity profiles of (2.16c). *The computations are for the square domain* $[0, \sqrt{\pi}] \times [0, \sqrt{\pi}]$.

α	λ_* (power-law)	λ_* (exponential)
0.5	1.523	1.314
1.0	2.485	2.005
2.0	5.607	4.589
3.0	10.85	10.21
4.0	18.67	21.89
5.0	29.04	44.61
6.0	41.67	83.31
7.0	56.38	132.6

profile in the unit disk, in Figure 9(a) and Figure 9(b), respectively, we show the close agreement between the full numerical value of λ_* and the asymptotic result (2.20).

For the square domain we use PLTMG (cf. [2]) to compute λ_* as a function of α. In Figure 10(a) we plot λ_* versus α for both the power-law and exponential profiles of (2.16c). In Table 3 we give numerical results for λ_* at different values of α for both

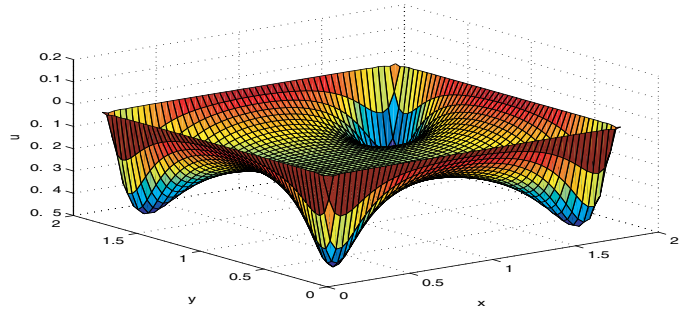


FIG. 11. *Plot of the numerical solution for u in the square domain at the fold point λ_* for the exponential permittivity profile with $\alpha = 8$. The maximum deflection now occurs near each of the four corners of the square, where the dielectric permittivity function f is the largest.*

profiles. The computations were done with 3200 finite elements. From this table we observe that λ_* increases rapidly with α. In Figure 10(b) we plot the pull-in distance $\max|u|$ versus α at $\lambda = \lambda_*$ for the exponential permittivity profile. This curve has the same qualitative shape as for the case of the unit disk shown in Figure 6(b). For the values of α shown in Figure 10(b) the maximum deflection occurs at the center of the square. For the power-law profile our numerical computations (not shown) indicate that, as for the case of the unit disk in Figure 6(a), the membrane deflection at $\lambda = \lambda_*$ is essentially independent of α, provided that α is not too large.

For $\alpha \gg 1$ the solution u under either a power-law or an exponential dielectric permittivity profile develops strong gradients in the localized regions where $f(x, y) \approx 1$. In contrast to the case of the unit disk, where $f \approx 1$ near the boundary $r = 1$, for the square domain we have $f \approx 1$ for $\alpha \gg 1$ only in small neighborhoods near each of the four corners of the square. Away from these corners we have $f \ll 1$ when $\alpha \gg 1$. In Figure 11, where we plot the numerical solution for u in the square domain at the fold point λ_* for the exponential permittivity profile with $\alpha = 8$, we observe that the maximum deflection now occurs near each of the corners of the square.

3. Touchdown behavior. We now study touchdown, or quenching, behavior for (1.1). The solution u of (1.1) is said to touchdown at finite time if the minimum value of u reaches -1 at some $t = T_* < \infty$. In section 3.1 we determine bounds on λ for which touchdown occurs in finite time. In section 3.2 we analytically construct the local touchdown profile for the case $f(x) \equiv 1$. For $f(x) \equiv 1$, in section 3.3 we briefly outline the construction of the local touchdown profile by using a formal center manifold analysis of a PDE, similar to that in [5] and [4], that results from a similarity group transformation of (1.1).

3.1. Bounds on touchdown behavior. Let μ_0 and ϕ_0 be the smallest eigenpair of (2.2). The first result is a minor modification of a key result in [6].

THEOREM 3.1. *Suppose that $f(x)$ satisfies*

$$0 < C_0 \le f(x) \le 1, \quad x \in \Omega, \tag{3.1}$$

and that $\lambda > \bar{\lambda}_1 \equiv \frac{4\mu_0}{27C_0}$. Then, the solution u of (1.1) reaches $u = -1$ at finite time.

Proof. Without loss of generality we assume that $\phi_0 > 0$ in Ω, and we normalize ϕ_0 so that $\int_\Omega \phi_0\, dx = 1$. Multiplying (1.1) by ϕ_0 and integrating over the domain, we

obtain

$$\frac{d}{dt}\int_\Omega \phi_0 u\,dx = \int_\Omega \phi_0 \Delta u\,dx - \int_\Omega \frac{\lambda\phi_0 f(x)}{(1+u)^2}\,dx. \tag{3.2}$$

Using Green's theorem, together with the lower bound in (3.1), we get

$$\frac{d}{dt}\int_\Omega \phi_0 u\,dx \le -\mu_0\int_\Omega \phi_0 u\,dx - \lambda C_0\int_\Omega \frac{\phi_0}{(1+u)^2}\,dx. \tag{3.3}$$

Next, we define an energy-like variable $E(t)$ by $E(t) = \int_\Omega \phi_0 u\,dx$, where $E(0) = 0$, so that

$$E(t) = \int_\Omega \phi_0 u\,dx \ge \inf_\Omega u \int_\Omega \phi_0\,dx = \inf_\Omega u. \tag{3.4}$$

Then, using Jensen's inequality on the second term on the right-hand side of (3.3), we obtain

$$\frac{dE}{dt} + \mu_0 E \le -\frac{\lambda C_0}{(1+E)^2}, \qquad E(0) = 0. \tag{3.5}$$

We then compare $E(t)$ with the solution $F(t)$ of

$$\frac{dF}{dt} + \mu_0 F = -\frac{\lambda C_0}{(1+F)^2}, \qquad F(0) = 0. \tag{3.6}$$

It then follows from standard comparison principles that $E(t) \le F(t)$ on their domains of existence. Therefore,

$$\inf_\Omega u \le E(t) \le F(t). \tag{3.7}$$

Next, we separate variables in (3.6) to determine t in terms of F. The touchdown time T_1 for F is obtained by setting $F = -1$ in the resulting formula. In this way, we get

$$T_1 \equiv \int_{-1}^{0} \left[\mu_0 s + \frac{\lambda C_0}{(1+s)^2}\right]^{-1} ds. \tag{3.8}$$

The touchdown time T_1 is finite when the integral in (3.8) converges. A simple calculation shows that this occurs when $\lambda > \bar{\lambda}_1 \equiv \frac{4\mu_0}{27C_0}$. Hence if T_1 is finite, then (3.7) implies that the touchdown time T_* of (1.1) must also be finite. Therefore, when $\lambda > \bar{\lambda}_1 = \frac{4\mu_0}{27C_0}$, we have that $T_* < T_1$, where T_1 is given in (3.8). □

Recalling Theorem 3.1 in [14], which was summarized in Theorem 2.1, we conclude that not only is there no steady-state solution for (1.1) when $\lambda > \bar{\lambda}_1$, but the corresponding time-dependent solution of (1.1) touches down in finite time. We are not able to obtain any theoretical information on touchdown behavior for the range $\lambda_* < \lambda < \bar{\lambda}_1$. The next result, using the approach of Theorem 2.2, establishes touchdown behavior for more general permittivity profiles $f(x)$, such as the power-law profile, that vanish at certain points in Ω.

THEOREM 3.2. *Suppose that f satisfies* (2.7)*, and that $\lambda > \bar{\lambda}_2$, where $\bar{\lambda}_2$ is defined in* (2.8)*. Then, the solution u of* (1.1) *reaches -1 at finite time.*

Proof. Let ϕ_0 and μ_0 be the smallest eigenpair of (2.2). We fix the sign $\phi_0 > 0$ in Ω, and we normalize ϕ_0 so that $\int_\Omega \phi_0\, dx = 1$. We multiply (1.1) by $\phi_0(1+u)^2$, and integrate the resulting equation over Ω to get

$$\frac{d}{dt}\int_\Omega \frac{\phi_0}{3}(1+u)^3\, dx = \int_\Omega \phi_0(1+u)^2\Delta u\, dx - \int_\Omega \lambda f\phi_0\, dx. \tag{3.9}$$

We calculate the first term on the right-hand side of (3.9) as in the proof of Theorem 2.2 to get

$$\begin{aligned}
&\frac{d}{dt}\int_\Omega \frac{\phi_0}{3}(1+u)^3\, dx \qquad\qquad (3.10a)\\
&\quad = -\int_\Omega \nabla u\cdot\nabla\left[\phi_0(1+u)^2\right]\, dx - \int_\Omega \lambda f\phi_0\, dx\\
&\quad = -\int_\Omega 2(1+u)\phi_0|\nabla u|^2\, dx - \int_\Omega \frac{1}{3}\nabla\phi_0\cdot\nabla\left[(1+u)^3\right]\, dx - \int_\Omega \lambda f\phi_0\, dx \qquad (3.10b)\\
&\quad \le -\frac{1}{3}\int_{\partial\Omega}\nabla\phi_0\cdot\hat{n}\, dS - \frac{\mu_0}{3}\int_\Omega (1+u)^3\phi_0\, dx - \int_\Omega \lambda f\phi_0\, dx. \qquad (3.10c)
\end{aligned}$$

Since $\int_\Omega \nabla\phi_0\cdot\hat{n}\, dS = -\mu_0$ and $u \ge -1$, we further estimate from (3.10c) that

$$\frac{dE}{dt} + \mu_0 E \le R, \qquad R \equiv \frac{\mu_0}{3} - \lambda\int_\Omega f\phi_0\, dx, \qquad E \equiv \frac{1}{3}\int_\Omega \phi_0(1+u)^3\, dx. \tag{3.11}$$

Next, we compare $E(t)$, which satisfies $E(0) = 1/3$, with the solution $F(t)$ of

$$\frac{dF}{dt} + \mu_0 F = R, \qquad F(0) = \frac{1}{3}. \tag{3.12}$$

By standard comparison principles and the definition of E, we obtain

$$\frac{1}{3}\inf_\Omega (1+u)^3 \le E(t) \le F(t). \tag{3.13}$$

Assume that $\lambda > \bar{\lambda}_2$, where $\bar{\lambda}_2$ is defined in (2.8). For this range of λ, we have that $R < 0$ in (3.11) and (3.12). For $R < 0$, we have that $F = 0$ at some finite time $t = T_2$. From (3.13), this implies that $E = 0$ at finite time. Thus, u has touchdown at some finite time $T_* < T_2$. Then, by calculating T_2 explicitly, we get the following bound on T_*:

$$T_* < T_2 \equiv -\frac{1}{\mu_0}\log\left[1 - \frac{\mu_0}{3\lambda}\left(\int_\Omega f\phi_0\, dx\right)^{-1}\right]. \tag{3.14}$$

The operation of a microvalve in MEMS technology explicitly exploits the existence of touchdown behavior in order to open and close a switch (see section 7.6 of [13]). The bounds on the touchdown time above relate to the time it takes to open such a valve, and thereby gives an estimate on the switching speed. To estimate the switching speed as a function of α, we label $I(\alpha) \equiv \int_\Omega f\phi_0\, dx$. For both the power-law and exponential permittivity profiles we calculate that $I'(\alpha) > 0$ and that $I(\alpha) \sim c\alpha^{-2}$ for $\alpha \gg 1$ for some $c > 0$. Therefore, from (3.14) we obtain that T_2 is an increasing function of α and that $T_2'(\alpha) \sim 2\alpha/(3c\lambda)$ for $\alpha \gg 1$. This suggests that the switching speed decreases as α increases. ∎

3.2. The touchdown profile $f(x) \equiv 1$: Transformed problem. We now construct a local expansion of the solution near the touchdown time and touchdown location by adapting the method of [8] used to analyze blow-up behavior. In the analysis below we assume that touchdown occurs at $x = 0$ and $t = T$. In the absence of diffusion, the time-dependent behavior of (1.1) is given by $u_t = -\lambda(1+u)^{-2}$. Integrating this differential equation and setting $u(T) = -1$, we get $(1+u)^3 = -3\lambda(t-T)$. This solution motivates the introduction of a new variable $v(x,t)$ defined in terms of $u(x,t)$ by

$$v = \frac{1}{3\lambda}(1+u)^3. \tag{3.15}$$

Notice that $u = -1$ maps to $v = 0$. In terms of v, (1.1) transforms exactly to

$$v_t = \Delta v - \frac{2}{3v}|\nabla v|^2 - 1, \quad x \in \Omega; \qquad v = \frac{1}{3\lambda}, \quad x \in \partial\Omega; \qquad v = \frac{1}{3\lambda}, \quad t = 0. \tag{3.16}$$

We will find a formal power series solution to (3.16) near $v = 0$ in dimension $N = 1$ and $N = 2$.

As in [8] we look for a locally radially symmetric solution to (3.16) in the form

$$v(x,t) = v_0(t) + \frac{r^2}{2!}v_2(t) + \frac{r^4}{4!}v_4(t) + \cdots, \tag{3.17}$$

where $r = |x|$. In dimension $N = 1$, such a form implies that the touchdown profile is locally even. We then substitute (3.17) into (3.16) and collect coefficients in r. In this way, we obtain the following coupled ODEs for v_0 and v_2:

$$v_0' = -1 + Nv_2, \qquad v_2' = -\frac{4}{3v_0}v_2^2 + \frac{(N+2)}{3}v_4. \tag{3.18}$$

We are interested in the solution to this system for which $v_0(T) = 0$, with $v_0' < 0$ and $v_2 > 0$ for $T - t > 0$ with $T - t \ll 1$. The system (3.18) has a closure problem in that v_2 depends on v_4. However, we will assume that $v_4 \ll v_2^2/v_0$ near the singularity. With this assumption, (3.18) reduces to

$$v_0' = -1 + Nv_2, \qquad v_2' = -\frac{4}{3v_0}v_2^2. \tag{3.19}$$

We now solve the system (3.19) asymptotically as $t \to T^-$ in a manner similar to that used in [8]. We first assume that $Nv_2 \ll 1$ near $t = T$. This leads to $v_0 \sim T - t$ and the following differential equation for v_2:

$$v_2' \sim \frac{-4}{3(T-t)}v_2^2 \quad \text{as } t \to T^-. \tag{3.20}$$

By integrating (3.20), we obtain that

$$v_2 \sim -\frac{3}{4\left[\log(T-t)\right]} + \frac{B_0}{\left[\log(T-t)\right]^2} + \cdots \quad \text{as } t \to T^-, \tag{3.21}$$

for some unknown constant B_0. From (3.21), we observe that the consistency condition that $Nv_2 \ll 1$ as $t \to T^-$ is indeed satisfied. Substituting (3.21) into (3.19) for

v_0, we obtain for $t \to T^-$ that

$$v_0' = -1 + N\left(-\frac{3}{4\left[\log(T-t)\right]} + \frac{B_0}{\left[\log(T-t)\right]^2} + \cdots\right). \tag{3.22}$$

Using the method of dominant balance, we look for a solution to (3.22) as $t \to T^-$ in the form

$$v_0 \sim (T-t) + (T-t)\left[\frac{C_0}{\left[\log(T-t)\right]} + \frac{C_1}{\left[\log(T-t)\right]^2} + \cdots\right], \tag{3.23}$$

for some C_0 and C_1 to be found. A simple calculation yields that

$$v_0 \sim (T-t) - \frac{3N(T-t)}{4|\log(T-t)|} - \frac{N\left(B_0 - 3/4\right)(T-t)}{|\log(T-t)|^2} + \cdots \quad \text{as } t \to T^-. \tag{3.24}$$

The local form for v near touchdown is $v \sim v_0 + r^2 v_0/2$. Using the leading term in v_2 from (3.21) and the first two terms in v_0 from (3.24), we obtain the local form

$$v \sim (T-t)\left[1 - \frac{3N}{4|\log(T-t)|} + \frac{3r^2}{8(T-t)|\log(T-t)|} + \cdots\right], \tag{3.25}$$

for $r \ll 1$ and $t - T \ll 1$. Finally, using the nonlinear mapping (3.15) relating u and v, we conclude that

$$u \sim -1 + \left[3\lambda(T-t)\right]^{1/3}\left(1 - \frac{3N}{4|\log(T-t)|} + \frac{3r^2}{8(T-t)|\log(T-t)|} + \cdots\right)^{1/3}. \tag{3.26}$$

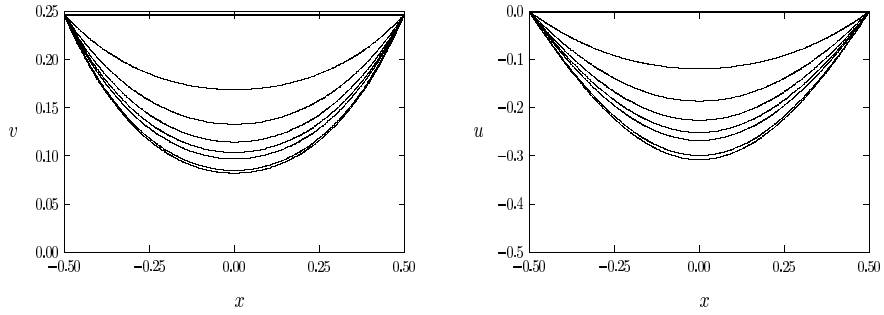
We note, as in [8], that if we use the local behavior $v \sim (T-t) + 3r^2/[8|\log(T-t)|]$, we get that

$$\frac{|\nabla v|^2}{v} \sim \left[\frac{2}{3}|\log(T-t)| + \frac{16(T-t)|\log(T-t)|^2}{9r^2}\right]^{-1}. \tag{3.27}$$

Hence, the term $|\nabla v|^2/v$ in (3.16) is bounded for any r, even as $t \to T^-$. This allows us to use a simple finite-difference scheme to compute numerical solutions to (3.16). With this observation, we now perform a few numerical experiments on the transformed problem (3.16). For the slab domain, we define v_j^m for $j = 1, \ldots, N+2$ to be the discrete approximation to $v(m\Delta t, -1/2 + (j-1)h)$, where $h = 1/(N+1)$ and Δt are the spatial and temporal mesh sizes, respectively. A second-order accurate in space and first-order accurate in time discretization of (3.16) is

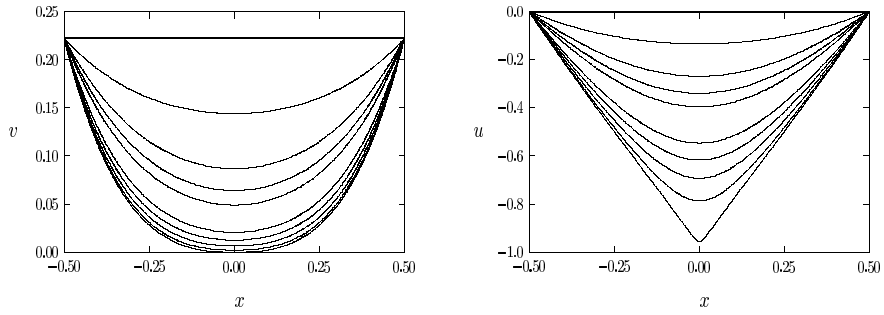
$$v_j^{m+1} = v_j^m + \Delta t\left(\frac{\left(v_{j+1}^m - 2v_j^m + v_{j-1}^m\right)}{h^2} - 1 - \frac{\left(v_{j+1}^m - v_{j-1}^m\right)^2}{6v_j^m h^2}\right), \quad j = 2, \ldots, N+1, \tag{3.28}$$

with $v_1^m = v_{N+2}^m = (3\lambda)^{-1}$ for $m \geq 0$. The initial condition is $v_j^0 = (3\lambda)^{-1}$ for $j = 1, \ldots, N+2$. The time-step Δt is chosen to satisfy $\Delta t < h^2/4$ for the stability of the discrete scheme.



(a) v versus x (b) u versus x

FIG. 12. *Experiment* 1: *For the slab domain and* $\lambda = 1.35 < \lambda_*$ *we plot* v *and* u *versus* x *at times* $t = (0, 0.1, 0.2, 0.3, 0.4, 0.5, 1.0, 3.0)$ *from the discrete scheme* (3.28) *with* $N = 200$ *and* $\Delta t = 0.6 \times 10^{-5}$. *Both* v *and* u *decrease towards a steady-state solution as* t *increases.*



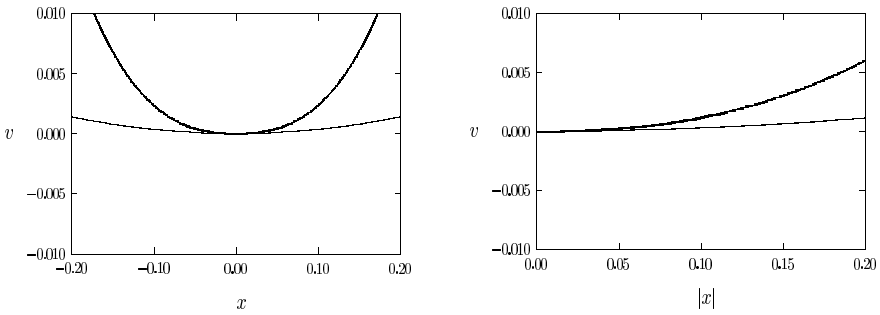
(a) v versus x (b) u versus x

FIG. 13. *Experiment* 2: *for the slab domain and* $\lambda = 1.5 > \lambda_*$ *we plot* v *and* u *versus* x *at times* $t = (0, 0.1, 0.3, 0.5, 0.7, 1.0, 1.04, 1.06, 1.07, 1.07364)$ *from the discrete scheme* (3.28) *with* $N = 200$, *and* $\Delta t = 0.6 \times 10^{-5}$. *For this data, there is touchdown in finite time.*

Experiment 1. We consider the slab domain $|x| \leq 1/2$ with $\lambda = 1.35$. We take $\Delta t = 0.60 \times 10^{-5}$ and $N = 200$, so that $h = 0.49751 \times 10^{-2}$. Since $\lambda < \lambda_* \approx 1.401$ from Table 1, we expect that the time-dependent solution will approach the steady-state solution on the lower branch of the $|u(0)|$ versus λ bifurcation diagram. This is shown in Figure 12(a) and Figure 12(b), where we plot v and $u = -1 + (3\lambda v)^{1/3}$ versus x, respectively.

Experiment 2. Next, we consider the slab domain with $\lambda = 1.5$. From Table 1 we note that $\bar{\lambda}_2 > \lambda > \bar{\lambda}_1 > \lambda_*$. Therefore, Theorem 3.1 guarantees touchdown in a finite time T_* with $T_* < T_1$, where $T_1 = 2.040$ as computed numerically from (3.8). Since $\lambda < \bar{\lambda}_2$, the bound T_2 for the touchdown time, as given in (3.14), is undefined. For the discrete scheme (3.28) we took $\Delta t = 0.60 \times 10^{-5}$ and $N = 200$, so that $h = 0.49751 \times 10^{-2}$. To determine the touchdown time accurately, we took time-steps smaller than this value of Δt when the minimum value of v dropped below some small threshold. In this way, we found that touchdown occurs at $x = 0$ and at $T_* \approx 1.07366$. In Figure 13(a) and Figure 13(b) we plot v and $u = -1 + (3\lambda v)^{1/3}$ versus x, respectively, showing touchdown behavior in finite time.

In Figure 14(a) we compare the numerical approximation for v with the local



(a) slab domain: $t = 1.0736$ (b) unit disk: $t = 0.7228$

FIG. 14. *Plot of discrete approximation for v (heavy solid curve) and the local approximation for v (lighter curve) given in* (3.25). (a) *Slab domain with* $T = 1.07366$. (b) *Unit disk with* $T = 0.722858$.

behavior (3.25) at $t = 1.0736$. In (3.25) we set $N = 1$ and use $T = 1.07366$ for the touchdown time. From this figure we observe that the local asymptotic result (3.25) compares favorably with the numerical result. Note that if we took a coarser mesh with $N = 150$ meshpoints, so that $h = 0.66225 \times 10^{-2}$, and chose $\Delta t = 0.1 \times 10^{-4}$, then the touchdown time would be $T_* \approx 1.07357$.

Experiment 3. Next, we consider the unit disk with $f(x) \equiv 1$ and $\lambda = 1.0$. From Table 1 we note that $\bar{\lambda}_2 > \lambda > \bar{\lambda}_1 > \lambda_*$. Therefore, Theorem 3.1 guarantees touchdown in a finite time T_* with $T_* < T_1$, where $T_1 = 1.140$ as computed numerically from (3.8). Since $\lambda < \bar{\lambda}_2$, T_2 in (3.14) is undefined. A second-order accurate in space and first-order accurate in time discrete approximation to (3.16), with spatial meshsize h, on $0 \le r \le 1$ and $t \ge 0$ is

(3.29a)
$$v_j^{m+1} = v_j^m + \Delta t \left(\frac{\left(v_{j+1}^m - 2v_j^m + v_{j-1}^m\right)}{h^2} + \frac{\left(v_{j+1}^m - v_{j-1}^m\right)}{2hr_j} - 1 - \frac{\left(v_{j+1}^m - v_{j-1}^m\right)^2}{6v_j^m h^2} \right), \quad j = 2, \ldots, N+1,$$

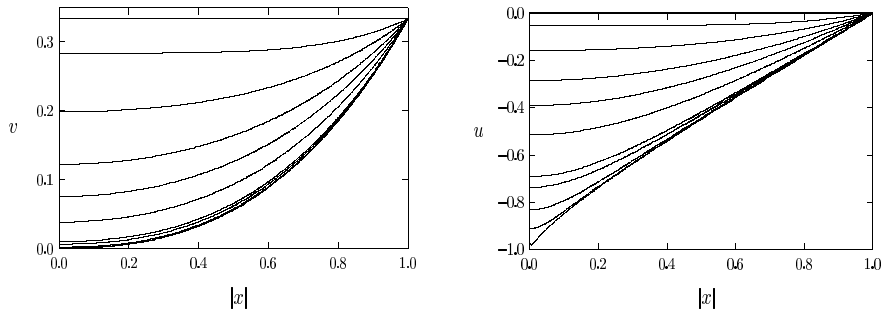
where $r_j = jh$. From [9, p. 50], the discrete approximation for v_1 at the origin $r = 0$ is

(3.29b)
$$v_1^{m+1} = v_1^m + \frac{4\Delta t}{h^2}\left(v_2^m - v_1^m\right).$$

The condition at $r = 1$ is $v_{N+2}^m = (3\lambda)^{-1}$. The results shown below are for $\Delta t = 0.6 \times 10^{-5}$ and $N = 200$, so that $h = 0.49751 \times 10^{-2}$. For these values, the touchdown time is found to be $T_* \approx 0.722858$.

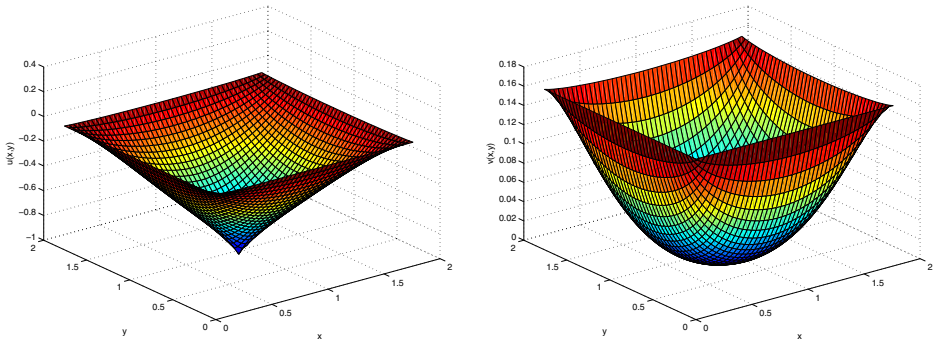
In Figure 15(a) and Figure 15(b) we plot v and $u = -1 + (3\lambda v)^{1/3}$, respectively, versus x, showing touchdown behavior in finite time. In Figure 14(b) we compare the numerical approximation for v with the local behavior (3.25) at $t = 0.7228$. In (3.25) we set $N = 2$ and use $T = 0.722858$ for the touchdown time.

Experiment 4. Finally, we give an example of touchdown behavior in the square domain $[0, \sqrt{\pi}] \times [0, \sqrt{\pi}]$ for the constant permittivity profile $f(x, y) \equiv 1$ with $\lambda = 2.0$. From Table 1 we note that $\bar{\lambda}_2 > \lambda > \bar{\lambda}_1 > \lambda_* \approx 0.857$. Therefore, Theorem 3.1 guarantees touchdown in a finite time T_* with $T_* < T_1$, where $T_1 = 0.2521$ as computed numerically from (3.8). Since $\lambda < \bar{\lambda}_2$, T_2 in (3.14) is undefined.



(a) v versus $|x|$ (b) u versus $|x|$

FIG. 15. *Experiment* 3: *For the unit disk and* $\lambda = 1.0$ *we plot* v *and* u *versus* $|x|$ *at times* $t = (0.05, 0.15, 0.30, 0.45, 0.60, 0.70, 0.71, 0.72, 0.7225, 0.722856)$. *For the discrete scheme* (3.29) *with* $N = 200$, *and* $\Delta t = 0.6 \times 10^{-5}$, *we compute the touchdown time* $T_* \approx 0.722858$.



(a) u versus (x, y) (b) v versus (x, y)

FIG. 16. *Experiment* 4: *For the square* $[0, \sqrt{\pi}] \times [0, \sqrt{\pi}]$ *and* $\lambda = 2.0$ *we show touchdown behavior. For the discrete scheme we used* $N = 150$ *meshpoints in the* x *and* y *directions and a time-step of* $\Delta t = 0.15 \times 10^{-4}$. (a) u *versus* (x, y) *at* $t = 0.1975$. (b) v *versus* (x, y) *at the same time.*

The discretization of (3.16) is similar to that given in (3.28). We use centered differences in x and y to compute discrete approximations to v_{xx} and v_{yy} for Δv. Centered differences are then used in x and y to compute $|\nabla v|^2$. An explicit Euler method is then used for the time integration. For a discretization of 150 meshpoints in each of the x and y directions, and with a time-step of $\Delta t = 0.15 \times 10^{-4}$, which is decreased near touchdown, we compute a touchdown time $T_* \approx 0.19751$. Notice that $T_1 = .2521$ is a reasonably good bound on the touchdown time. The touchdown point $(x_0, y_0) = (.880, .880)$ is at the center of the square. In Figure 16(a) we plot the numerically computed u versus (x, y) for $t = 0.1975$, which is very close to the singularity time. A plot of v versus (x, y) is shown in Figure 16(b).

3.3. The touchdown profile $f(x) \equiv 1$: Center manifold analysis. A different approach to determining the local touchdown profile when $f(x) \equiv 1$ is based on the center manifold analysis of a PDE that results from a similarity group transformation of (1.1). This approach was used in [5] for the case $N = 1$. A closely related method was used in [4] to determine the local blow-up profile for a semilinear heat

equation. We now briefly outline the results that can be derived this way. The first step is to introduce new variables by

$$u = -1 + (T-t)^{1/3} w\,(y,s)\,, \qquad s \equiv -\log(T-t), \qquad y \equiv \frac{x}{\sqrt{T-t}}. \tag{3.30}$$

With this transformation, (1.1) becomes

$$w_s = \frac{1}{\rho}\nabla\cdot(\rho\nabla w) + \frac{w}{3} - \frac{\lambda}{w^2}, \qquad \rho \equiv e^{-|y|^2/4}. \tag{3.31}$$

The touchdown profile as $t \to T^-$ is determined by the large s behavior of (3.31). For $s \gg 1$ and $|y|$ bounded, we have that $w = w_\infty + v$, where $v \ll 1$ and $w_\infty \equiv (3\lambda)^{1/3}$. Keeping the quadratic terms in v, we get

$$v_s = \frac{1}{\rho}\nabla\cdot(\rho\nabla w) + v + \gamma v^2 + O(v^3), \qquad \gamma = -(3\lambda)^{-1/3}. \tag{3.32}$$

As shown in [4] (see also [5]), the nullspace of the linearized operator in (3.32) is three-dimensional when $N = 2$ and is one-dimensional when $N = 1$. By projecting the nonlinear term in (3.32) against the nullspace of the linearized operator, the following far-field behavior of v for $s \to +\infty$ and $|y|$ bounded was obtained (see (1.7), (1.8) of [4]):

$$v \sim \frac{1}{4\gamma s}\left(1 - \frac{y^2}{2}\right), \quad N = 1; \qquad v \sim \frac{1}{2\gamma s}\left(1 - \frac{|y|^2}{2}\right), \quad N = 2. \tag{3.33}$$

The local touchdown profile is then obtained from $w \sim w_\infty + v$, (3.30), and (3.33), which yields

$$u \sim -1 + [3\lambda(T-t)]^{1/3}\left(1 - \frac{N}{4|\log(T-t)|} + \frac{|x|^2}{8(T-t)|\log(T-t)|}\right). \tag{3.34}$$

By making a binomial approximation of (3.26), it is easy to see that (3.26) agrees asymptotically with (3.34). A rigorous derivation of (3.34) for the case $N = 1$, using this type of center manifold analysis, was given in [5]. We also remark that the spatially independent term in (3.34) was proved rigorously in [3].

4. Touchdown behavior: Variable permittivity. In this section we obtain some numerical and formal asymptotic results for touchdown behavior associated with a spatially variable permittivity profile in a slab domain. With the transformation

$$v = \frac{1}{3\lambda}(1+u)^3, \tag{4.1}$$

the problem (1.1) for u in the slab domain, with permittivity profile $f(x)$, transforms exactly to

$$v_t = v_{xx} - \frac{2}{3v}v_x^2 - f(x), \quad |x| < \frac{1}{2}; \qquad v = \frac{1}{3\lambda}, \quad x = \pm\frac{1}{2}; \qquad v = \frac{1}{3\lambda}, \quad t = 0. \tag{4.2}$$

We now use the formal power series method of section 3.2 to locally construct a power series solution to (4.2) near the unknown touchdown point x_0 and the unknown

touchdown time T. We first assume that $f(x)$ is analytic at $x = x_0$ with $f(x_0) > 0$, so that for $x - x_0 \ll 1$ it has the convergent series expansion

$$f(x) = f_0 + f_0^{'}(x - x_0) + \frac{f_0^{''}(x - x_0)^2}{2} + \cdots , \tag{4.3}$$

where $f_0 \equiv f(x_0)$, $f_0^{'} \equiv f^{'}(x_0)$, and $f_0^{''} \equiv f^{''}(x_0)$. Near $x = x_0$, we look for a touchdown profile for (4.2) in the form

$$v(x,t) = v_0(t) + \frac{(x - x_0)^2}{2!}v_2(t) + \frac{(x - x_0)^3}{3!}v_3(t) + \frac{(x - x_0)^4}{4!}v_4(t) + \cdots . \tag{4.4}$$

In order for v to be a touchdown profile, it is clear that we must require that

$$\lim_{t \to T^-} v_0 = 0, \qquad v_0 > 0 \quad \text{for } t < T, \qquad v_2 > 0 \quad \text{for } t - T \ll 1. \tag{4.5}$$

Substituting (4.4) and (4.3) into (4.2), we equate powers of $x - x_0$ to obtain

$$v_0^{'} = -f_0 + v_2, \qquad v_2^{'} = -\frac{4v_2^2}{3v_0} + v_4 - f_0^{''}, \qquad v_3 = f_0^{'}. \tag{4.6}$$

As in section 3.2, we assume that $v_2 \ll 1$ and $v_4 \ll 1$ as $t \to T^-$. This yields that $v_0 \sim f_0(T - t)$ and

$$v_2^{'} \sim -\frac{4v_2^2}{3f_0(T - t)} - f_0^{''}. \tag{4.7}$$

For $t \to T^-$, we obtain from a simple dominant balance argument that

$$v_2 \sim -\frac{3f_0}{4\left[\log(T - t)\right]} + \cdots \quad \text{as } t \to T^-. \tag{4.8}$$

By substituting (4.8) into (4.6) for v_0 and integrating the resulting expression, we obtain

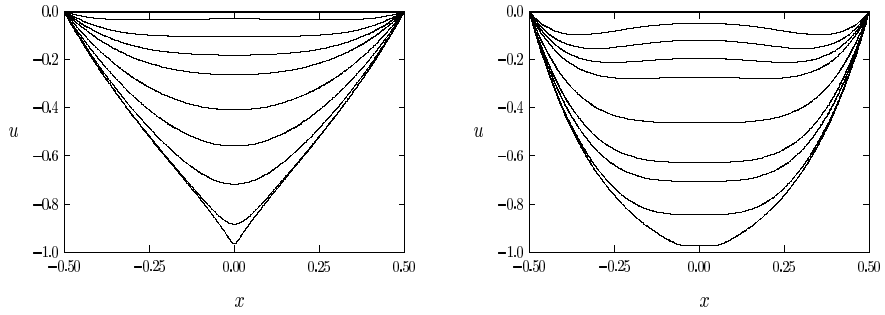
$$v_0 \sim f_0\,(T - t) + \frac{-3f_0(T - t)}{4|\log(T - t)|} + \cdots \quad \text{for } t \to T^-. \tag{4.9}$$

Next, we substitute (4.8), (4.9), and (4.6) for v_3 into (4.4) to obtain the local touchdown behavior

$$v \sim f_0\,(T - t)\left[1 - \frac{3}{4|\log(T - t)|} + \frac{3(x - x_0)^2}{8(T - t)|\log(T - t)|} + \frac{f_0^{'}(x - x_0)^3}{6f_0(T - t)} + \cdots\right], \tag{4.10}$$

for $(x - x_0) \ll 1$ and $t - T \ll 1$. Finally, using the nonlinear mapping (4.1) relating u and v, we conclude that

$$u \sim -1 + [3f_0\lambda(T - t)]^{1/3}\left(1 - \frac{3}{4|\log(T - t)|} + \frac{3(x - x_0)^2}{8(T - t)|\log(T - t)|} + \frac{f_0^{'}(x - x_0)^3}{6f_0(T - t)} + \cdots\right)^{1/3}. \tag{4.11}$$



(a) $\alpha = 2.0$, $\lambda = 5.0$ (b) $\alpha = 10.0$, $\lambda = 22.0$

FIG. 17. *Exponential permittivity profile.* (a) *Plot of u versus x at different times for $\alpha = 2.0$ and $\lambda = 5.0$. The touchdown time is $T_* \approx 0.1332$ and $\lambda_* \approx 2.14$.* (b) *Plot of u versus x at different times for $\alpha = 10.0$ and $\lambda = 22.0$. The touchdown time is $T_* \approx 0.1497$ and $\lambda_* \approx 10.4$. For both cases, the touchdown point is $x_0 = 0$.*

Here $f_0 \equiv f(x_0)$ and $f_0^{'} \equiv f^{'}(x_0)$.

Since $f(x) > 0$ for the exponential profile $f(x) = e^{\alpha(x^2-1/4)}$ of (2.16a), then (4.11) holds for some touchdown point x_0 and touchdown time T. If the touchdown point is at $x_0 = 0$, then (4.11) holds for $f_0 = e^{-\alpha/4}$, and $f_0^{'} = 0$. For two sets of α and λ, in Figure 17 we plot the numerically computed u versus x at different times showing touchdown behavior for the exponential permittivity profile. The bounds T_1 and T_2 on the touchdown time, given in (3.8) and (3.14), together with the numerically computed touchdown time T_* and saddle-node value λ_* are as follows:

(4.12a)

$$\alpha = 2.0, \quad \lambda = 5.0; \qquad \lambda_* \approx 2.14, \quad T_1 = 0.1697, \quad T_2 = 0.4030, \quad T_* \approx 0.1332,$$

(4.12b)

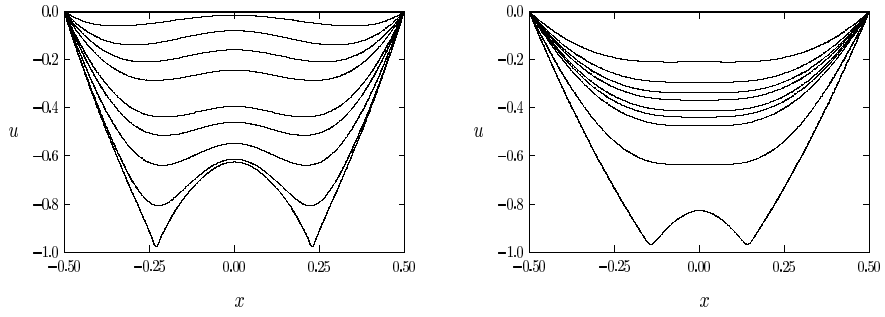
$$\alpha = 10.0, \quad \lambda = 22.0; \qquad \lambda_* \approx 10.40, \quad T_1 = 0.5321, \quad T_2 = 0.3281, \quad T_* \approx 0.1497.$$

In obtaining (4.12) we discretized (4.2) in a similar manner as in (3.28). The discrete approximation to u was then obtained from (4.1). The computations were done with a time-step of $\Delta t = 0.6\times 10^{-5}$ and with $N = 200$ meshpoints, so that $h = 0.4975\times 10^{-2}$. From Figure 17 we observe that touchdown occurs at $x_0 = 0$. For $\alpha = 10$, the touchdown profile is much flatter than that for $\alpha = 2$. This is because $f(0) = e^{-\alpha/4}$ is a decreasing function of α.

We remark that the touchdown profile (4.11) also holds for the power-law profile $f(x) = |2x|^\alpha$ of (2.16a) whenever the touchdown point x_0 is not at the origin, i.e., $x_0 \neq 0$. If this occurs, then (4.11) holds with

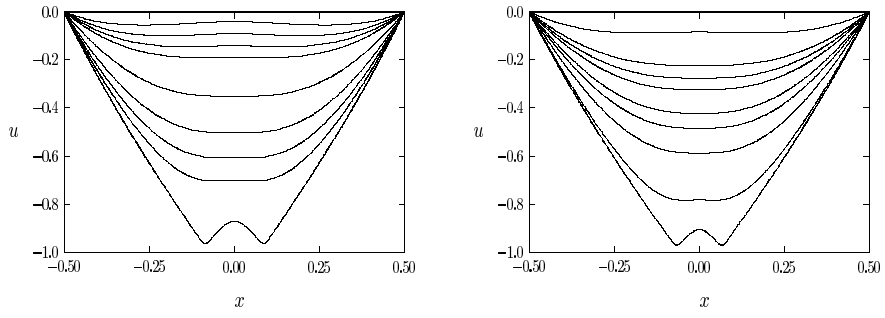
$$f_0 = |2x_0|^\alpha, \qquad f_0^{'} = 2\alpha|2x_0|^{\alpha-1}. \tag{4.13}$$

In Figure 18 we plot the numerically computed u versus x at different times, and for different sets of α and λ, showing touchdown behavior for the power-law profile. In this figure, the touchdown time T_* and the saddle-node value λ_* are shown for each parameter set. From these numerical results we observe that touchdown seems to occur at two points, symmetrically located about the origin. For each of the



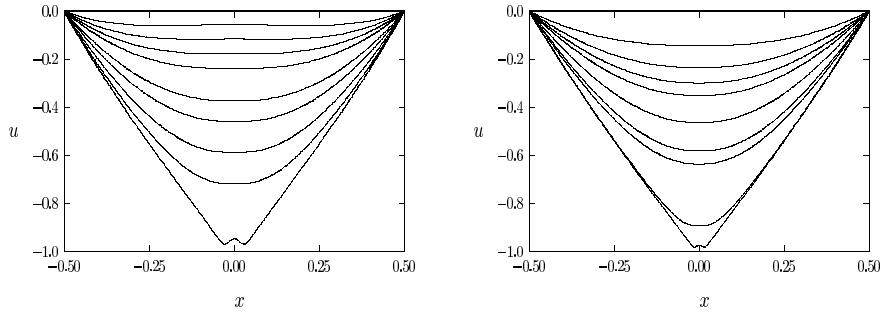
(a) $\alpha = 1.0$, $\lambda = 10.0$

(b) $\alpha = 1.0$, $\lambda = 4.5$



(c) $\alpha = 0.4$, $\lambda = 4.0$

(d) $\alpha = 0.4$, $\lambda = 3.0$



(e) $\alpha = 0.16$, $\lambda = 3.5$

(f) $\alpha = 0.16$, $\lambda = 2.0$

FIG. 18. *Power-law permittivity profile: Plots of u versus x at different times for the values of* α *and* λ *shown in the figure captions. The values for saddle-node point* λ_**, the touchdown time* T_**, and the touchdown points* x_0 *are as follows.* (a) $\lambda_* \approx 4.2$, $T_* \approx 0.1257$, $x_0 = \pm 0.226$. (b) $\lambda_* \approx 4.2$, $T_* \approx 1.887$, $x_0 = \pm 0.147$. (c) $\lambda_* \approx 2.41$, $T_* \approx 0.2366$, $x_0 = \pm 0.087$. (d) $\lambda_* \approx 2.41$, $T_* \approx 0.4857$, $x_0 = \pm 0.067$. (e) $\lambda_* \approx 1.77$, $T_* \approx 0.174$, $x_0 = \pm 0.027$. (f) $\lambda_* \approx 1.77$, $T_* \approx 0.746$, $x_0 = \pm 0.012$.

computations, we have taken $N = 200$ meshpoints, so that $h = 0.4975 \times 10^{-2}$, and a time step $\Delta t = 0.6 \times 10^{-5}$.

Next, we perform a more delicate computational experiment to determine whether touchdown can occur at $x = 0$ for the power-law profile. We take $\alpha = 0.01$ and $\lambda = 2.0$. Since $\alpha \ll 1$, this example represents a small perturbation of the constant

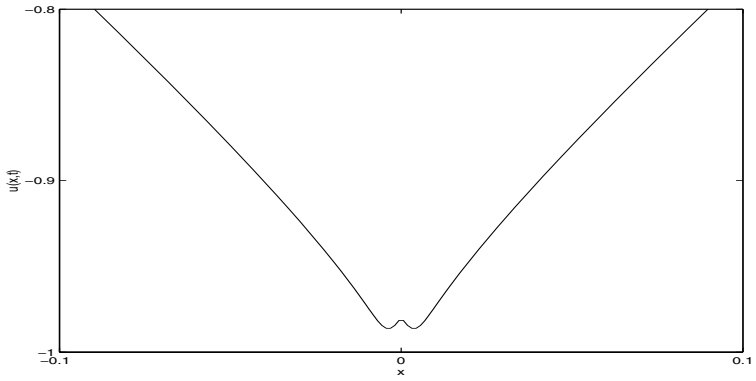


FIG. 19. *Plot of u versus x at $t = 0.34213$ near the touchdown region for the power-law profile with $\alpha = 0.01$ and $\lambda = 2.0$. Touchdown does not occur at $x = 0$, but rather at two points on either side of $x = 0$. The touchdown time is $T_* \approx 0.3422$.*

permittivity profile $f(x) \equiv 1$. Using $N = 800$ meshpoints, in Figure 19 we plot u versus x at $t = 0.34213$ in a neighborhood of the origin. The touchdown time is found to be $T_* \approx 0.3422$. From this figure, we observe that touchdown does not occur at $x = 0$. These computational results suggest the possibility that touchdown cannot occur at a point x_0, where $f(x_0) = 0$. We first investigate this possibility by using a formal power series analysis. Then, at the end of this section we prove a result in this direction. A consequence of this analysis is that touchdown at $x_0 = 0$ is impossible for the power-law profile $f(x) = |2x|^\alpha$.

We first assume that $f(x)$ is analytic at $x = 0$, with $f(0) = 0$ and $f'(0) = 0$, so that $f(x) = f_0 x^2 + O(x^3)$ as $x \to 0$ with $f_0 > 0$. We then look for a power series solution to (4.2) as in (4.4). In place of (4.6) for v_3, we get $v_3 = 0$, and

$$v_0' = v_2, \qquad v_2' = -\frac{4v_2^2}{3v_0} + v_4 - 2f_0. \tag{4.14}$$

Assuming that $v_4 \ll 1$ as before, we can combine the equations in (4.14) to get

$$v_0'' = -\frac{4(v_0')^2}{3v_0} - 2f_0. \tag{4.15}$$

By solving (4.15) with $v_0(T) = 0$, we obtain the exact solution

$$v_0 = -\frac{3f_0}{11}(T-t)^2, \qquad v_2 = \frac{6f_0}{11}(T-t). \tag{4.16}$$

Since the criteria (4.5) are not satisfied, the form (4.16) does not represent a touchdown profile centered at $x = 0$.

A similar calculation can be done for the case where $f(x)$ is analytic at $x = 0$, with $f(0) = 0$ and $f'(0) = f_0 > 0$. From a power series expansion solution centered at $x = 0$, and assuming that $v_4 \ll 1$, we get $v_3 = f_0$ and

$$v_0'' = -\frac{4(v_0')^2}{3v_0}, \qquad v_2 = v_0'. \tag{4.17}$$

In terms of some constant A, the explicit solution to (4.17) with $v_0(T) = 0$ is

$$v_0 = A\,(T-t)^{3/7}\,, \qquad v_2 = -\frac{3A}{7}\,(T-t)^{-4/7}\,. \tag{4.18}$$

Since v_0 and v_2 have opposite signs as $t \to T^-$, the criteria (4.5) do not hold, and we do not have touchdown at $x = 0$. These formal calculations suggest the general result that touchdown cannot occur at a point $x = x_0$, where $f(x_0) = 0$. Without loss of generality, we assume that $x_0 = 0$. Our final result is as follows.

THEOREM 4.1. *Let $u(x,t)$ be a solution of*

$$u_t = u_{xx} - \frac{\lambda f(x)}{u^2}, \qquad |x| \le \frac{1}{2}, \quad 0 < t < T; \qquad u\left(\pm\frac{1}{2}, t\right) = 1, \qquad u(x,0) = 1. \tag{4.19}$$

Here $f(x)$ satisfies (2.7)*, and u touches down at the finite time T. If $f(0) = 0$, then $x_0 = 0$ cannot be a touchdown point of $u(x,t)$ at finite time T.*

Proof. Set $v = u_t$. Then we calculate

$$v_t = v_{xx} + \frac{2\lambda f(x)}{u^3} v, \quad |x| \le \frac{1}{2}, \quad 0 < t < T; \qquad v\left(\pm\frac{1}{2}, t\right) = 0, \quad v(x,0) \le 0. \tag{4.20}$$

Here $\frac{2\lambda f(x)}{u^3}$ is a locally bounded function. By the strong maximum principle, we conclude that

$$u_t = v < 0, \qquad |x| < \frac{1}{2}, \quad 0 < t < T. \tag{4.21}$$

Therefore, since $f(0) = 0$, we have as $t \to T^-$ that $u_{xx} = u_t < 0$ at $x = 0$. From this result, and from the smoothness of $u(x,t)$, we deduce that when $t \to T^-$, there exists an $\bar{x} \neq 0$ such that $u(0,t) > u(\bar{x},t)$. This shows that $x_0 = 0$ cannot be a touchdown point of $u(x,t)$ at finite time T. ∎

5. Conclusion. We have analyzed some properties of the pull-in voltage instability for (1.1) in terms of a spatially variable dielectric permittivity profile for the thin elastic membrane. Bounds on the pull-in voltage were given in section 2, and sufficient conditions for finite-time touchdown were obtained in section 3, together with bounds on the touchdown time. From these bounds, and from numerical computations, it was shown that by appropriately tailoring the dielectric permittivity of the thin membrane the pull-in voltage and the pull-in distance can both be increased. For the special case of a power-law permittivity profile in a slab domain, this conclusion was first obtained in [14]. For voltages that exceed the pull-in voltage threshold, the local touchdown profile was calculated asymptotically in sections 3 and 4 for spatially uniform and spatially nonuniform permittivity profiles, respectively.

An interesting open problem is to formulate an optimization problem for the pull-in distance associated with the steady-state problem (1.1), whereby an optimum permittivity profile f can be computed numerically for a given set of design constraints on both the stable operating range of the applied voltage and maximum value of V that is available by the power supply.

Another way of tailoring the pull-in voltage, without introducing a spatially nonuniform permittivity profile, is to rigidly attach the thin membrane near the

region where the deflection would otherwise be largest. Mathematically this corresponds to considering (1.1) with $f(x,y) \equiv 1$, in a domain Ω punctured by a small patch Ω_ε of area $O(\varepsilon^2) \ll 1$, where $u = 0$ for $x \in \Omega_\varepsilon$. An asymptotic theory for the location of saddle-node bifurcation values for general classes of semilinear problems in such singularly perturbed domains was developed in [18]. For a MEMS device, symmetry-breaking properties of radially symmetric solutions for an annular domain were computed numerically in [15]. For this type of modification of (1.1), it would be interesting to obtain an analytical theory for the pull-in voltage instability.

Finally, it would be interesting to analyze pull-in voltage and touchdown behavior for extensions of the basic model (1.1), whereby the upper surface is modeled by an elastic plate of nonzero rigidity and inertial effects are considered. The resulting model for the deflection of a thin plate that has a spatially uniform permittivity profile involves the Biharmonic operator Δ^2 and takes the following form for some $\beta > 0$ and $\delta > 0$ (see (7.50) of [13]):

$$\beta \frac{\partial^2 u}{\partial t^2} + \frac{\partial u}{\partial t} - \Delta u + \delta \Delta^2 u = -\frac{\lambda}{(1+u)^2}, \quad x \in \Omega; \qquad u = 0, \quad (x,y) \in \partial\Omega; \qquad u(x,y,0) = 0. \tag{5.1}$$

Appendix. Derivation of the membrane deflection equation. Following the analysis in [14] and [6], we now outline the derivation of the membrane deflection equation (1.1). Referring to Figure 1, the electrostatic potential is assumed to satisfy Laplace's equation in the gap between the fixed plate and the lower surface of the membrane. Inside the thin membrane, the dielectric permittivity $\varepsilon_2 = \varepsilon_2(x,y)$ can exhibit a spatial variation. On the upper surface of the membrane, a fixed voltage V is imposed. Therefore, in dimensionless variables, the problem for the electrostatic potential is

$$\frac{\partial^2 \psi}{\partial z^2} + \delta^2 \left(\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} \right) = 0, \quad (x,y) \in \Omega, \quad 0 \le z \le \hat{u} - l, \tag{A.1a}$$

$$\varepsilon_2 \frac{\partial^2 \psi}{\partial z^2} + \delta^2 \left(\frac{\partial}{\partial x}\left(\varepsilon_2 \frac{\partial \psi}{\partial x} \right) + \frac{\partial}{\partial y}\left(\varepsilon_2 \frac{\partial \psi}{\partial y} \right) \right) = 0, \quad (x,y) \in \Omega, \quad \hat{u} - l \le z \le \hat{u} + l, \tag{A.1b}$$

$$\psi = 0, \quad z = 0 \text{ (ground plate)}; \qquad \psi = 1, \quad z = \hat{u} + l \text{ (upper membrane surface)}, \tag{A.1c}$$

together with the continuity of the potential and the displacement fields across $z = \hat{u} - l$. Here ψ is the dimensionless potential scaled with respect to the applied voltage V, x and y are scaled with respect to the length L of the undeformed plate Ω, z is a vertical coordinate scaled with respect to the undeformed gap-size d, $2l$ is the thickness of the membrane, and $\delta \equiv d/L \ll 1$ is the device aspect ratio. The deflection of the membrane is denoted by $\hat{u}$, with $\hat{u} = 1$ on $\partial\Omega$ denoting the undeflected state. Note that $\hat{u} = 0$ corresponds to the touching of the membrane and the lower plate and that $\hat{u}$ is scaled in the same manner as z.

In the small aspect ratio limit $\delta \ll 1$, the asymptotic solution for ψ that is continuous across $z = \hat{u} - l$ is

$$\psi = \begin{cases} \psi_L \frac{z}{\hat{u}-l}, & 0 \le z \le \hat{u} - l, \\ 1 + \frac{(1-\psi_L)}{2l}\left(z - (\hat{u} + l) \right), & \hat{u} - l \le z \le \hat{u} + l. \end{cases} \tag{A.2}$$

To ensure that the displacement field is continuous across $z = \hat{u} - l$ to leading order in δ, we must impose that $\varepsilon_0 \psi_z|_- = \psi_2 \psi_z|_+$, where the plus or minus signs indicate that ψ_z is to be evaluated on the upper or lower side, respectively, of the bottom surface $z = \hat{u} - l$ of the membrane. This condition determines ψ_L in (A.2) as

$$\psi_L = \left[1 + \frac{2l}{\hat{u} - l}\left(\frac{\varepsilon_0}{\varepsilon_2}\right)\right]^{-1}. \tag{A.3}$$

From (A.2) and (A.3), the electric field in the z-direction inside the membrane is independent of z, and is given by

$$\psi_z = \frac{\varepsilon_0}{\varepsilon_2(\hat{u} - l)}\left[1 + \frac{2l}{\hat{u} - l}\frac{\varepsilon_0}{\varepsilon_2}\right]^{-1} \sim \frac{\varepsilon_0}{\varepsilon_2 \hat{u}} \quad \text{for } l \ll 1. \tag{A.4}$$

The coupling of the electrostatic field to the deflection of the membrane was modeled in [6] by a dimensionless damped wave equation of the form

$$\gamma^2 \frac{\partial^2 \hat{u}}{\partial t^2} + \frac{\partial \hat{u}}{\partial t} - \Delta \hat{u} = -\lambda \left(\frac{\varepsilon_2}{\varepsilon_0}\right)\left[\delta^2 |\nabla_\perp \psi|^2 + \left(\frac{\partial \psi}{\partial z}\right)^2\right], \quad (x, y) \in \Omega, \quad \hat{u} - l \le z \le \hat{u} + l. \tag{A.5}$$

Here λ is defined in (1.2), the time t is scaled with respect to the strength of the damping, and $\nabla_\perp$ denotes the gradient in the x- and y-directions only. By substituting (A.4) into (A.5), and letting $\delta \ll 1$, we obtain

$$\gamma^2 \frac{\partial^2 \hat{u}}{\partial t^2} + \frac{\partial \hat{u}}{\partial t} - \Delta \hat{u} \sim -\lambda \frac{\varepsilon_0}{\varepsilon_2 \hat{u}^2}, \quad (x, y) \in \Omega; \qquad \hat{u} = 1, \quad (x, y) \in \partial\Omega. \tag{A.6}$$

We then define $u \equiv \hat{u} - 1$, so that $u = 0$ is the undeflected state. Finally, assuming that the damping force dominates the inertial force so that $\gamma \ll 1$, as was done in [6], (A.6) reduces to the membrane deflection equation (1.1) of section 1.

REFERENCES

[1] U. Ascher, R. Christiansen, and R. Russell, *Collocation software for boundary value ODE's*, Math. Comp., 33 (1979), pp. 659–679.

[2] R. E. Bank, *PLTMG: A Software Package for Solving Elliptic Partial Differential Equations, User's Guide* 8.0, Software Environ. Tools, SIAM, Philadelphia, PA, 1998.

[3] M. Fila and J. Hulshof, *A note on the quenching rate*, Proc. Amer. Math. Soc., 112 (1991), pp. 473–477.

[4] S. Filippas and R. V. Kohn, *Refined asymptotics for the blow up of* $u_t - \Delta u = u^p$, Comm. Pure Appl. Math., 45 (1992), pp. 821–869.

[5] S. Filippas and J. S. Guo, *Quenching profiles for one-dimensional semilinear heat equations*, Quart. Appl. Math., 51 (1993), pp. 713–729.

[6] G. Flores, G. A. Mercado, and J. A. Pelesko, *Dynamics and touchdown in electrostatic MEMS*, in Proceedings of the 2003 International Conference on MEMS, NANO, and Smart Systems, Banff, AB, Canada 2003, IEEE Computer Society, Los Alamitos, CA, 2003, pp. 182–187.

[7] J. S. Guo, *On the quenching behavior of the solution of a semilinear parabolic equation*, J. Math. Anal. Appl., 151 (1990), pp. 58–79.

[8] J. B. Keller and J. Lowengrub, *Asymptotic and numerical results for blowing-up solutions to semilinear heat equations*, in Proceedings of the meeting on Singularities in Fluids, Plasmas, and Optics (Heraklion 1992), NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 404, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993, pp. 11–38.

[9] K. W. Morton and D. F. Mayers, *Numerical Solution of Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1994.
[10] H. C. Nathanson, W. E. Newell, R. A. Wickstrom, and J. R. Davis, *The resonant gate transistor*, IEEE Trans. Elect. Devices, 14 (1967), pp. 117–133.
[11] J. A. Pelesko, *Multiple Solutions in Electrostatic MEMS*, in Proceedings of the 4th International Conference on Modeling and Simulation of Microsystems, Hilton Head, SC, 2001, NSTI, Cambridge, MA, 2001, pp. 290–293.
[12] J. A. Pelesko and A. A. Triolo, *Nonlocal Problems in MEMS Device Control*, J. Eng. Math., 41 (2001), pp. 345–366.
[13] J. A. Pelesko and D. H. Bernstein, *Modeling MEMS and NEMS*, Chapman Hall, London, and CRC Press, Boca Raton, FL, 2002.
[14] J. A. Pelesko, *Mathematical modeling of electrostatic MEMS with tailored dielectric properties*, SIAM J. Appl. Math., 62 (2002), pp. 888–908.
[15] J. A. Pelesko, D. Bernstein, and J. McCuan, *Symmetry and symmetry breaking in electrostatic MEMS*, in Proceedings of the 6th International Conference on Modeling and Simulation of Microsystems, San Francisco, CA, 2003, NSTI, Cambridge, MA, 2003, pp. 304–307.
[16] I. Stackgold, *Green's Functions and Boundary Value Problems*, Wiley, New York, 1998.
[17] G. I. Taylor, *The coalescence of closely spaced drops when they are at different electric potentials*, Proc. Roy. Soc. A, 306 (1968), pp. 423–434.
[18] M. J. Ward, W. D. Henshaw, and J. B. Keller, *Summing logarithmic expansions for singularly perturbed eigenvalue problems*, SIAM J. Appl. Math., 53 (1993), pp. 799–828.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 339–360

ON RECOVERING THE SHAPE OF A DOMAIN FROM THE TRACE OF THE HEAT KERNEL*

Z. SCHUSS† AND A. SPIVAK‡

Abstract. The problem of recovering geometric properties of a domain from the trace of the heat kernel for an initial-boundary value problem arises in NMR microscopy and other applications. It is similar to the problem of "hearing the shape of a drum," for which a Poisson-type summation formula relates geometric properties of the domain to the eigenvalues of the Dirichlet or Neumann problems for the Laplace equation. It is well known that the area, circumference, and the number of holes in a planar domain can be recovered from the short-time asymptotics of the solution of the initial-boundary value problem for the heat equation. It is also known that the length spectrum of closed billiard ball trajectories in the domain is contained in the spectral density of the Laplace operator with the given boundary conditions in the domain, from which the short-time hyperasymptotics of the trace of the heat kernel can be obtained by the Laplace transform. However, the problem of recovering these lengths from measured values of the trace of the heat kernel (the "resurgence" problem) is unresolved. In this paper we develop a simple algorithm for extracting the lengths from the short-time hyperasymptotic expansion of the trace. We give an alternative construction of the short-time expansion of the trace by constructing a ray approximation to the heat kernel for a planar domain with Dirichlet or Neumann boundary conditions. We evaluate the trace by introducing the rays as global coordinates.

Key words. heat kernel, trace, short-time asymptotics, eigenvalues

AMS subject classifications. 35K20, 35J25, 35C20, 35P20

DOI. 10.1137/S0036139903424928

1. Introduction.

The problem of recovering geometric properties of a domain from NMR measurements arises in oil explorations and in noninvasive microscopy of cell structure [1], as well as in other applications. In these measurements the trace of the heat kernel for the initial value problem with reflecting (Neumann) boundary conditions is measured directly. The problem is analogous to "hearing the shape of a drum," where the solution of the wave equation in the domain is measured directly (it is "heard"). This problem consists in recovering geometrical properties of a domain from the eigenvalues of the Dirichlet or Neumann problems for the Laplace equation in a bounded domain.

Much attention has been devoted in the literature to the recovery problem (see [2], [3], [4], [5], [6], [7], [8], [9] for some history and early results; for more recent work see [10], [11] and the references therein). The mathematical statement of the problem is as follows. Green's function for the heat equation in a smooth planar domain Ω, with homogeneous Dirichlet boundary conditions, satisfies

$$\frac{\partial G(\boldsymbol{y},\boldsymbol{x},t)}{\partial t} = D\Delta_{\boldsymbol{y}} G(\boldsymbol{y},\boldsymbol{x},t) \quad \text{for} \quad \boldsymbol{y},\boldsymbol{x}\in\Omega,\ t>0, \tag{1.1}$$

*Received by the editors March 26, 2003; accepted for publication (in revised form) May 6, 2005; published electronically November 15, 2005.
http://www.siam.org/journals/siap/66-1/42492.html

†Department of Mathematics, Tel-Aviv University, Ramat-Aviv, Tel-Aviv 69978, Israel (schuss@post.tau.ac.il). This author was partially supported by a research grant from the Foundation for Basic Research administered by the Israel Academy of Science and by a research grant from the US-Israel Binational Science Foundation.

‡Department of Sciences, Academic Institute of Technology, P.O. Box 305, Holon 58102, Israel (spivak@hait.ac.il).

$$G(\boldsymbol{y},\boldsymbol{x},0)=\delta(\boldsymbol{y}-\boldsymbol{x}), \tag{1.2}$$

$$G(\boldsymbol{y},\boldsymbol{x},t)=0 \quad \text{for} \quad \boldsymbol{y}\in\partial\Omega,\ \boldsymbol{x}\in\Omega,\ t>0. \tag{1.3}$$

Since D can be eliminated from (1.1) by scaling it into t, we assume henceforward that $D=1$. The function $G(\boldsymbol{x},\boldsymbol{x},t)\,d\boldsymbol{x}$ is the probability of return to $\boldsymbol{x}\,d\boldsymbol{x}$ at time t of a free Brownian particle that starts at the point $\boldsymbol{x}$ at time $t=0$ and diffuses in Ω with diffusion coefficient 1, with absorption at the boundary $\partial\Omega$. If it is reflected at $\partial\Omega$, rather than absorbed, the Dirichlet boundary condition (1.3) is replaced with the Neumann condition [12]

$$\frac{\partial G(\boldsymbol{y},\boldsymbol{x},t)}{\partial\boldsymbol{\nu}(\boldsymbol{y})}=0 \quad \text{for} \quad \boldsymbol{y}\in\partial\Omega,\ \boldsymbol{x}\in\Omega,\ t>0, \tag{1.4}$$

where $\boldsymbol{\nu}(\boldsymbol{y})$ is the unit outer normal at the boundary point $\boldsymbol{y}$. The trace of the heat kernel is defined as

$$P(t)=\int_\Omega G(\boldsymbol{x},\boldsymbol{x},t)\,d\boldsymbol{x} \tag{1.5}$$

and can be represented by the Dirichlet series

$$P(t)=\sum_{n=1}^{\infty}e^{-\lambda_n t}, \tag{1.6}$$

where λ_n are the eigenvalues of Laplace equation with the Dirichlet or Neumann boundary conditions (1.3) or (1.4), respectively.

It has been shown by Kac [2] that for a domain Ω with smooth boundary $\partial\Omega$, the leading terms in the expansion of $P(t)$ in powers of $\sqrt{t}$ are

$$P_{\text{Kac}}(t)\sim\frac{|\Omega|}{4\pi t}-\frac{|\partial\Omega|}{8\sqrt{\pi t}}+\frac{1}{6}(1-r)+O\left(\sqrt{t}\right) \quad \text{for } t\to 0,$$

where $|\Omega|$ denotes the area of Ω, $|\partial\Omega|$ denotes the arclength of $\partial\Omega$, and r is the number of holes in Ω. The full short-time asymptotic power series expansion of $P(t)$ in the form

$$P(t)\sim\sum_{n=0}^{\infty}a_n t^{n/2-1}$$

can be deduced from the large s expansion of the Laplace transform

$$g(s)=\int_0^\infty\exp\{-s^2t\}\left(P(t)-\frac{a_0}{t}\right)dt, \quad \left(a_0=\frac{|\Omega|}{4\pi}\right)$$

in inverse powers of s. Such an expansion was given by Stewartson and Waechter [3] in the form

$$\hat{g}(s)\sim\sum_{n=1}^{\infty}\frac{c_n}{s^n},$$

where

$$c_n=a_n\Gamma\left(\frac{n}{2}\right). \tag{1.7}$$

The constants c_n are computable functionals of the curvature of the boundary. The full expansion is denoted

$$P_{\mathrm{SW}}(t) \sim \frac{|\Omega|}{4\pi t} - \frac{|\partial\Omega|}{8\sqrt{\pi t}} + \frac{1}{6}(1-r) + \sum_{n=3}^{\infty} a_n t^{n/2-1} \quad \text{for } t \to 0. \tag{1.8}$$

If the boundary is not smooth but has cusps and corners, the expansion contains a term of the order $t^{-\nu}$, where ν is a number between 0 and 1/2.

The Stewartson–Waechter expansion was used in [10] to deduce further geometric properties of Ω by extending $g(s)$ into the complex plane. Examples were given in [10] of the resurgence of the length spectrum of closed billiard ball trajectories in the domain. It was conjectured in [10, eq. (4)] that

$$a_n = \frac{\alpha\Gamma(n-\beta+1)}{\Gamma\left(\frac{n}{2}\right) l^{n-2}}, \tag{1.9}$$

where α and β are constants of order unity and l is the shortest accessible geodesic (as defined in [10]).

The full length spectrum of closed geodesics on a compact Riemannian manifold without boundary Ω appeared in the short-time asymptotic expansion given in [5], [6],

$$P(t) \sim \frac{1}{\sqrt{\pi t}} \sum_{n=0}^{\infty} P_n(\sqrt{t})\, e^{-\delta_n^2/t} \quad \text{for } t \to 0, \tag{1.10}$$

where δ_n are the lengths of closed geodesics on Ω and $P_n(x)$ are power series in x.

Using a different approach, based on the expansion of the spectral density [7], [8], [10]

$$d(s) = \sum_{n=1}^{\infty} \delta(s-\lambda_n) = \bar{d}(s) + d_{\mathrm{OSC}}(s), \tag{1.11}$$

where the nonoscillatory and oscillatory parts are, respectively,

$$\bar{d}(s) \sim \frac{|\Omega|}{\pi}, \quad d_{\mathrm{OSC}}(s) \sim \Re\mathrm{e} \sum_j A_j(s) e^{-il_j\sqrt{s}} \quad \text{for} \quad s \to \infty, \tag{1.12}$$

the second sum extends over the periodic orbits of billiard balls in Ω, and l_j are their lengths. The coefficients $A_j(s)$ depend on the stability of the orbits (see [10, eqs. (11) and (12)]).

Using the identity

$$\int_0^{\infty} \exp\left\{-st - \frac{\delta_n^2}{t}\right\} dt = \frac{2\delta_n K_1(2\delta_l\sqrt{s})}{\sqrt{s}} \sim \frac{\sqrt{\pi\delta_n}}{s} e^{-2\delta_l\sqrt{s}} \quad \text{for} \quad s \to \infty \tag{1.13}$$

and extending formally the asymptotic relation (1.12) to the complex plane, we identify

$$\delta_n = \frac{l_n}{2}. \tag{1.14}$$

In this paper, we adopt a direct approach to the hyperasymptotic short-time expansion of the trace, rather than expanding its Laplace transform. The results can be generalized to higher dimensions in a straightforward manner. We construct the expansion in the form

$$P(t) \sim P_{\rm SW}(t) + \frac{1}{\sqrt{\pi t}} \sum_{n=1}^{\infty} P_n(\sqrt{t})\, e^{-\delta_n^2/t} \quad \text{for } t \to 0, \tag{1.15}$$

where δ_n, ordered by magnitude, are determined directly from the expansion to be related to l_n by (1.14), and $P_n(x)$ are power series in x. Transcendentally small terms may be, in fact, quite large and make a finite contribution to the expansion (1.15) [13]. Indeed, given $P(t)$, e.g., from NMR measurements, we describe a simple numerical algorithm for recovering δ_n from $P(t)$.

To construct the expansion, we use the short-time *ray asymptotic approximation* to the heat kernel [14], [15], [16], [17] to evaluate its trace. Specifically, we use the rays as global coordinates to expand the double integral (1.5) asymptotically beyond all orders for short times. We show that the transcendentally small terms are due to rays reflected in the boundary, much like in the geometric theory of diffraction [18], [19], [20]. In particular, the smallest exponent δ_1 is the width of the narrowest bottleneck in the domain. For the particular case of a circular domain, we find that all diffractive closed trajectories contribute to the transcendentally small terms (see also [10, sect. 3]).

In section 2, we explain the ray approximation and the evaluation of the trace in a one-dimensional example. In section 3, we generalize the ray approximation to higher dimensions, and in section 4 we use it to evaluate the trace for planar domains. In section 5, we work out an algorithm for the numerical evaluation of the exponents δ_n from the given trace $P(t)$. Finally, in section 6 we carry out explicit computations for special domains.

2. The one-dimensional case. The solution of the heat equation in an interval can be constructed by the method of images. Specifically, the Green function of the problem satisfies

$$\frac{\partial G(y,x,t)}{\partial t} = \frac{\partial^2 G(y,x,t)}{\partial y^2} \quad \text{for } 0 < x, y < a,\ t > 0, \tag{2.1}$$

$$G(y,x,0) = \delta(y-x) \qquad \text{for } 0 < x, y < a, \tag{2.2}$$

$$\left(\frac{\partial}{\partial y}\right)^k G(0,x,t) = \left(\frac{\partial}{\partial y}\right)^k G(a,x,t) = 0 \quad \text{for } 0 < x < a,\ t > 0,\ k = 0, 1. \tag{2.3}$$

The method of images gives the representation

(2.4)

$$G(y,x,t) = \frac{1}{2\sqrt{\pi t}} \sum_{n=-\infty}^{\infty} \left[\exp\left\{ -\frac{(y-x+2na)^2}{4t} \right\} - (-1)^k \exp\left\{ -\frac{(y+x+2na)^2}{4t} \right\} \right], \quad (k = 0, 1).$$

Note that if the infinite series is truncated after a finite number of terms, the boundary conditions are satisfied only in an asymptotic sense as $t \to 0$. That is, the boundary

values of the truncated solution decay exponentially fast in t^{-1} as $t \to 0$, and the exponential rate increases together with the number of retained terms.

The trace is given by

$$\begin{aligned}\int_0^a G(x,x,t)\,dx &= \frac{1}{2\sqrt{\pi t}}\int_0^a \sum_{n=-\infty}^{\infty}\left[\exp\left\{-\frac{(na)^2}{t}\right\}+(-1)^k\exp\left\{-\frac{(x+na)^2}{t}\right\}\right]dx\\ &= \frac{1}{2\sqrt{\pi t}}\sum_{n=-\infty}^{\infty}\left[a\exp\left\{-\frac{(na)^2}{t}\right\}+(-1)^k\int_0^a\exp\left\{-\frac{(x+na)^2}{t}\right\}dx\right]\\ &= \frac{a}{2\sqrt{\pi t}}\sum_{n=-\infty}^{\infty}\exp\left\{-\frac{(na)^2}{t}\right\}+\frac{(-1)^k}{2}\\ &= \frac{a}{2\sqrt{\pi t}}+\frac{(-1)^k}{2}+\frac{a}{2\sqrt{\pi t}}\sum_{n\neq 0}\exp\left\{-\frac{(na)^2}{t}\right\},\quad (k=0,1). \qquad (2.5)\end{aligned}$$

On the other hand,

$$\int_0^a G(x,x,t)\,dx = \sum_{n=1}^{\infty} e^{-\lambda_n t}, \qquad (2.6)$$

where $\{\lambda_n\}$ are the eigenvalues of the homogeneous Dirichlet or Neumann problem for the operator d^2/dx^2 in the interval $[0,a]$. Thus

$$\sum_{n=1}^{\infty} e^{-\lambda_n t} = \frac{a}{2\sqrt{\pi t}}+\frac{(-1)^k}{2}+\frac{a}{2\sqrt{\pi t}}\sum_{n\neq 0}\exp\left\{-\frac{(na)^2}{t}\right\},\quad (k=0,1). \qquad (2.7)$$

If instead of a single interval of length a we consider the heat equation in a set Ω consisting of K disjoint intervals of lengths l_j, $(j=1,\dots,K)$, respectively, the resulting expansion is

$$\sum_{n=1}^{\infty} e^{-\lambda_n t} = \frac{\sum_{j=1}^K l_j}{2\sqrt{\pi t}}+(-1)^k\frac{2K}{4}+\sum_{j=1}^K\frac{l_j}{2\sqrt{\pi t}}\sum_{n\neq 0}\exp\left\{-\frac{(nl_j)^2}{t}\right\},\quad (k=0,1). \qquad (2.8)$$

The numerator in the first term on the right-hand side of (2.8) can be interpreted as the "area" of Ω, so we denote it $\sum_{j=1}^K l_j = |\Omega|$. The number $2K$ is the number of boundary points of Ω, which can be interpreted as the "circumference" of the boundary, so we denote it $|\partial\Omega| = 2K$. The exponents in the sum on the right-hand side of (2.8) can be interpreted as the "widths" of the components of Ω. Clearly, for small t, the term containing the smallest width, $r = \min_{1\le j\le K} l_j$, will dominate the sum. Thus we can rewrite (2.8) as

$$\sum_{n=1}^{\infty} e^{-\lambda_n t} = \frac{|\Omega|}{2\sqrt{\pi t}}-\frac{|\partial\Omega|}{4}+\frac{mr}{\sqrt{\pi t}}\exp\left\{-\frac{r^2}{t}\right\}+\sum_{l_j>r}\frac{l_j}{2\sqrt{\pi t}}\sum_{n\neq 0}\exp\left\{-\frac{(nl_j)^2}{t}\right\}, \qquad (2.9)$$

where m is the number of the shortest intervals in Ω.

Equation (2.9) can be viewed as the short-time asymptotic expansion of the sum on the left-hand side of the equation. The algebraic part of the expansion consists of the first two terms, and all other terms are transcendentally small. The geometric information in the various terms of the expansion consists of the "area" of Ω and the

"circumference" $|\partial\Omega|$ in the algebraic part of the expansion. The transcendental part of the expansion is dominated by the term containing the smallest "width" of the domain, r.

The geometric information about Ω contained in the algebraic part is the information given in the "Can one hear the shape of a drum?" expansions [2], [3]. The geometric information contained in the transcendentally small terms in (2.9) can be understood as follows. The terms nl_j in the exponents are the lengths of closed trajectories of billiard balls in Ω, or the lengths of closed rays reflected at the boundaries, as in [7].

The representation (2.4) can be constructed as a short-time approximation to the solution of the heat equation (2.1)–(2.3) by the *ray method* [14]. In this method the solution is constructed in the form

$$G(y,x,t) = e^{-S^2(y,x)/4t} \sum_{n=0}^{\infty} Z_n(y,x) t^{n-1/2}. \tag{2.10}$$

Substituting the expansion (2.10) into the heat equation (2.1) and ordering terms by orders of magnitude for small t, we obtain at the leading order the *ray equation*, also called the *eikonal equation*,

$$\left| \frac{\partial S(y,x)}{\partial y} \right|^2 = 1, \tag{2.11}$$

and at the next orders the *transport equations*

$$\begin{aligned} &2\frac{\partial S(y,x)}{\partial y}\frac{\partial Z_n(y,x)}{\partial y} + Z_n(y,x)\left(\frac{\partial^2 S(y,x)}{\partial y^2} + \frac{2n}{S(y,x)}\right), \\ &\frac{2}{S(y,x)}\frac{\partial^2 Z_{n-1}(y,x)}{\partial y^2}, \quad n = 0,1,\dots . \end{aligned} \tag{2.12}$$

Denoting

$$p(y,x) = \frac{\partial S(y,x)}{\partial y},$$

we write the equations of the *characteristics*, or *rays* of the eikonal equation (2.11), as [15]

$$\frac{\partial y(\tau,x)}{\partial t} = 2p, \quad \frac{dp(\tau)}{dt} = 0, \quad \frac{dS(\tau)}{d\tau} = 2p^2(\tau) \tag{2.13}$$

with the initial conditions

$$y(0,x) = x, \quad p(0) = \pm 1, \quad S(0) = 0.$$

The condition $S(0) = 0$ is implied by the initial condition $G(x,y,0) = \delta(x-y)$. The solutions are given by

$$y(\tau,x) = x + 2p\tau, \quad p(\tau) = \pm 1, \quad S(\tau) = 2\tau = \pm(y-x). \tag{2.14}$$

Thus $S(y,x)$ is the length of the ray from y to x. We denote this solution by $S_0(y,x)$. It is easy to see that the solution of the transport equations corresponding to $S_0(y,x)$

is given by $Z_0(y,x) = \text{const}$, and $Z_n(y,x) = 0$ for all $n \geq 1$. The initial condition (2.2) implies that

$$Z_0(y,x) = \frac{1}{2\sqrt{\pi}}.$$

Combined into (2.10) this solution gives Green's function for the heat equation on the entire line,

$$G_0(y,x,t) = \frac{1}{2\sqrt{\pi t}} \exp\left\{-\frac{(y-x)^2}{4t}\right\},$$

which is the positive term corresponding to $n = 0$ in the expansion (2.4).

The ray from x to y is not the only one emanating from x. There are rays emanating from x that end at y after reflection in the boundary. Thus the ray from x that reaches y after it is reflected at the boundary 0 has length $y + x$. Therefore there is another solution of the eikonal equation, $S_1(y,x)$, which is the length of the reflected ray, given by

$$S_1(y,x) = y + x.$$

The ray from x that reaches y after it is reflected at the boundary a has length $2a - x - y$. The ray from x to 0, then to a, and then to y has length $2a + x - y$. Thus the lengths of all rays that reach y from x after any number of reflections in the boundary generate solutions of the eikonal equation, which are the lengths of the rays, which in turn generate solutions of the heat equation. We denote them by $S_k(y,x)$ with some ordering. The corresponding solutions of the transport equation are

$$Z_{0,k}(y,x) = \frac{C_k}{2\sqrt{\pi}},$$

where C_k are constant. They are chosen so that the sum of all the ray solutions,

$$G_k(y,x,t) = \frac{Z_{0,k}(y,x)}{\sqrt{t}} e^{-S_k^2(y,x)/4t},$$

satisfies the boundary conditions (2.3). Note that for all $k \neq 0$

$$G_k(y,x,t) \to 0 \quad \text{as } t \to 0.$$

This construction recovers the solution (2.4).

3. The ray method for short-time asymptotics of Green's function. The ray method consists in the construction of Green's function $G(\boldsymbol{y},\boldsymbol{x},t)$ (1.1)–(1.3) in the asymptotic form

$$G(\boldsymbol{y},\boldsymbol{x},t) \sim e^{-S^2(\boldsymbol{y},\boldsymbol{x})/4t} \sum_{n=0}^{\infty} Z_n(\boldsymbol{y},\boldsymbol{x}) t^{n-1}. \tag{3.1}$$

The function $S(\boldsymbol{y},\boldsymbol{x})$ is the solution of the *eikonal equation*

$$\left|\nabla_{\boldsymbol{y}} S(\boldsymbol{y},\boldsymbol{x})\right|^2 = 1, \tag{3.2}$$

and the functions $Z_n(\boldsymbol{y},\boldsymbol{x})$ solve the *transport equations*

$$2\nabla_{\boldsymbol{y}} S(\boldsymbol{y},\boldsymbol{x})\cdot\nabla_{\boldsymbol{y}} Z_n(\boldsymbol{y},\boldsymbol{x}) + Z_n(\boldsymbol{y},\boldsymbol{x})\left[\Delta_{\boldsymbol{y}} S(\boldsymbol{y},\boldsymbol{x}) + \frac{2n-1}{S(\boldsymbol{y},\boldsymbol{x})}\right]$$

$$= \frac{2}{S(\boldsymbol{y},\boldsymbol{x})}\Delta_{\boldsymbol{y}} Z_{n-1}(\boldsymbol{y},\boldsymbol{x}) \quad \text{for } n = 0,1,2,\dots. \tag{3.3}$$

The eikonal equation (3.2) is solved by the *method of characteristics* [15]. The characteristics, called *rays*, satisfy the differential equations

$$\frac{d\boldsymbol{y}(\tau,\boldsymbol{x})}{d\tau} = 2\nabla_{\boldsymbol{y}} S(\boldsymbol{y}(\tau,\boldsymbol{x}),\boldsymbol{x}), \quad \frac{d\nabla_{\boldsymbol{y}} S(\boldsymbol{y}(\tau,\boldsymbol{x}),\boldsymbol{x})}{d\tau} = 0, \quad \frac{dS(\boldsymbol{y}(\tau,\boldsymbol{x}),\boldsymbol{x})}{d\tau} = 2. \tag{3.4}$$

The initial condition (1.2) implies that the rays emanate from the point $\boldsymbol{x}$. Thus we choose the initial conditions

$$\boldsymbol{y}(0,\boldsymbol{x}) = \boldsymbol{x}, \quad \nabla_{\boldsymbol{y}} S(\boldsymbol{y}(0,\boldsymbol{x}),\boldsymbol{x}) = \boldsymbol{\nu}, \quad S(\boldsymbol{y}(0,\boldsymbol{x}),\boldsymbol{x}) = 0, \tag{3.5}$$

where $\boldsymbol{\nu}$ is a constant vector of unit length. The solution is given by

$$\boldsymbol{y}(\tau,\boldsymbol{x}) = \boldsymbol{x} + 2\boldsymbol{\nu}\tau, \quad S(\boldsymbol{y},\boldsymbol{x}) = |\boldsymbol{y}-\boldsymbol{x}| = 2\tau, \quad \nabla_{\boldsymbol{y}} S(\boldsymbol{y},\boldsymbol{x}) = \boldsymbol{\nu}. \tag{3.6}$$

The pair $(\tau,\boldsymbol{\nu})$ determines uniquely the point $\boldsymbol{y} = \boldsymbol{y}(\tau,\boldsymbol{x})$ and the value of $S(\boldsymbol{y},\boldsymbol{x})$ at the point. The parameter τ is half the distance from $\boldsymbol{y}$ to $\boldsymbol{x}$ or half the length of the ray from $\boldsymbol{x}$ to $\boldsymbol{y}$. The vector $\boldsymbol{\nu}$ is the unit vector in the direction from $\boldsymbol{x}$ to $\boldsymbol{y}$.

The function $Z_0(\boldsymbol{y},\boldsymbol{x})$ is easily seen to be a constant, $1/4\pi$, and $Z_n(\boldsymbol{y},\boldsymbol{x}) = 0$ for all $n > 0$. This construction recovers the solution of the heat equation in the entire plane and disregards the boundary $\partial\Omega$, because in the plane every point can be seen from every other point by a straight ray. Note that to calculate the function $P(t)$ in (1.5), only the values of $S(\boldsymbol{x},\boldsymbol{x})$ and $Z_0(\boldsymbol{x},\boldsymbol{x})$ are needed. Thus $S(\boldsymbol{x},\boldsymbol{x}) = 0$, and the first approximation to $G(\boldsymbol{x},\boldsymbol{x},t)$ is

$$G(\boldsymbol{x},\boldsymbol{x},t) = \frac{1}{4\pi t};$$

hence the first approximation to $P(t)$ is

$$P_0(t) = \frac{|\Omega|}{4\pi t}.$$

There is another solution of the eikonal equation (3.2) constructed along rays that emanate from $\boldsymbol{x}$ but reach $\boldsymbol{y}$ after they are reflected in $\partial\Omega$ [14]. The law of reflection is determined from the boundary conditions. Dirichlet and Neumann boundary conditions imply that the angle of incidence equals that of reflection [14]. Similarly, there are solutions of the eikonal equation that are the lengths of rays that emanate from $\boldsymbol{x}$ and reach $\boldsymbol{y}$ after any number of reflections in $\partial\Omega$. We denote these solutions $S_k(\boldsymbol{y},\boldsymbol{x})$ with some ordering. Thus the full ray expansion of Green's function has the form

$$G(\boldsymbol{y},\boldsymbol{x},t) \sim \sum_{k=1}^{\infty} e^{-S_k^2(\boldsymbol{y},\boldsymbol{x})/4t} Z_k(\boldsymbol{y},\boldsymbol{x},t), \tag{3.7}$$

where

$$Z_k(\boldsymbol{y},\boldsymbol{x},t) = \sum_{n=0}^{\infty} Z_{n,k}(\boldsymbol{y},\boldsymbol{x}) t^{n-1}.$$

As above, each one of the series

$$e^{-S_k^2(\boldsymbol{y},\boldsymbol{x})/4t} Z_k(\boldsymbol{y},\boldsymbol{x},t)$$

is called a *ray solution* of the diffusion equation. The boundary values of $Z_k(\boldsymbol{y},\boldsymbol{x},t)$ are chosen so that $G(\boldsymbol{y},\boldsymbol{x},t)$ in (3.7) satisfies the imposed boundary condition. In particular, the values of $S_k(\boldsymbol{x},\boldsymbol{x})$ are the lengths of all rays that emanate from $\boldsymbol{x}$ and are reflected from the boundary back to $\boldsymbol{x}$. Note that sums of ray solutions satisfy boundary conditions only at certain points.

To fix the ideas, we consider first simply connected domains. We denote

$$S_0(\boldsymbol{y},\boldsymbol{x}) = |\boldsymbol{x}-\boldsymbol{y}|$$

and

$$G_0(\boldsymbol{y},\boldsymbol{x},t) = \frac{1}{4\pi t} e^{-S_0^2(\boldsymbol{y},\boldsymbol{x})/4t}.$$

We first consider solutions corresponding to rays that are reflected only once at the boundary and, in particular, rays that are reflected back from the boundary to the points of their origin. Such rays hit the boundary at right angles (see Figure 1 and [14]). If there is only one minimal eikonal $S_1(\boldsymbol{x},\boldsymbol{x}) > 0$, we say that $\boldsymbol{x}$ is a *regular* point of Ω. If there is more than one minimal eikonal $S_1(\boldsymbol{x},\boldsymbol{x})$, we say that $\boldsymbol{x}$ is a *critical* point of Ω. We denote by Γ the locus of critical points in Ω. The eikonal $S_1(\boldsymbol{y},\boldsymbol{x})$ is the length of the shortest ray from $\boldsymbol{x}$ to $\boldsymbol{y}$ with one reflection in the boundary such that the ray from $\boldsymbol{x}$ to the boundary does not intersect Γ. For $\boldsymbol{x} = \boldsymbol{y}$ the eikonal $S_1(\boldsymbol{x},\boldsymbol{x})$ is twice the distance of $\boldsymbol{x}$ to the boundary. We denote by $\boldsymbol{x}'$ the orthogonal projection of $\boldsymbol{x}$ on the boundary along the shortest normal from $\boldsymbol{x}$ to the boundary. When $\boldsymbol{y} = \boldsymbol{x}'$,

$$S_1(\boldsymbol{x}',\boldsymbol{x}) = S_0(\boldsymbol{x}',\boldsymbol{x}) = |\boldsymbol{x}-\boldsymbol{x}'|\,. \tag{3.8}$$

The function

$$G_1(\boldsymbol{y},\boldsymbol{x},t) = e^{-S_1^2(\boldsymbol{y},\boldsymbol{x})/4t} Z_1\left(\boldsymbol{y},\boldsymbol{x},t\right)$$

has to be chosen so that $G_0(\boldsymbol{x}',\boldsymbol{x},t) - G_1(\boldsymbol{x}',\boldsymbol{x},t) = 0$. In view of (3.8), we have to choose

$$Z_1\left(\boldsymbol{x}',\boldsymbol{x},t\right) = \frac{1}{4\pi t}.$$

When $\boldsymbol{y}''$ is the other boundary point on the normal from $\boldsymbol{x}'$ to $\boldsymbol{x}$, we have

$$G_0(\boldsymbol{y}'',\boldsymbol{x},t) - \ G_1(\boldsymbol{y}'',\boldsymbol{x},t) = \frac{1}{4\pi t} e^{-|x-y''|^2/t} - e^{-\left(|x'-x|+|y''-x'|\right)^2/t} Z_1\left(\boldsymbol{y}'',\boldsymbol{x},t\right). \tag{3.9}$$

Next, we consider in $\Omega - \Gamma$ the minimal among the remaining eikonals $S_k(\boldsymbol{x},\boldsymbol{x}) > S_1(\boldsymbol{x},\boldsymbol{x})$ and denote it $S_2(\boldsymbol{x},\boldsymbol{x})$. This eikonal is twice the length of a ray that emanates from $\boldsymbol{x}$, intersects Γ once, and intersects the boundary $\partial\Omega$ at right angles at a point, denoted $\boldsymbol{x}''$. The eikonal $S_2(\boldsymbol{y},\boldsymbol{x})$ is the length of the ray from $\boldsymbol{x}$ to $\boldsymbol{y}$ with one reflection in the boundary such that the ray from $\boldsymbol{x}$ to the boundary intersects Γ once. When $\boldsymbol{y} = \boldsymbol{x}''$,

$$S_2(\boldsymbol{x}'',\boldsymbol{x}) = S_0(\boldsymbol{x}'',\boldsymbol{x}) = |\boldsymbol{x}-\boldsymbol{x}''|\,. \tag{3.10}$$

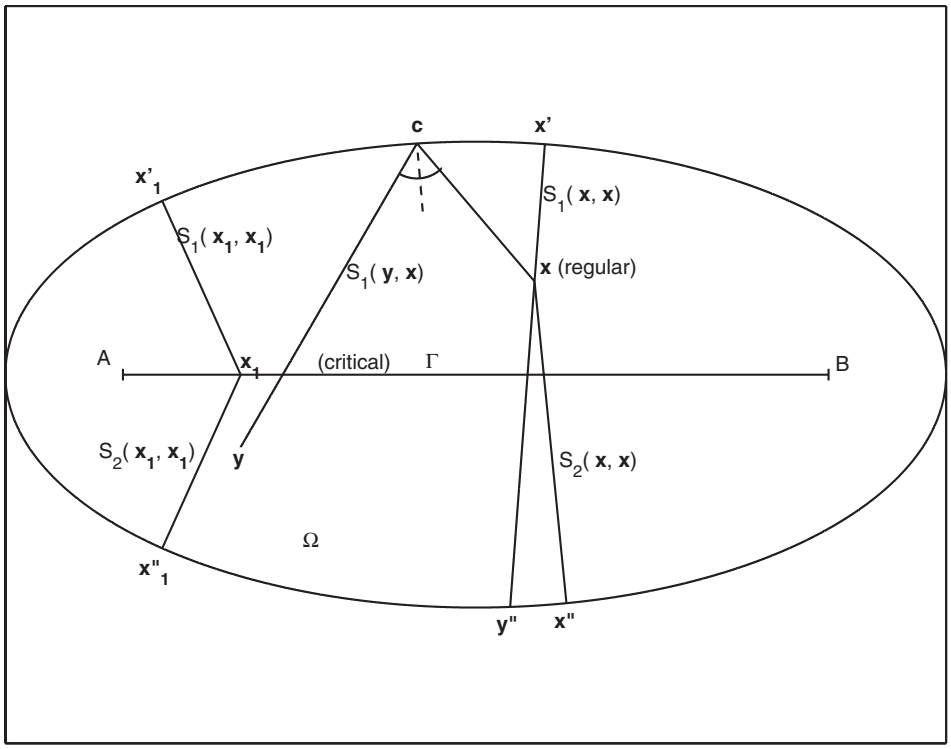


FIG. 1. *The locus of critical points, Γ, is the segment AB. The first eikonal is $S_1(\boldsymbol{y},\boldsymbol{x}) = |\boldsymbol{x}-\boldsymbol{c}|+|\boldsymbol{c}-\boldsymbol{y}|$. It is defined as the shortest reflected ray from $\boldsymbol{x}$ to $\boldsymbol{y}$, such that $\boldsymbol{x}-\boldsymbol{c}$ does not intersect Γ. For $\boldsymbol{x}=\boldsymbol{y}$ the diagonal values are $S_1(\boldsymbol{x},\boldsymbol{x}) = 2|\boldsymbol{x}-\boldsymbol{x'}|$. The diagonal values of the second eikonal are $S_2(\boldsymbol{x},\boldsymbol{x}) = 2|\boldsymbol{x}-\boldsymbol{x''}|$. The vectors $\boldsymbol{x}-\boldsymbol{x'}$ and $\boldsymbol{x}-\boldsymbol{x''}$ are orthogonal to the boundary. For $\boldsymbol{x}_1\in\Gamma$ the two eikonals are equal.*

When $\boldsymbol{y}'$ is the other boundary point on the normal that emanates from $\boldsymbol{x}''$ (see Figure 2), we have

$$S_2(\boldsymbol{y}',\boldsymbol{x}) = |\boldsymbol{x}-\boldsymbol{x}''|+|\boldsymbol{y}'-\boldsymbol{x}''|\,.$$

In general, $\boldsymbol{x}'\neq\boldsymbol{y}'$ and $\boldsymbol{x}''\neq\boldsymbol{y}''$. However, if the ray is a 2-periodic orbit (that hits the boundary at only 2 points), $\boldsymbol{x}'=\boldsymbol{y}'$ and $\boldsymbol{x}''=\boldsymbol{y}''$ so that

$$S_2(\boldsymbol{y}'',\boldsymbol{x}) = S_0(\boldsymbol{y}'',\boldsymbol{x}) = |\boldsymbol{x}-\boldsymbol{y}''|$$

and

$$S_2(\boldsymbol{y}',\boldsymbol{x}) = |\boldsymbol{x}-\boldsymbol{x}''|+|\boldsymbol{y}''-\boldsymbol{x}'|\,.$$

Since

$$|\boldsymbol{x}-\boldsymbol{y}''| < |\boldsymbol{x}-\boldsymbol{x}'|+|\boldsymbol{y}''-\boldsymbol{x}'| < |\boldsymbol{x}-\boldsymbol{x}''|+|\boldsymbol{y}''-\boldsymbol{x}'|$$

for all regular points $\boldsymbol{x}$, the order of magnitude of the boundary error (3.10) decreases if we use the approximation

$$G_0(\boldsymbol{y},\boldsymbol{x},t) \sim G_0(\boldsymbol{y},\boldsymbol{x},t) - G_1(\boldsymbol{y},\boldsymbol{x},t) - G_2(\boldsymbol{y},\boldsymbol{x},t) \tag{3.11}$$

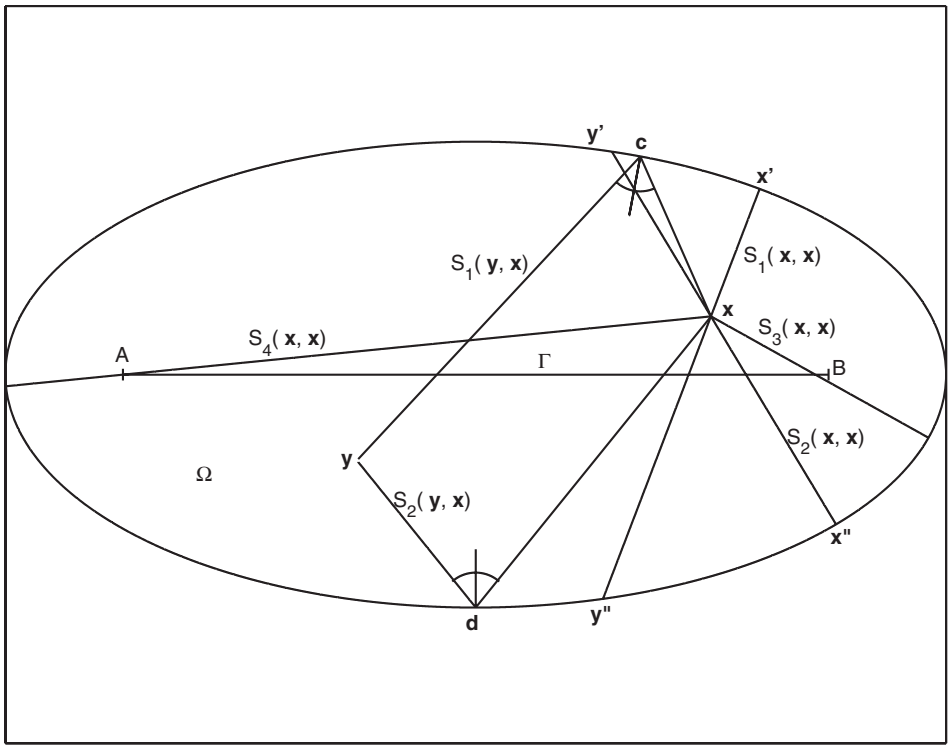


FIG. 2. *The second eikonal* $S_2(\boldsymbol{y}, \boldsymbol{x}) = |\boldsymbol{x} - \boldsymbol{d}| + |\boldsymbol{d} - \boldsymbol{y}|$. *It is defined as the shortest reflected ray such that* $\boldsymbol{x} - \boldsymbol{d}$ *intersects* Γ. *The eikonals* $S_3(\boldsymbol{x}, \boldsymbol{x})$ *and* $S_4(\boldsymbol{x}, \boldsymbol{x})$ *are ordered according to magnitude.*

with

$$Z_2(\boldsymbol{y}'', \boldsymbol{x}, t) = Z_1(\boldsymbol{y}'', \boldsymbol{x}, t) = Z_0(t).$$

4. The trace. To find the short-time asymptotics of the Dirichlet series (1.6), as given in (1.5),

$$P(t) = \int_\Omega G(\boldsymbol{x}, \boldsymbol{x}, t)\, d\boldsymbol{x},$$

we use the ray expansion (3.7) for the evaluation of the integral. We retain in the resulting expansion only terms that are transcendentally small, since all algebraic terms are contained in the expansion (1.8).

4.1. Simply connected domains. We note that according to Sard's theorem, Γ is a set of measure zero and that all points in the domain $\Omega - \Gamma$ are regular. For any point $\boldsymbol{x} \in \Omega$, we denote by $r_1(\boldsymbol{x})$ its distance to the boundary and note that $S_1(\boldsymbol{x}, \boldsymbol{x}) = 2r_1(\boldsymbol{x})$. We also denote by $s_1(\boldsymbol{x})$ the arclength at the boundary point $\boldsymbol{x}'$ (the orthogonal projection of $\boldsymbol{x}$ on $\partial\Omega$ along the shortest normal from $\boldsymbol{x}$ to $\partial\Omega$), measured from a boundary point where the arclength is set to 0 (see Figure 3).

It follows that the change of variables in $\Omega - \Gamma$, given by

$$\boldsymbol{x} \to (r_1(\boldsymbol{x}), s_1(\boldsymbol{x})), \tag{4.1}$$

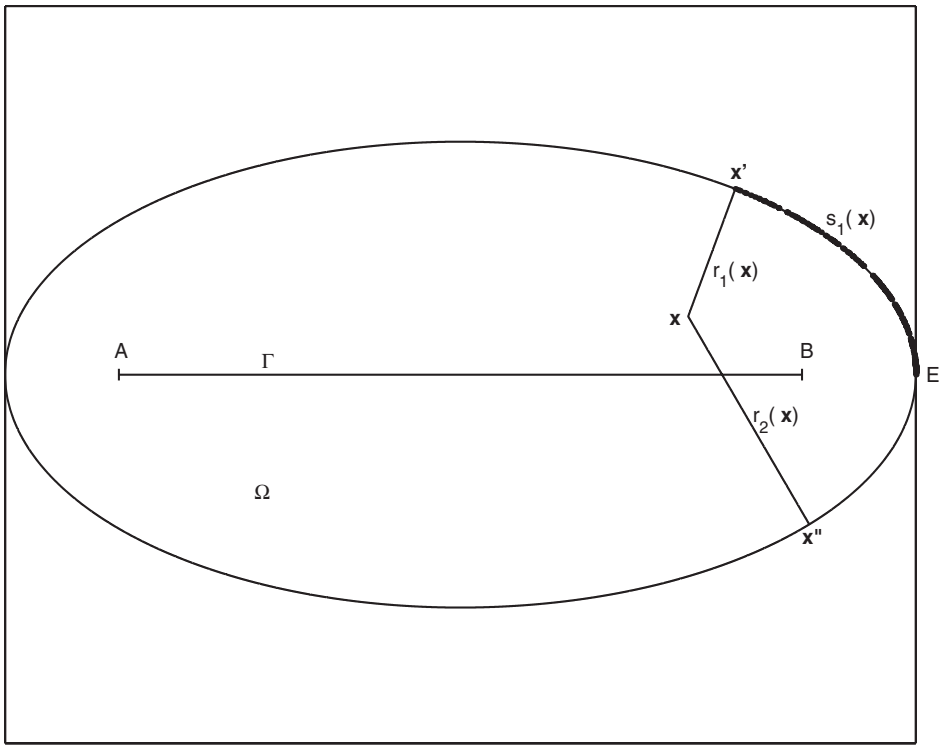


FIG. 3. *The arclength $s_1(\boldsymbol{x})$ is measured from the point E. Both transformations $\boldsymbol{x} \to (r_1(\boldsymbol{x}), s_1(\boldsymbol{x}))$ and $\boldsymbol{x} \to (r_2(\boldsymbol{x}), s_1(\boldsymbol{x}))$ are one-to-one mappings of $\Omega - \Gamma$. The images are given in Figure 4.*

is a one-to-one mapping of $\Omega - \Gamma$ onto a strip $0 \le r_1 \le r_1(s_1)$, $0 \le s_1 \le L$, where $r_1(s_1)$ is the distance from the boundary point corresponding to arclength s_1 to Γ.

We evaluate the integral over Ω separately for each summand k in the expansion (3.7). In this notation, we can write

$$\begin{aligned}
&\int_\Omega G_1(\boldsymbol{x}, \boldsymbol{x}, t)\, d\boldsymbol{x} \\
&= \int_\Omega e^{-[S_1(\boldsymbol{x},\boldsymbol{x})]^2/4t} \sum_{n=0}^{\infty} Z_{n,1}(\boldsymbol{x}, \boldsymbol{x}) t^{n-1}\, d\boldsymbol{x} \\
&= \int_0^L ds \int_0^{r_1(s_1)} e^{-r_1^2/t} J_1(r_1, s_1) Z_1(r_1, s_1, t)\, dr_1,
\end{aligned} \tag{4.2}$$

where $J_1(r_1, s_1)$ is the Jacobian of the transformation and

$$Z_1(r_1, s_1, t) = \sum_{n=0}^{\infty} Z_{n,1}(\boldsymbol{x}, \boldsymbol{x}) t^{n-1}.$$

Note that the Jacobian vanishes neither inside $\Omega - \Gamma$ nor at $r_1 = 0$, because the transformation is one-to-one in $\Omega - \Gamma$; however, it does on Γ.

We set $S_2(\boldsymbol{x},\boldsymbol{x}) = 2r_2(\boldsymbol{x})$ and use it as a coordinate. We use $s_1(\boldsymbol{x})$ as the other coordinate of the point $\boldsymbol{x} \in \Omega - \Gamma$. Note that while $r_2(\boldsymbol{x})$ is the length of the longer normal from $\boldsymbol{x}$ to $\partial\Omega$ (the one that intersects Γ), the other coordinate is the arclength corresponding to the shorter normal from $\boldsymbol{x}$ to $\partial\Omega$ (the one that does not intersect Γ). The transformation

$$\boldsymbol{x} \to (r_2(\boldsymbol{x}), s_1(\boldsymbol{x})) \tag{4.3}$$

maps $\Omega - \Gamma$ onto the strip $r\,(s_1) \le r_2 \le l(s_1),\ 0 \le s_1 \le L$, where $l(s_1)$ is the length of the segment of the normal that starts at the boundary point $r_1 = 0, s_1$ and ends at its other intersection point with the boundary. This mapping is one-to-one as well. It follows that

$$\begin{aligned}
&\int_\Omega G_2(\boldsymbol{x},\boldsymbol{x},t)\,d\boldsymbol{x} \\
&= \int_\Omega e^{-[S_2(\boldsymbol{x},\boldsymbol{x})]^2/4t} \sum_{n=0}^{\infty} Z_{n,2}(\boldsymbol{x},\boldsymbol{x}) t^{n-1}\,d\boldsymbol{x} \\
&= \int_0^L ds_1 \int_{r(s_1)}^{l(s_1)} e^{-r_2^2/t} J_2(r_2,s_1) Z_2(r_2,s_1,t)\,dr_2,
\end{aligned} \tag{4.4}$$

where

$$Z_2(r_2,s_1,t) = \sum_{n=0}^{\infty} Z_{n,2}(\boldsymbol{x},\boldsymbol{x}) t^{n-1}.$$

Note that for $\boldsymbol{x}$ on Γ both transformations, (4.1) and (4.3) are identical and

$$J_2(r_2,s_1)Z_2(r_2,s_1,t) = J_1(r_1,s_1)Z_1(r_1,s_1,t).$$

It follows that (4.2) and (4.4) combine to give

$$\begin{aligned}
&\int_\Omega \left[G_1(\boldsymbol{x},\boldsymbol{x},t) + G_2(\boldsymbol{x},\boldsymbol{x},t)\right]\,d\boldsymbol{x} \\
&= \int_0^L \int_0^{l(s)} e^{-r^2/t} J(r,s) Z(r,s,t)\,dr\,ds,
\end{aligned} \tag{4.5}$$

where $s = s_1,\ r = r_1,\ J = J_1$, and $Z = Z_1$ for $0 < r < r_1(s_1)$, and $s = s_1,\ r = r_2,\ J = J_2$, and $Z = Z_2$ for $r_2(s_1) < r < l(s_1)$. Thus the domain of integration of the function $e^{-r^2/t}J(r,s)Z(r,s,t)$ in (4.5) is the domain enclosed by the s_1-axis and the upper curve in Figure 4. Now, for $t \ll 1$, we write the inner integral on the right-hand side of (4.5) as

$$\begin{aligned}
&\int_0^{l(s)} e^{-r^2/t} J(r,s) Z(r,s,t)\,dr = \sqrt{\frac{\pi t}{2}}\operatorname{erf}\left(\frac{l(s)}{\sqrt{t}}\right) J(0,s)Z(0,s,t)\left(1 + O\left(\sqrt{t}\right)\right) \\
&= \sqrt{\frac{\pi t}{2}}\left(1 - \frac{\exp\left\{-\dfrac{l^2(s)}{t}\right\}\sqrt{t}}{l(s)}\right) J(0,s)Z(0,s,t)\left(1 + O\left(\sqrt{t}\right)\right).
\end{aligned}$$

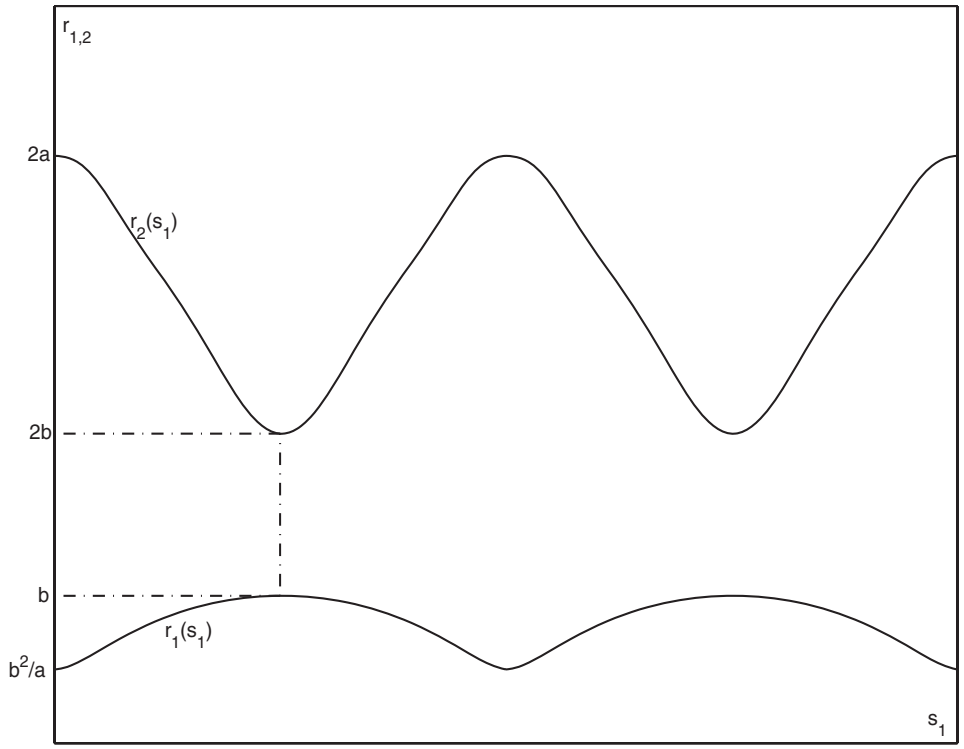


FIG. 4. *The domain* Ω *is the ellipse* $\frac{x^2}{a^2}+\frac{y^2}{b^2}\le 1$. *The domain enclosed between the* s_1*-axis and the lower curve is the image of the ellipse under the transformation* (4.1)*, and the domain enclosed between the upper and the lower curves is its image under* (4.3).

Recall that $J(0,s)Z(0,s,t)\neq 0$. Only the exponentially small terms have to be considered, because the algebraic terms are included in the Stewart–Waechter expansion. Thus

$$\int_0^L\int_0^{l(s)} e^{-r^2/t}J(r,s)Z(r,s,t)\,dr\,ds-\int_0^L\sqrt{\frac{\pi t}{2}}J(0,s)Z(0,s,t)\left(1+O\left(\sqrt{t}\right)\right)\,ds$$

$$=-\int_0^L \exp\left\{-\frac{l^2(s)}{t}\right\}\frac{J(0,s)Z(0,s,t)}{l(s)}O\left(t\right)\,ds\quad\text{for } t\ll 1.$$

Evaluating the last integral by the Laplace method, we find that each point s_i, which is an extremum point of $l(s)$, contributes an exponential term of the form

$$\exp\left\{-\frac{l^2(s_i)}{t}\right\}\frac{J(0,s_i)Z(0,s_i,t)}{l(s_i)}O\left(t^\nu\right), \tag{4.6}$$

where $\nu\ge 0$ and $O\left(t^\nu\right)$ depend on the type of the critical point s_i, and so also on the local behavior of $l^2(s)$ near s_i. The expression (4.6) means that some of the δ_n's in the expansion (1.15) are the extremal values $l(s_i)$ and their multiples. These are half the lengths of the 2-periodic orbits of a billiard ball in Ω (see Figure 5). In particular, the shortest neck is given by

$$\delta_1=\frac{l}{2}. \tag{4.7}$$

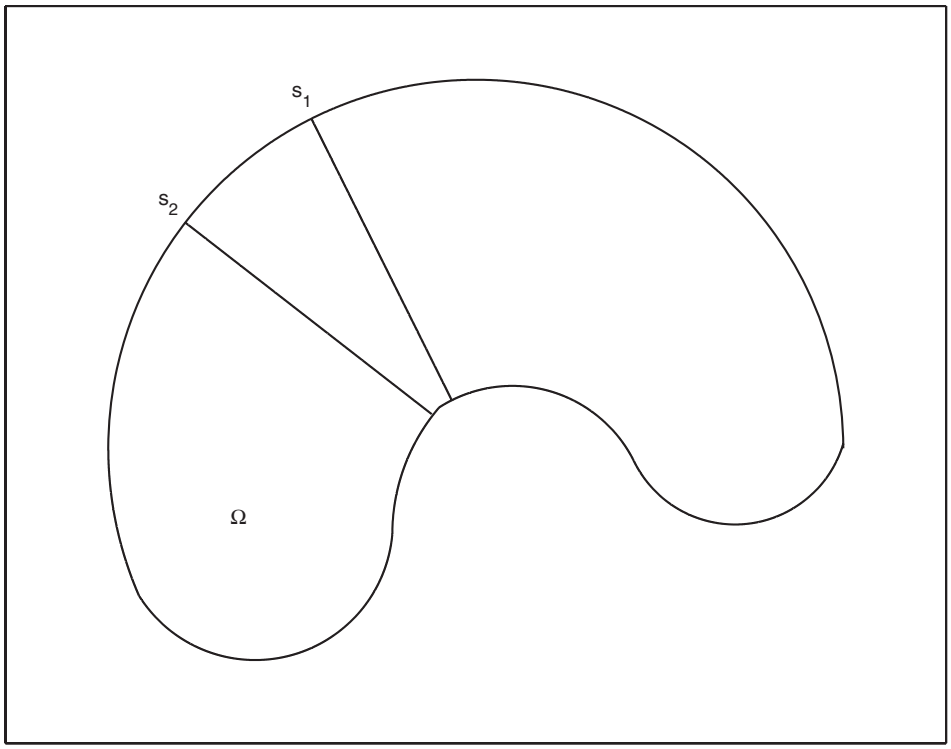


FIG. 5. *The rays emanating from the boundary points s_1 and s_2 are orthogonal to the boundary at both ends. They are 2-periodic orbits.*

The 2-periodic orbits of the ellipse are the major axes, which correspond to the lowest and highest points of the top curve in Figure 4. There are other exponents as well, as discussed below.

The preexponential terms in the expression (4.6) influence the factors $P_n(\sqrt{t})$ in (1.15). For example, if $l'(s_i) = 0,\ l''(s_i) \neq 0$, then $\nu = 3/2$. If the boundary is flatter, then $1 \leq \nu < 3/2$.

In addition to the 2-periodic orbits, there are ray solutions corresponding to rays from $\boldsymbol{x}$ to $\boldsymbol{y}$ that are reflected any number of times in the boundary. There are eikonals from $\boldsymbol{x}$ to $\boldsymbol{y}$ in Ω with $N-1$ different vertices on the boundary, which have N vertices on $\partial\Omega$ if $\boldsymbol{x} = \boldsymbol{y}$ and $\boldsymbol{x} \in \partial\Omega$ (this is a periodic orbit with $N-1$ reflections). Among these periodic orbits, there are eikonals $S_N(\boldsymbol{x}, \boldsymbol{x})$ with extremal length, denoted $S_{N,j},\ (j = 1, \ldots)$. At points $\boldsymbol{x} \in \Omega$ on a 2-periodic orbit the eikonal $S_N(\boldsymbol{x}, \boldsymbol{x})$, which now has $N-1$ vertices on the boundary, may reduce to the 2-periodic orbit with N reflections. Therefore the change of variables $\boldsymbol{x} \to (S_N(\boldsymbol{x}, \boldsymbol{x}), s(\boldsymbol{x}))$ will map the domain into a strip with extremal widths that are the differences between the lengths $S_{N,j}$ and the length of a 2-periodic orbit with N reflections. It follows that the evaluation of the trace by the Laplace method leads to exponents which are the extremal lengths of periodic orbits with any number of reflections.

For example, there is an eikonal in a circle (centered at the origin) that is the ray from $\boldsymbol{x}$ to $\boldsymbol{y}$ with two reflections in the boundary (see Figure 6). For $\boldsymbol{x} = \boldsymbol{y}$ it is the equilateral triangle (see Figure 7) with circumference

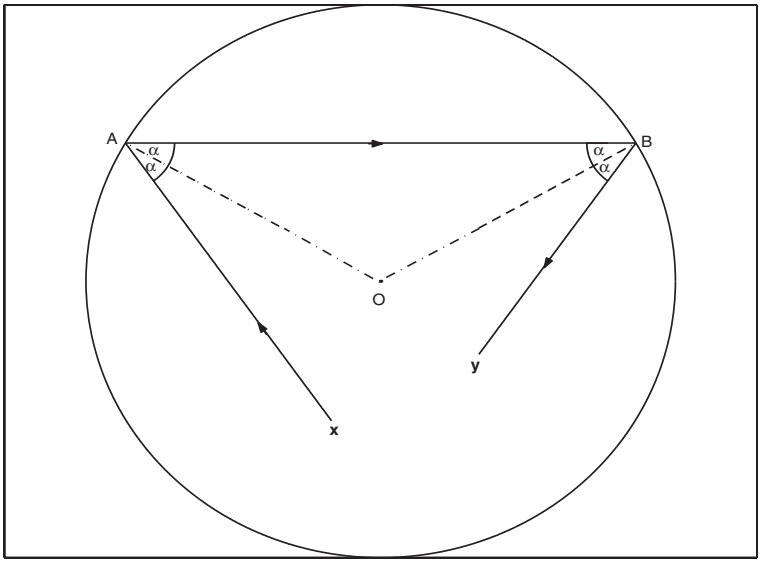


FIG. 6. *The eikonal $S_3(\boldsymbol{y},\boldsymbol{x})$ with two reflections in the circle.*

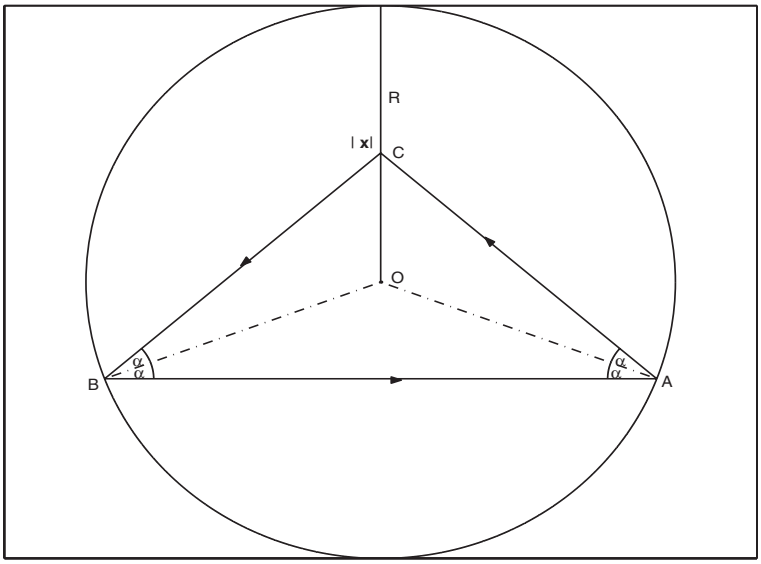


FIG. 7. *The eikonal $S_3(\boldsymbol{x},\boldsymbol{x})$ with two reflections, where $|\boldsymbol{x}| = OC$.*

$$S(\boldsymbol{x},\boldsymbol{x}) = R\left(2\frac{\sqrt{2|\boldsymbol{x}|^2+1+\sqrt{8|\boldsymbol{x}|^2+1}}}{\sqrt{4|\boldsymbol{x}|^2+1+\sqrt{8|\boldsymbol{x}|^2+1}}} + \sqrt{4|\boldsymbol{x}|^2+2+2\sqrt{8|\boldsymbol{x}|^2+1}}\right).$$

The eikonal $S_3(\boldsymbol{x},\boldsymbol{y})$ reduces to a 2-periodic orbit with two reflections if $\boldsymbol{x} = \boldsymbol{y} = 0$ (the center of the circle). If $\boldsymbol{x}$ is on the circumference, the eikonal becomes the isosceles

triangle with one vertex at $\boldsymbol{x}$. To evaluate the contribution of the corresponding ray solution to the trace, we use this eikonal as a coordinate that varies between $4R$, the length of the 2-periodic orbit with two reflections, and $3\sqrt{3}R$, the circumference of the inscribed isosceles triangle. The contribution of this integral to the exponential sum in (1.15) contains exponents that are both lengths. Similarly, the 2-periodic orbit with three reflections has length $6R$, while the periodic orbit with three reflections at three different points has length $4\sqrt{2}R < 6R$. Obviously, if the preexponential factors vanish, these exponents do not appear in the expansion.

4.2. Multiply connected domains. Once again, we first consider rays from $\boldsymbol{x}$ to $\boldsymbol{y}$ that are reflected only once in the boundary. For every connected component of $\partial\Omega$, denoted $\partial\Omega_i$ $(i = 1, \ldots, I)$, a point $\boldsymbol{x}$ in Ω is regular with respect to $\partial\Omega_i$ if there is only one minimal eikonal $S_{i,1}(\boldsymbol{x}, \boldsymbol{x}) > 0$ with one reflection at $\partial\Omega_i$. We denote by Γ_i the locus of the irregular points of Ω with respect to $\partial\Omega_i$.

As above, we define in $\Omega - \Gamma_i$ the minimal eikonal with one reflection in $\partial\Omega_i$ such that $S_{i,2}(\boldsymbol{x}, \boldsymbol{x}) > S_{i,1}(\boldsymbol{x}, \boldsymbol{x})$. We construct an approximation

$$G(\boldsymbol{y}, \boldsymbol{x}, t) \sim G_0(\boldsymbol{y}, \boldsymbol{x}, t) + \sum_{k=1}^{2} \sum_{i=1}^{I} G_{i,k}(\boldsymbol{y}, \boldsymbol{x}, t),$$

where $G_{i,k}(\boldsymbol{y}, \boldsymbol{x}, t)$ are ray solutions with eikonals $S_{i,k}(\boldsymbol{y}, \boldsymbol{x})$ and $Z_{i,k}(\boldsymbol{y}, \boldsymbol{x})$ chosen so as to minimize the boundary values of the sum at the boundary points of rays orthogonal to the boundary, as above. The trace of the double sum is calculated by introducing the change of variables $2r_{i,k}(\boldsymbol{x}) = S_{i,k}(\boldsymbol{x}, \boldsymbol{x})$ and arclength $s_i(\boldsymbol{x})$ in $\partial\Omega_i$, as above. The Laplace evaluation of the integrals produces exponents that are the 2-periodic orbits in Ω.

Eikonals with two or more reflections contribute exponents that are lengths of extremal closed orbits with any number of reflections in the boundary, as in the case of simply connected domains. Thus the exponents δ_n in (1.15) consist of half the lengths of 2-periodic orbits in Ω and their multiples, and extremal lengths of closed periodic orbits with any number of reflections in the boundary and their multiples.

5. Recovering $\boldsymbol{\delta_1}$ from $\boldsymbol{P(t)}$. We denote the sum of $N \geq 3$ terms in (1.8)

$$Q_N(t) \sim \frac{|\Omega|}{4\pi t} - \frac{|\partial\Omega|}{8\sqrt{\pi t}} + \frac{1}{6}(1 - r) + \sum_{n=3}^{N} a_n t^{n/2-1} \quad \text{for } t \to 0. \tag{5.1}$$

To recover the geometrical information from the expansion (1.15), given the (measured) function $P(t)$, we note that

$$|\Omega| = \lim_{t\to 0} 4\pi t P(t), \quad |\partial\Omega| = -\lim_{t\to 0} 8\sqrt{\pi t}\left[P(t) - \frac{|\Omega|}{4\pi t}\right], \tag{5.2}$$

and so on. This way any number of terms in the expansion (1.8) can be determined.

Once $Q_n(t)$ has been determined, we can write

$$t[P(t) - Q_{2n-1}(t)] \sim a_{2n} t^n \left(1 + O\left(\sqrt{t}\right)\right) + \frac{P_1(0)}{\sqrt{\pi}} \left(1 + O\left(\sqrt{t}\right)\right) e^{-\delta_1^2/t}, \tag{5.3}$$

where the first $O\left(\sqrt{t}\right)$ may depend on n. According to (1.9) and to Stirling's formula,

$$a_{2n} = \frac{\alpha}{l^{2n-2}} O\left(\frac{\Gamma(2n)}{\Gamma(n)}\right) = O\left(AB^n n^n\right) \quad \text{for} \quad n \to \infty, \tag{5.4}$$

where

$$A = l^2\alpha, \quad B = \frac{4}{el^2}. \tag{5.5}$$

We will show that for some $\delta > 0$

$$\max_{0\le t\le\delta}\left\{-t\log\left|a_{2n}t^n\left(1+O\left(\sqrt{t}\right)\right)+\frac{P_1(0)}{\sqrt{\pi}}\left(1+O\left(\sqrt{t}\right)\right)e^{-\delta_1^2/t}\right|\right\}\to\delta_1^2 \quad \text{as} \quad n\to\infty.$$

Indeed, setting $b = \frac{P_1(0)}{\sqrt{\pi}}$, assuming $A, B, b > 0$, and defining

$$g(t) = -t\log\left(AB^n n^n t^n + be^{-\delta_1^2/t}\right), \tag{5.6}$$

we set $nt = u$ and rewrite (5.6) as

(5.7)

$$g(t) = -u\log\sqrt[n]{A\left(\frac{4}{el^2}\right)^n u^n + be^{-n\delta_1^2/u}} \to -u\max\left\{\log\frac{4}{el^2}, -\frac{\delta_1^2}{u}\right\} \quad \text{as} \quad n\to\infty$$

$$= \max\left\{\delta_1^2, e\left(\frac{l}{2}\right)^2(-y\log y)\right\},$$

where

$$y = \frac{u}{e}\left(\frac{2}{l}\right)^2.$$

Because

$$-y\log y \le e^{-1},$$

we can write, according to (4.7),

$$\max_{0<t<\delta} g(t) \to \max\left\{\delta_1^2, \left(\frac{l}{2}\right)^2\right\} = \delta_1^2 = \left(\frac{l}{2}\right)^2. \tag{5.8}$$

Note that the maximum is achieved at

$$t_{\max} = \frac{1}{en} \to 0 \quad \text{as} \quad n\to\infty. \tag{5.9}$$

Equation (5.8) implies the following algorithm for determining δ_1 from the given values of $P(t)$. First, use the method of (5.2) to construct $Q_n(t)$ for a given n and then find

$$\max_t\{-t\log|t[P(t)-Q_n(t)]|\} = \delta_1^2 + o(1) \quad \text{as} \quad n\to\infty.$$

This algorithm works quite well with MAPLE.

6. Discussion. We illustrate our expansion for a disk, whose boundary has only one connected component and a single critical point. We consider points $\boldsymbol{x} = (x_1, y_1)$ and $\boldsymbol{y} = (x_2, y_2)$ inside a circle of radius R centered at the origin. The leading-order eikonal is

$$S_0(\boldsymbol{y}, \boldsymbol{x}) = |\boldsymbol{x} - \boldsymbol{y}|.$$

When both $\boldsymbol{x}$ and $\boldsymbol{y}$ are on the x-axis, we have $y_1 = y_2 = 0$ and $S_0(\boldsymbol{y}, \boldsymbol{x}) = |x_1 - x_2|$. Denoting $\boldsymbol{x}_1 = (x_1, 0)$ and $\boldsymbol{x}_2 = (x_2, 0)$, we see that the values of the eikonal on the x-axis are $S_0(\boldsymbol{x}_1, \boldsymbol{x}_2) = |x_1 - x_2|$. We assume that $x_1 > 0$. The boundary values of the eikonal are

$$S_0(\boldsymbol{x}_1, \boldsymbol{x}_2) = R - x_1 \quad \text{at} \quad \boldsymbol{x}_2 = (R, 0)$$

and

$$S_0(\boldsymbol{x}_1, \boldsymbol{x}_2) = R + x_1 \quad \text{at} \quad \boldsymbol{x}_2 = (-R, 0).$$

Thus the leading-order ray approximation to Green's function $G(\boldsymbol{y}, \boldsymbol{x}, t)$,

$$G_0(\boldsymbol{y}, \boldsymbol{x}, t) = \frac{1}{4\pi t} e^{-S_0^2(\boldsymbol{y}, \boldsymbol{x})/4t},$$

misses the boundary conditions when $\boldsymbol{x}$ and $\boldsymbol{y}$ are on the x-axis, giving

$$G_0(\boldsymbol{x}_1, \boldsymbol{x}_2, t) = \frac{1}{4\pi t} e^{-(R - x_1)^2/4t} \quad \text{at} \quad \boldsymbol{x}_2 = (R, 0)$$

and

$$G_0(\boldsymbol{x}_1, \boldsymbol{x}_2, t) = \frac{1}{4\pi t} e^{-(R + x_1)^2/4t} \quad \text{at } \boldsymbol{x}_2 = (-R, 0). \tag{6.1}$$

The next eikonal, denoted $S_1(\boldsymbol{y}, \boldsymbol{x})$, is given on the x-axis by $S_1(\boldsymbol{x}_1, \boldsymbol{x}_2) = 2R - x_1 - x_2$, and its boundary values are

$$S_1(\boldsymbol{x}_1, \boldsymbol{x}_2) = R - x_1 \quad \text{at} \quad \boldsymbol{x}_2 = (R, 0)$$

and

$$S_1(\boldsymbol{x}_1, \boldsymbol{x}_2) = 3R - x_1 \quad \text{at} \quad \boldsymbol{x}_2 = (-R, 0).$$

Thus the approximation of Green's function $G(\boldsymbol{y}, \boldsymbol{x}, t)$,

$$G(\boldsymbol{y}, \boldsymbol{x}, t) \sim G_0(\boldsymbol{y}, \boldsymbol{x}, t) - G_1(\boldsymbol{y}, \boldsymbol{x}, t),$$

corresponding to the ray solutions $G_0(\boldsymbol{y}, \boldsymbol{x}, t)$ and

$$G_1(\boldsymbol{x}, \boldsymbol{y}, t) = Z_1(\boldsymbol{x}, \boldsymbol{y}, t)\, e^{-S_1^2(\boldsymbol{y}, \boldsymbol{x})/4t},$$

will satisfy the boundary condition at (x_1, R) if $Z_1(\boldsymbol{y}, \boldsymbol{x}, t)$ is chosen so that

$$Z_1(\boldsymbol{x}_1, \boldsymbol{x}_2, t) = \frac{1}{4\pi t} \quad \text{at} \quad \boldsymbol{x}_2 = (R, 0).$$

However, this approximation does not satisfy the boundary condition at $\boldsymbol{x}_2 = (-R, 0)$. The error in the boundary values at $\boldsymbol{x}_2 = (-R, 0)$ is

$$\begin{aligned} &G_0(\boldsymbol{x}_1, \boldsymbol{x}_2, t) - G_1(\boldsymbol{x}_1, \boldsymbol{x}_2, t) \\ &= \frac{1}{4\pi t} e^{-(R+x_1)^2/4t} - Z_1(\boldsymbol{x}_1, \boldsymbol{x}_2, t)\, e^{-4(R-x_1)^2/4t} \quad \text{at } \boldsymbol{x}_2 = (-R, 0) \end{aligned}$$

and is of the same order of magnitude as that of the leading-order approximation (6.1). To make up for the missed boundary condition, the further approximation

$$G(\boldsymbol{y}, \boldsymbol{x}, t) \sim G_0(\boldsymbol{y}, \boldsymbol{x}, t) - G_1(\boldsymbol{y}, \boldsymbol{x}, t) - G_2(\boldsymbol{y}, \boldsymbol{x}, t) \tag{6.2}$$

can be used, with

$$G_2(\boldsymbol{y}, \boldsymbol{x}, t) = Z_2(\boldsymbol{y}, \boldsymbol{x}, t)\, e^{-s_1^2(\boldsymbol{y}, \boldsymbol{x})/4t},$$

where on the x-axis

$$s_1(\boldsymbol{x}_1, \boldsymbol{x}_2) = 2R + x_1 + x_2$$

and

$$Z_2(\boldsymbol{x}_1, \boldsymbol{x}_2, t) = \frac{1}{4\pi t} \quad \text{at} \quad \boldsymbol{x}_2 = (-R, 0).$$

This eikonal corresponds to rays with two reflections in the boundary. The approximation (6.2) decreases the error in the boundary condition at $\boldsymbol{x}_2 = (-R, 0)$ to

$$-Z_1(\boldsymbol{x}_1, \boldsymbol{x}_2, t)\, e^{-4(R-x_1)^2/4t}$$

but misses the boundary condition at $\boldsymbol{x}_2 = (R, 0)$ with error

$$\begin{aligned} &G_0(\boldsymbol{x}_1, \boldsymbol{x}_2, t) - G_1(\boldsymbol{x}_1, \boldsymbol{x}_2, t) - G_2(\boldsymbol{x}_1, \boldsymbol{x}_2, t) \\ &= -Z_2(\boldsymbol{x}_1, \boldsymbol{x}_2, t)\, e^{-(3R+x_1)^2/4t} \quad \text{at} \quad \boldsymbol{x}_2 = (R, 0). \end{aligned}$$

This process gives successive approximations to Green's function with errors that decrease at transcendental rather than algebraic rates.

The approximation to the trace produced by $G_0(\boldsymbol{y}, \boldsymbol{x}, t)$ is the first algebraic term in the expansion (1.8). The contributions of the terms $-G_1(\boldsymbol{y}, \boldsymbol{x}, t)$ and $-G_2(\boldsymbol{y}, \boldsymbol{x}, t)$ in the approximation (6.2) of terms that are $O(\sqrt{t}e^{-R^2/t})$ are identical but with opposite signs, and thus they cancel each other. The second term contributes a negative term that is $O(\sqrt{t}e^{-4R^2/t})$. The term $O(\sqrt{t}e^{-R^2/t})$ for small t corresponds to $O(\frac{1}{s}e^{-2R\sqrt{s}})$ for large positive s in the Laplace plane. The number $2R$ is the length of the periodic orbit of a billiard ball bouncing inside a circle with the center removed, that is, inside the domain $\Omega - \Gamma$, where the set of critical points Γ consists of the center. Similarly, the term $O(\sqrt{t}e^{-4R^2/t})$ for small t corresponds to $O(\frac{1}{s}e^{-4R\sqrt{s}})$ for large positive s in the Laplace plane. The number $4R$ is the length of the minimal periodic orbit of a billiard ball bouncing inside a disk. We conclude that the conjecture of [10] should be supplemented with the orbit of length $2R$.

If Ω is an annulus between two concentric circles, of radii a and b, respectively, $(a > b)$, the two connected components of the boundary are the two circles, and there are no critical points in the domain relative to either one of them. In this case $\delta_1 = (a - b)$.

If Ω is the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} < 1$$

with $a > b$, the locus of critical points relative to the boundary is the segment

$$\Gamma = \left[-\frac{a^2 - b^2}{a}, \frac{a^2 - b^2}{a} \right]$$

on the x-axis. The segment Γ is the short diagonal of the evolute of the ellipse (the asteroid $(ax)^{2/3} + (by)^{2/3} = (a^2 - b^2)^{2/3}$). For the ellipse there are exponents in (1.15) which are $\delta_1 = 2b$ and its multiples and $\delta_2 = 2a$ and its multiples, as well as extremal periodic orbits with any number of reflections in the boundary.

Finally, we observe that if the boundary is reflecting (i.e., a homogeneous Neumann boundary condition), the exponential decay rate of the transcendental terms in the expansion of the trace is the same as in the case of absorbing boundary (homogeneous Dirichlet boundary condition). In this case the second term in the expansion (1.15) changes sign.

Obviously, rays that are reflected from the boundary more than once also give rise to ray solutions. The number of ray solutions needed in the expansion (3.7) is determined by the required degree of asymptotic approximation of the boundary conditions. If only a finite sum of ray solutions satisfies the boundary conditions, the sum (3.7) is finite. Otherwise, additional ray solutions improve the degree of approximation of the boundary conditions, as described in the one-dimensional ray expansion in section 2.

The derivation of the algorithm described in section 5 suggests further conjectures of the type (1.9) [10] concerning the coefficients of the power series $P_n(x)$ in the asymptotic series (1.15). They should relate the rate of growth of the coefficients of $P_1(x)$ to δ_2, and so on. This will make the evaluation of δ_n possible for $n > 1$, as above.

Finally, the asymptotic convergence of the ray expansion follows from the maximum principle for the heat equation in a straightforward manner.

REFERENCES

[1] P. T. Callaghan, *Principles of Nuclear Magnetic Resonance Microscopy*, Oxford University Press, New York, 1991.

[2] M. Kac, *Can one hear the shape of a drum?*, Amer. Math. Monthly, 73 (1966), pp. 1–23.

[3] K. Stewartson and R. T. Waechter, *On hearing the shape of a drum: Further results*, Math. Proc. Cambridge Philos. Soc., 69 (1971), pp. 353–363.

[4] P. Greiner, *An asymptotic expansion for the heat equation*, Arch. Rational Mech. Anal., 41 (1971), pp. 163–218.

[5] Y. Colin de Verdière, *Spectre du Laplacien et longueurs des géodésiques périodiques* I, Compositio Math., 27 (1973), pp. 83–106.

[6] Y. Colin de Verdière, *Spectre du Laplacien et longueurs des géodésiques périodiques* II, Compositio Math., 27 (1973), pp. 159–184.

[7] R. Balian and C. Bloch, *Distribution of eigenfrequencies for the wave equation in a finite domain*: III. *Eigenfrequency density oscillations*, Ann. Physics, 69 (1972), pp. 76–160.

[8] M. C. Gutzwiller, *Chaos in Classical and Quantum Mechanics*, Springer, New York, 1990.

[9] V. Guillemin and R. Melrose, *The Poisson summation formula for manifolds with boundary*, Adv. in Math., 32 (1979), pp. 204–232.

[10] M. V. Berry and C. J. Howls, *High orders of the Weyl expansion for quantum billiards: Resurgence of periodic orbits, and the Stokes phenomenon*, Proc. Roy. Soc. London Ser. A, 447 (1994), pp. 527–555.

[11] S. Zelditch, *Spectral determination of analytic bi-axisymmetric plane domains*, Geom. Funct. Anal., 10 (2000), pp. 628–677.
[12] Z. Schuss, *Theory and Applications of Stochastic Differential Equations*, Wiley Series in Probability and Statistics, John Wiley and Sons, New York, 1980.
[13] R. E. Meyer, *Exponential asymptotics*, SIAM Rev., 22 (1980), pp. 213–224.
[14] J. Cohen and R. Lewis, *Ray method for the asymptotic solution of the diffusion equation*, J. Inst. Math. Appl., 3 (1967), pp. 266–290.
[15] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Wiley, New York, 1989.
[16] J. K. Cohen, F. Hagin, and J. B. Keller, *Short time asymptotic expansion of solutions of parabolic equations*, J. Math. Anal. Appl., 38 (1972), pp. 82–91.
[17] C. Tier and J. B. Keller, *Asymptotic analysis of diffusion equations in population genetics*, SIAM J. Appl. Math., 34 (1978), pp. 549–576.
[18] B. D. Seckler and J. B. Keller, *Geometrical theory of diffraction in inhomogeneous media,* J. Acoust. Soc. Amer., 31 (1959), pp. 192–205.
[19] J. B. Keller, *Geometrical theory of diffraction*, J. Opt. Soc. Amer., 52 (1962), pp. 116–130.
[20] R. M. Lewis and J. B. Keller, *Asymptotic methods for partial differential equations: The reduced wave equation and Maxwell's equations*, in Surveys in Applied Mathematics, Vol. 1, J. B. Keller, G. Papanicolau, and D. McLaughlin, eds., Plenum, New York, 1995, pp. 1–82.

SIAM J. APPL. MATH.
Vol. 66, No. 1, pp. 361–364

ERRATUM: SINGULARITY FORMATION IN CHEMOTAXIS—A CONJECTURE OF NAGAI*

MATTHEW A. HALVERSON†, HOWARD A. LEVINE†, AND JOANNA RENCŁAWOWICZ‡

Abstract. In [H. A. Levine and J. Rencławowicz, *SIAM J. Appl. Math.*, 65 (2004), pp. 336–360] we considered the problem $u_t = u_{xx} - (uv_x)_x, v_t = u - av$ on the interval $I = [0, 1]$, where $u_x, v_x = 0$ at the end points, $u(x, 0), v(x, 0)$ are prescribed, and $a > 0$. (It was claimed in that article that there were solutions that blow up in finite time in every neighborhood of the spatially homogeneous steady state $(u, v) = (\mu, \mu/a)$ if $\mu > a$.) Here we correct an estimate and reduce Nagai's conjecture to the following statement. Let $\sigma = a/(\mu - a), \rho_1 = 1$. If $\lim_{n\to+\infty} \rho_n$ exists, where for $n \geq 2$, $\rho_n^n \equiv 1/(n-1)\sum_{j=1}^{n-1}(1 + \sigma/j)\rho_j^j \rho_{n-j}^{n-j}$, then the blow up assertion holds.

Key words. chemotaxis, finite time singularity formation, Keller–Segel model

AMS subject classifications. 35K55, 92C17

DOI. 10.1137/050631550

1. Introduction. In [1] we studied the system $u_t = u_{xx} - (uv_x)_x, v_t = u - av$ on the interval $I = [0, 1]$, where $u_x, v_x = 0$ at the end points, $u(x, 0), v(x, 0)$, are prescribed, and $a > 0$. Nagai and Nakaki [2] showed that there are solutions that are unbounded in finite *or* in infinite time.[1] We claimed that there were initial conditions for which solutions failed to exist for all time. In our proof we used a differential inequality, the derivation of which was unfortunately flawed. We correct this and make more precise the statement proved in [1].

2. Approximate solution. The notation of [1] is in force here. Because system $u_t = u_{xx} - (uv_x)_x, v_t = u - av$ is autonomous, we can assume the initial values are prescribed at $t = 0$ and that the blow up time, when it exists, is positive. As in [1], define, for any sequence $z(t) = \{z_n(t)\}_{n=1}^{\infty}$, $\mathcal{G}_n(z, z') = (1/2)C^2 n\{(\mathcal{M}z * z')_n + n\frac{a}{2}(z*z)_n\}$ and $\mathcal{H}_n(z, z') = (1/2)C^2 n\{[(T_n\mathcal{M}z, z') - (\mathcal{M}z, T_n z')] + an(z, T_n z)\}$, where $\mathcal{M}z(t) = \{nz_n(t)\}_{n=1}^{\infty}$ and $T_k z(t) = \{z_{n+k}(t)\}_{n=1}^{\infty}$. Here $|z| = \{|z_n|\}_{n=1}^{\infty}$ and $(z*w)_n = \sum_{k=1}^{n-1} z_k w_{n-k}$. (The sum is zero if $n = 1$.)

The infinite system of ordinary differential equations for the cosine coefficients $h(t) = \{h_n(t)\}_{n=1}^{\infty}$ is[2]

$$\mathfrak{L}_n h_n \equiv h_n'' + (C^2n^2 + a)h_n' - (\mu - a)C^2n^2h_n = \mathcal{G}_n(h, h') + \mathcal{H}_n(h, h').$$

The infinite system of ordinary differential equations satisfied by the cosine coefficients for the approximate problem, $g(t) = \{g_n(t)\}_{n=1}^{\infty}$, satisfies $\mathfrak{L}_n g_n = \mathcal{G}_n(g, g')$. The

*Received by the editors May 13, 2005; accepted for publication May 24, 2005; published electronically November 15, 2005.
http://www.siam.org/journals/siap/66-1/63155.html

†Department of Mathematics, Iowa State University, Ames, IA 50011 (mhalver@iastate.edu, halevine@iastate.edu).

‡Institute of Mathematics, Polish Academy of Sciences, Śniadeckich 8, 00-956, Warsaw, Poland (jr@impan.gov.pl).

[1]The Nagai conjecture states that if $\mu > a$, there are spatially nonhomogeneous solutions beginning in every small neighborhood of $(\mu, \mu/a)$ which cannot exist for all time.

[2]The spatially homogeneous solution is given by $V(t) = \mu/a + (v_0 - \mu/a)\exp(-at)$, $U(t) = \mu$. One sets $\psi(x, t) = v(x, t) - V(t)$, $u(x, t) = \mu + \psi_t + a\psi$. Then $h(t)$ is the sequence of cosine coefficients for $\psi(x, t)$.

particular sequence $g(t) \equiv \{g_n(t) = a_n e^{n\lambda t}\}_{n=1}^{\infty}$ satisfies this system for $a_1 > 0$, and for $n \geq 2$ and any integer $M > 0$ with $C = 2\pi M$, $\mu > a$ if

$$2\lambda[n - a/(4\pi^2 M^2)]a_n = \frac{1}{n-1}\sum_{k=1}^{n-1}[\lambda(n-k)k + ak]a_k a_{n-k}, \tag{2.1}$$

where λ is the positive root of $\lambda^2 + (4\pi^2 M^2 + a)\lambda - (\mu - a)4\pi^2 M^2 = 0$. There are positive constants a, b, ϵ, δ with $a\epsilon^n \leq na_n \leq b\delta^n$ for all positive integers [1]. From this, it follows that $\liminf_{n\to+\infty}[(-\ln na_n)/(n\lambda)] \equiv \underline{T_b}$ and $\limsup_{n\to+\infty}[(-\ln na_n)/(n\lambda)] \equiv \overline{T_b}$ are finite. Hence there is a subsequence $\{a_{n_k}\}_{k=1}^{\infty}$ such that $\lim_{k\to+\infty}[(-\ln n_k a_{n_k})/(n_k\lambda)] \equiv \underline{T_b}$. For this sequence, $\lim_{k\to+\infty} n_k a_{n_k} \exp(n_k\lambda\underline{T_b}) = 1$. Set $a_n = (A_n/n)\exp(-n\lambda\underline{T_b})$. On the subsequence, $A_{n_k} \to 1$ and

$$\lim_{t\uparrow\underline{T_b}} \sum_{k=1}^{\infty} A_{n_k} e^{-n_k\lambda(\underline{T_b}-t)} = +\infty \text{ and } \lim_{t\uparrow\underline{T_b}} \sum_{k=1}^{\infty} \frac{A_{n_k} e^{-n_k\lambda(\underline{T_b}-t)}}{{n_k}^{1+\delta}} < +\infty \tag{2.2}$$

(for any $\delta > 0$).

Now $\underline{T_b}$ must be the blow up time for the approximate solution $g(t)$ in the space $\ell_1^1(0,\underline{T_b}) \times \ell^1(0,\underline{T_b})$. (A sequence $\{a_n\}$ is in ℓ_1^1 if $\{na_n\}$ is in ℓ^1.) To see this, note that as long as t is in the existence interval,

$$\begin{aligned} \|\mathcal{M}g(t)\|_{\ell^1} + \|g'(t)\|_{\ell^1} &= \sum_{n=1}^{\infty} na_n(1+\lambda)e^{n\lambda t} \geq (1+\lambda)\sum_{k=1}^{\infty} n_k a_{n_k} e^{n_k\lambda t} \\ &= (1+\lambda)\sum_{k=1}^{\infty} A_k e^{-n_k\lambda(T_b - t)}. \end{aligned} \tag{2.3}$$

Consequently, from the first equation in (2.2), $g(\cdot)$ must blow up at some time, possibly earlier than $\underline{T_b}$. If $t < \underline{T_b}$, then $\liminf_{n\to+\infty}[(-\ln na_n)/(n\lambda)] \equiv \underline{T_b} > \underline{T_b} - \delta > t$ for some positive δ. Therefore, for sufficiently large N, $\sum_{n=N}^{\infty} na_n e^{n\lambda t} \leq \sum_{n=N}^{\infty} ne^{-n\lambda(\underline{T_b}-\delta-t)} < \infty$.

Set $\sigma = a/\lambda$. Let $\{\ln[na_n/(2a_1^n)]/n\}_{n=1}^{\infty} = \{\ln A_n/n\}_{n=1}^{\infty} \equiv \{p_n/n\}_{n=1}^{\infty}$. The p_n satisfy $p_1 = -\ln 2$, and for $n \geq 2$, $[1 - a/(4\pi^2 M^2 n)]e^{p_n} = \frac{1}{n-1}\sum_{j=1}^{n-1}(1 + \sigma/j)e^{(p_j + p_{n-j})}$. Then we have the following theorem.

THEOREM 1 (Nagai's conjecture). *Let $\lim_{n\to+\infty}\frac{p_n}{n}$ exist. The corresponding solution of the Nagai problem for which $h_n(0) = g_n(0)$ and $h_n'(0) = g_n'(0)$ for all n cannot both exist and be ℓ^1 regular on $[0,\infty)$. (A solution of the Nagai–Nakaki problem is ℓ^1 regular on an interval $I = [0, T_b)$ if it exists there and if $(\|\mathcal{M}h(s)\|_{\ell^1} + \|h'(s)\|_{\ell^1})$ is uniformly bounded on compact subsets I.)*

3. Estimate. Inequality (7.5) of [1] is incorrect. The correct form of the upper bound for the norm of $g - h \equiv w$, $\|\mathcal{M}w(t)\|_{\ell^1} + \|w'(t)\|_{\ell^1}$, is based on the following (infinite) system of ordinary differential equations:

$$\mathfrak{L}_n w_n = \mathcal{G}_n(h-g, h') + \mathcal{G}_n(g, h'-g') + \mathcal{H}_n(h,h') = \mathcal{G}_n(w,h') + \mathcal{G}_n(g,w') + \mathcal{H}_n(h,h') \tag{3.1}$$

and, for some $B > 0$ depending perhaps on τ but not on w, w', h, h', g, g', is given by

(3.2)

$$\|\mathcal{M}w(t)\|_{\ell^1} + \|w'(t)\|_{\ell^1} \le I(t) + J(t) + B\int_0^t \frac{(\|\mathcal{M}h(s)\|_{\ell^1} + \|h'(s)\|_{\ell^1})^2}{\sqrt{t-s}}\,ds + B\int_0^t \frac{(\|\mathcal{M}w(s)\|_{\ell^1} + \|w'(s)\|_{\ell^1})(\|\mathcal{M}h(s)\|_{\ell^1} + \|h'(s)\|_{\ell^1})}{\sqrt{t-s}}\,ds,$$

where

$$I(t)+J(t) \equiv \int_0^t \sum_{n=1}^{\infty} \mathcal{M}(|g'|*\mathcal{M}|w|)_n e^{-dn^2(t-s)}\,ds + \int_0^t \sum_{n=1}^{\infty} \mathcal{M}^2(|g|*|w|)_n e^{-dn^2(t-s)}\,ds,$$

and where $d > 0$ is the positive constant in [1, Lemma 1]. We have

$$\begin{aligned}
I(t) &= \int_0^t \sum_{n=1}^{\infty}\sum_{k=1}^{n-1} n(n-k)|g'_k||w_{n-k}|e^{-dn^2(t-s)}\,ds \\
&= \int_0^t \sum_{k=1}^{\infty}\sum_{n=1}^{\infty} n(n+k)|g'_k||w_n|e^{-d(n+k)^2(t-s)}\,ds \\
&\le \int_0^t \sum_{k=1}^{\infty} |g'_k|e^{-(d/2)k^2(t-s)} \left[\sum_{n=1}^{\infty}(n+k)e^{-(d/2)(n+k)^2(t-s)} n|w_n|\right] ds \\
&\le c\int_0^t \frac{\sum_{k=1}^{\infty}|g'_k|e^{-(d/2)k^2(t-s)}}{\sqrt{t-s}}\|\mathcal{M}w(s)\|_{\ell_1}\,ds \\
&\le c\int_0^t \sum_{k=1}^{\infty} A_k e^{-(d/2)k^2(t-s)-\lambda k(T_b-s)}\frac{\|\mathcal{M}w(s)\|_{\ell_1}}{\sqrt{t-s}}\,ds \\
&\le c\int_0^t \left\{\sum_{k=1}^{\infty} A_k e^{-[(d/2)k^2+k\lambda](t-s)}\right\}\frac{\|\mathcal{M}w(s)\|_{\ell_1}}{\sqrt{t-s}}\,ds \equiv c\int_0^t \mathcal{W}(t-s)\frac{\|\mathcal{M}w(s)\|_{\ell_1}}{\sqrt{t-s}}\,ds.
\end{aligned}$$

In the same manner,

$$\begin{aligned}
J(t) &= \int_0^t \sum_{n=1}^{\infty}\sum_{k=1}^{n-1} n^2|g_k||w_{n-k}|e^{-dn^2(t-s)}\,ds \\
&\le \int_0^t \sum_{k=1}^{\infty}\sum_{n=1}^{\infty}(n+k)^2|g_k||w_n|e^{-d(n+k)^2(t-s)}\,ds \\
&\le \int_0^t \sum_{k=1}^{\infty}|kg_k|e^{-(d/2)k^2(t-s)}\left[\sum_{n=1}^{\infty}\frac{(n+k)^2}{kn}e^{-(d/2)(n+k)^2(t-s)}n|w_n|\right] ds.
\end{aligned}$$

From the inequality $(k+l)/kl \le 2$,

$$J(t) \le c'\int_0^t \left\{\sum_{k=1}^{\infty} A_k e^{-[(d/2)k^2+k\lambda](t-s)}\right\}\frac{\|\mathcal{M}w(s)\|_{\ell_1}}{\sqrt{t-s}}\,ds \equiv c'\int_0^t \mathcal{W}(t-s)\frac{\|\mathcal{M}w(s)\|_{\ell_1}}{\sqrt{t-s}}\,ds.$$

In view of (2.2), $\lim_{t\uparrow T_b}\sum_{k=1}^{\infty} A_k k^{-2}e^{-k\lambda(T_b-t)} < +\infty$. Thus $\mathcal{W}(t)$ is in every $L^p[0,T_b]$ space for $1 \le p < \infty$. With $f(t) = \|\mathcal{M}w(t)\|_{\ell^1} + \|w'(t)\|_{\ell^1}$, we see $f(t) \le \int_0^t \mathcal{W}(t-s)f(s)/\sqrt{t-s}\,ds + \Phi(h(t))$. From Hölder's inequality with $1/p+1/r+1/q = 1$

and $1 < r < 2$, there is a constant $K > 0$ such that $f(t) \le K[\int_0^t f(s)^q\, ds]^{1/q} + \Phi(h(t))$ on $[0, T_b)$. From Gronwall's inequality, if h is global, $f(t)$ is bounded on $[0, T_b)$. From the first sum in (2.2) and the triangle inequality, this is impossible.

Other minor errors in [1]. Page 345, equation (5.1): Replace $k(w_k g_{n+k} + h_k w_{n+k})$ by $n(w_k g_{n+k} + h_k w_{n+k})$. Page 349, equation in line 13: $c\sqrt{t}$ should be replaced by $c \sup_{[0,T]} \sqrt{t}$.

REFERENCES

[1] H. A. Levine and J. Rencławowicz, *Singularity formation in chemotaxis—a conjecture of Nagai*, SIAM J. Appl. Math., 65 (2004), pp. 336–360.

[2] T. Nagai and T. Nakaki, *Stability of constant steady states and existence of unbounded solutions in time to a reaction-diffusion equation modelling chemotaxis*, Nonlinear Anal., 58 (2004), pp. 657–681.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 365–383

THE INVERSE CONDUCTIVITY PROBLEM WITH AN IMPERFECTLY KNOWN BOUNDARY*

VILLE KOLEHMAINEN†, MATTI LASSAS‡, AND PETRI OLA§

Abstract. We show how to eliminate the error caused by an incorrectly modeled boundary in electrical impedance tomography (EIT). In practical measurements, one usually lacks exact knowledge of the boundary. Because of this, the numerical reconstruction from the measured EIT data is done using a model domain that represents the best guess for the true domain. However, it has been noticed that an inaccurate model of the boundary causes severe errors for the reconstructions. We introduce a new algorithm to find a deformed image of the original isotropic conductivity based on the theory of Teichmüller spaces, and we implement it numerically.

Key words. inverse conductivity problem, electrical impedance tomography, unknown boundary, Teichmüller mapping

AMS subject classifications. 35J25, 30C75

DOI. 10.1137/040612737

1. Introduction.

We consider the electrical impedance tomography (EIT) problem, i.e., the determination of an unknown conductivity distribution inside a domain, for example the human thorax, from voltage and current measurements made on the boundary. Mathematically this is formulated as follows: Let $\Omega \subset \mathbb{R}^2$ be the measurement domain, and denote by $\gamma = (\gamma^{ij})$ the symmetric matrix describing the conductivity in Ω. We assume that the matrix has components in $L^\infty(\Omega)$ and that it is strictly positive definite; that is, for some $c > 0$ we have $\langle \xi, \gamma(x)\xi \rangle \geq c||\xi||^2$ for all $x \in \Omega$. The electrical potential u satisfies in Ω the equation

$$\nabla \cdot \gamma \nabla u = 0. \tag{1.1}$$

To uniquely fix the solution u it is enough to give its value on the boundary. Let this be $u|_{\partial\Omega} = f \in H^{1/2}(\partial\Omega)$, where $H^{1/2}(\partial\Omega)$ is the Sobolev space. Then (1.1) has a unique weak solution $u \in H^1(\Omega)$.

Our boundary data is the map that takes the voltage distribution f on the boundary for all f to the corresponding current flux through the boundary, $\nu \cdot \gamma \nabla u$, where ν is the exterior unit normal to Ω. Mathematically this amounts to the knowledge of the Dirichlet–Neumann map Λ corresponding to γ, i.e., the map taking the Dirichlet boundary values to the corresponding Neumann boundary values of the solution to (1.1),

$$\Lambda_\gamma : u|_{\partial\Omega} \mapsto \sum_{i,j=1}^{2} \nu_i \, \gamma^{ij} \frac{\partial u}{\partial x_j}\bigg|_{\partial\Omega}.$$

*Received by the editors August 3, 2004; accepted for publication (in revised form) April 14, 2005; published electronically November 22, 2005. This work was supported by the Academy of Finland projects 203985, 72434, and 102175.
http://www.siam.org/journals/siap/66-2/61273.html
†Department of Applied Physics, University of Kuopio, P. O. Box 1627, FIN-70211, Finland (Ville.Kolehmainen@uku.fi).
‡Institute of Mathematics, Helsinki University of Technology, P. O. Box 1100, FIN-02015, Finland (Matti.Lassas@hut.fi).
§Department of Mathematics and Statistics, Rolf Nevanlinna Institute, University of Helsinki, Helsinki P. O. Box 4, FIN-00014, Finland (Petri.Ola@rni.helsinki.fi).

This defines a bounded operator $\Lambda_\gamma : H^{1/2}(\partial\Omega) \to H^{-1/2}(\partial\Omega)$. The symmetric quadratic form corresponding to Λ_γ,

$$\Lambda_\gamma[h,h] := \int_{\partial\Omega} h\,\Lambda_\gamma h\,dS = \int_{\partial\Omega} \nabla u\cdot\gamma\nabla u\,dx, \tag{1.2}$$

equals in physical terms the power needed to maintain the potential h on $\partial\Omega$.

When γ is a scalar-valued function times an identity matrix, we say that the conductivity is *isotropic*. As usual, conductivities that may be matrix-valued are referred to as *anisotropic* conductivities. The EIT problem is to reconstruct γ from Λ_γ. The problem was originally proposed by Calderón [6] and then solved in dimensions three and higher for isotropic smooth conductivities in [18]. The two-dimensional case that is relevant to us was solved by Nachman [13] for isotropic conductivities assuming $\gamma \in W^{2,p}$, $p > 1$, and then finally for general L^∞-smooth isotropic conductivities by Astala and Päivärinta in a celebrated paper [4].

The conductivity equation is invariant under deformations of the domain Ω in the following sense. If F is a diffeomorphism taking Ω to some other domain $\widetilde{\Omega}$, then $u \circ F^{-1}$ will satisfy the conductivity equation in $\widetilde{\Omega}$ with conductivity

$$\widetilde{\gamma}(x) = \left.\frac{F'(y)\,\gamma(y)\,(F'(y))^t}{|\det F'(y)|}\right|_{y=F^{-1}(x)}, \tag{1.3}$$

where F' is the Jacobi matrix of map F and u is a solution of $\nabla\cdot\gamma\nabla u = 0$ in Ω. We say that $\widetilde{\gamma}$ is the push forward of γ by F and denote it by $\widetilde{\gamma} = F_*\gamma$. Note that all this is well defined for general matrix-valued γ. For us the starting point is the trivial observation that even if γ is isotropic, the deformed conductivity $\widetilde{\gamma}$ will not in general be isotropic. The boundary measurements are invariant: When $f : \partial\Omega \to \partial\widetilde{\Omega}$ is the restriction of $F : \Omega \to \widetilde{\Omega}$, we say that $\widetilde{\Lambda} = f_*\Lambda_\gamma$,

$$((f_*\Lambda_\gamma)h)(x) = (\Lambda_\gamma(h\circ f))(y)|_{y=f^{-1}(x)}, \quad h \in H^{1/2}(\partial\widetilde{\Omega}),$$

is the push forward of Λ_γ in f. As seen in [17], it turns out that $f_*\Lambda_\gamma = \Lambda_{F_*\gamma}$.

The fact that the anisotropic conductivity equation and the boundary measurements are invariant has the important consequence that the EIT problem with an anisotropic conductivity is not uniquely solvable, even though the isotropic problem is; see [17].

In practice, when solving the EIT problem in a given domain Ω, one typically seeks the isotropic conductivity that minimizes

$$||\Lambda_{meas} - \Lambda_\gamma||^2 + \alpha||\gamma||_X^2 \tag{1.4}$$

for γ defined in terms of some triangulation of Ω as, e.g., a piecewise constant function and $||\cdot||_X$ is some regularization norm [10]. Here Λ_{meas} is the measurement of the Dirichlet–Neumann map that contains measurement errors.

In practice, one of the key difficulties in solving the EIT problem is that the domain Ω may not be known accurately. It has been noticed that the use of a slightly incorrect model for Ω, i.e., a slightly incorrect model of the boundary, causes serious errors in reconstructions; see, e.g., [9, 1, 8]. As an example, consider the EIT measurements of pulmonary function from the human thorax. The measurement electrodes are attached on the skin of the patient around the thorax. In principle, an exact parameterization for the shape of the thorax could be obtained from other medical

imaging modalities such as magnetic resonance imaging (MRI) or computerized tomography (CT). However, in most cases this data is not available, and one has to resort to some approximate thorax model. Further, the shape of the thorax varies between breathing states, and it is also dependent on the orientation of the patient. Thus, the thorax geometry is known inaccurately even in best-case scenarios.

In this paper our aim is to propose a method to overcome the problem that the boundary and its parameterization are not exactly known. The set-up of the problem we consider is the following.

We want to recover the unknown conductivity γ in Ω from the measurements of the Dirichlet-to-Neumann map, and we assume a priori that γ is isotropic. We assume that $\partial\Omega$ and Λ_γ are not known. Instead, let Ω_m, called the model domain, be our best guess for the domain, and let $f_m : \partial\Omega \to \partial\Omega_m$ be a diffeomorphism modeling the approximate knowledge of the boundary. As the data for the inverse problem, we assume that we are given the boundary of the model domain, $\partial\Omega_m$, and the map $\Lambda_m := (f_m)_*\Lambda_\gamma$ on $\partial\Omega_m$. Note that we have simplified the problem by assuming that the only error in Λ_m comes from the imperfect knowledge of the boundary.

This set-up is motivated by the fact that the quadratic form corresponding to Λ_m,

$$\Lambda_m[h,h] = \int_{\partial\Omega_m} h\,\Lambda_m h\,dS = \int_{\partial\Omega} (h\circ f_m)\,\Lambda_\gamma(h\circ f_m)\,dS, \qquad h \in H^{1/2}(\partial\Omega_m),$$

represents the power needed to maintain the potential $h \circ f_m$ on $\partial\Omega$.

Since Λ_m usually does not correspond to any isotropic conductivity because of the deformation done when going from the original domain Ω to Ω_m, we obtain an erroneous solution γ when solving the minimization problem (1.4). This means that a systematic error in the domain model causes a systematic error in the reconstruction. In particular, local changes of the conductivity often give rise to nonlocalized changes in reconstructions due to the above systematic error. Thus the spatial resolution of details of reconstructions are often weak. This is clearly seen in practical measurements; see, e.g., [8].

We note that in solving the minimization problem (1.4) one could forget the assumption that γ is isotropic and find the minimizer in the set of anisotropic conductivities. However, the anisotropic inverse problem has a nonunique solution, and as the minimization problem is highly nonconvex, the minimization would be hard; as usual, forgetting existing a priori information makes the solution significantly worse.

To formulate our main results, let us define certain concepts. We start with the maximal anisotropy of an anisotropic conductivity.

DEFINITION 1.1. *Let $\gamma^{jk}(x)$ be an $L^\infty(\Omega)$-smooth matrix-valued conductivity in Ω, and let $\lambda_1(x)$ and $\lambda_2(x)$, $\lambda_1(x) \geq \lambda_2(x)$, be the eigenvalues of matrix $\gamma^{jk}(x)$. We define the maximal anisotropy of a conductivity to be $K(\gamma)$ given by*

$$K(\gamma) = \sup_{x\in\Omega} K(\gamma,x), \quad \textit{where} \quad K(\gamma,x) = \frac{\sqrt{L(x)}-1}{\sqrt{L(x)}+1}, \quad L(x) = \frac{\lambda_1(x)}{\lambda_2(x)}.$$

We call the function $K(\gamma,x)$ the anisotropy of γ at x. Here sup *denotes the essential supremum.*

Sometimes, to indicate the domain Ω, we denote $K(\gamma) = K_\Omega(\gamma)$. As a particularly important example needed later, let us consider the conductivity matrices of the form

$$\widehat{\gamma}(x) = \eta(x) R_{\theta(x)} \begin{pmatrix} \lambda^{1/2} & 0 \\ 0 & \lambda^{-1/2} \end{pmatrix} R_{\theta(x)}^{-1}, \tag{1.5}$$

where $\lambda \geq 1$ is a constant, $\eta(x) \in \mathbb{R}_+$ is a real-valued function, and $R_{\theta(x)}$ is a rotation matrix corresponding to angle $\theta(x)$, where

$$R_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

We denote such conductivities by $\widehat{\gamma} = \widehat{\gamma}_{\lambda,\theta,\eta}$. These conductivities have the anisotropy $K(\widehat{\gamma}, x) = c_\lambda = (\lambda^{1/2}-1)/(\lambda^{1/2}+1)$ at every point, and thus their maximal anisotropy is $K = c_\lambda$. We call such conductivities $\widehat{\gamma}$ *uniformly anisotropic conductivities.*

THEOREM 1.2. *Let Ω be a bounded simply connected $C^{1,\alpha}$-domain with $\alpha > 0$. Assume that $\gamma \in C^{0,\alpha}(\overline{\Omega})$ is a (possibly) anisotropic conductivity and Λ_γ its Dirichlet–Neumann map. Let Ω_m be a model of the domain satisfying the same regularity assumptions as Ω, and $f_m : \partial\Omega \to \partial\Omega_m$ be a $C^{1,\alpha}$-smooth diffeomorphism.*

Assume that we are given $\partial\Omega_m$ and $\Lambda_m = (f_m)_*\Lambda_\gamma$. Then we have the following:

1. *There is a unique anisotropic conductivity $\widehat{\gamma} \in L^\infty(\Omega_m, \mathbb{R}^{2\times 2})$ such that if γ_1 is an anisotropic conductivity in Ω_m for which $\Lambda_{\gamma_1} = \Lambda_m$, then $K(\gamma_1) \geq K(\widehat{\gamma})$.*
2. *Let $\lambda \geq 1$ be such that $K(\widehat{\gamma}) = (\lambda^{1/2}-1)/(\lambda^{1/2}+1)$. Then there are unique $\theta \in L^\infty(\Omega_m, S^1)$ and $\eta \in L^\infty(\Omega_m, \mathbb{R}_+)$ such that $\widehat{\gamma} = \widehat{\gamma}_{\lambda,\theta,\eta}$.*

Theorem 1.2 can be interpreted by saying that we can find a unique conductivity in Ω_m that is as close as possible to being isotropic. For an isotropic conductivity, the assumption on smoothness of conductivity can be relaxed, as can be seen from the following theorem, which is the main result of the paper.

THEOREM 1.3. *Let Ω, Ω_m, and f_m be as in Theorem* 1.2. *Let $\gamma \in L^\infty(\overline{\Omega})$ be an isotropic conductivity. Assume that we are given $\partial\Omega_m$ and $\Lambda_m = (f_m)_*\Lambda_\gamma$. Then results* 1 *and* 2 *of Theorem* 1.2 *are valid.*

Theorem 1.2 yields immediately the following algorithm for finding $\widehat{\gamma}$.

Remark 1. The conductivity $\widehat{\gamma} = \widehat{\gamma}_{\lambda,\eta,\theta}$ can obtained using the unique solution of the minimization problem

$$\min_{(\lambda,\theta,\eta)\in S} \lambda, \quad \text{where } S = \left\{ (\lambda,\theta,\eta) \in [1,\infty) \times L^\infty \times L^\infty \mid \Lambda_{\widehat{\gamma}(\lambda,\theta,\eta)} = \Lambda_m \right\}. \tag{1.6}$$

Later, in implementation of the algorithm we approximate the problem (1.6) with the regularized minimization problem

$$\min_{(\lambda,\theta,\eta)} ||\Lambda_{\widehat{\gamma}(\lambda,\theta,\eta)} - \Lambda_m||^2_{L(H^{1/2}(\partial\Omega_m), H^{-1/2}(\partial\Omega_m))} + \varepsilon_1 f(\lambda) + \varepsilon_2 ||\theta||^2_{H^1} + \varepsilon_2 ||\eta||^2_{H^1}, \tag{1.7}$$

where $f : [1,\infty) \to \mathbb{R}_+$ is a convex function that has its minimum near $\lambda = 1$ and $\lim_{t\to 1} f(t) = \lim_{t\to\infty} f(t) = \infty$ and $\varepsilon_1, \varepsilon_2, \varepsilon_3 > 0$.

The proof of Theorem 1.2 is based on the theory of quasi-conformal maps. There are several equivalent definitions for these maps, and we will present the one based on a partial differential equation (Beltrami equation) in section 2. However, the quasi-conformal maps also have a geometric definition. Indeed, they are generalizations of conformal maps that take infinitesimal disks at z to infinitesimal disks at $f(z)$, and the radii gets dilated by $|f'(z)|$. Analogously, a homeomorphic map is quasi-conformal on a domain Ω if infinitesimal disks at any $z \in \Omega$ get mapped to infinitesimal ellipsoids at $f(z)$. The ratio of the larger semiaxis to the smaller semiaxis is called the dilation of f at z, and the supremum of dilatations over Ω is the maximal dilation. This

dilatation of infinitesimal discs is in fact the reason why isotropic conductivities change to anisotropic ones in push forwards with quasi-conformal maps.

The crucial fact that we use in proving Theorem 1.2 is a result of Strebel [16] that, roughly speaking, says that among all quasi-conformal self-maps of the unit disk to itself with a given sufficiently smooth boundary value there is a unique one with the minimal maximal dilation. This will yield that corresponding to the given boundary modeling map $f_m : \partial\Omega \to \partial\Omega_m$ there is a unique map $F : \Omega \to \Omega_m$ having the minimal maximal dilation. We will show that this leads to the following result.

PROPOSITION 1.4. *Let Ω, Ω_m, f_m, and an isotropic conductivity γ satisfy the assumptions of Theorem* 1.3. *Then there is a unique map $F : \Omega \to \Omega_m$, depending only on $f : \partial\Omega \to \partial\Omega_m$, such that for the uniformly anisotropic conductivity $\widehat{\gamma}$ corresponding to γ in Theorem* 1.2 *we have*

$$\det(\widehat{\gamma}(x))^{1/2} = \gamma(F^{-1}(x)). \tag{1.8}$$

Proposition 1.4 can interpreted as saying that, solving the minimization problem (1.6), we can find the function $(\det \widehat{\gamma}(x))^{1/2}$ in Ω_m that represents a deformed image of original conductivity γ in the unknown domain Ω, and that the deformation depends only on the error made in modeling the boundary, not on the conductivity in Ω.

In particular, this turns out to be useful as local perturbations of conductivity remain local in reconstruction: If we consider one fixed boundary modeling map $f_m : \partial\Omega \to \partial\Omega_m$ but two isotropic conductivities γ_1 and $\gamma_2 = \gamma_1 + \sigma$ in Ω, then the reconstructions $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ obtained by Theorem 1.2 corresponding to γ_1 and γ_2 satisfy

$$\det(\widehat{\gamma}_2(x))^{1/2} - \det(\widehat{\gamma}_1(x))^{1/2} = \sigma(F^{-1}(x)).$$

The paper is organized as follows. In section 2 we show how to use isothermal coordinates to push forward an anisotropic conductivity to an isotropic one. There we pay close attention to the smoothness required from γ and Ω and introduce the necessary background from the theory of anisotropic inverse problems. We apply this in section 3 to prove our main results using the existence of a Teichmüller mapping. In section 4 we consider physically realistic measurements, i.e., the so-called complete electrode model. The numerical implementation for the complete electrode model is then described in the last sections.

2. Quasi-conformal maps and solvability of the inverse problem with anisotropic conductivity. It is a classical result that every Riemannian surface is locally conformal to a Euclidean plane: This corresponds to choosing the coordinate system to be *isothermal* [19, section 5.10]. Similarly, every anisotropic conductivity matrix can be transformed into an isotropic conductivity. We identify the plane $\mathbb{R}^2$ with the complex plane $\mathbb{C}$. We use the class $L_{\infty,\alpha}(\Omega)$ that consists of conductivities $\gamma = a\,\sigma$, where $a \in L^\infty(\Omega)$ satisfies $a(x) \geq c > 0$ and σ is a $C^{1,\alpha}(\overline{\Omega})$-smooth symmetric positive definite matrix; i.e., we allow arbitrary L^∞-smooth conformal transformations of $C^{0,\alpha}$-smooth background conductivities. We need the smoothness of the background conductivity above to guarantee the existence of the unique extremal conductivity, but this is a conformally invariant procedure, and hence the smoothness of the conformal factor a plays no role, as we will see below. Notice also that by taking $\sigma = id$, we have an L^∞ isotropic conductivity.

LEMMA 2.1. *Let Ω be a bounded simply connected $C^{1,\alpha}$-domain with $\alpha > 0$. Assume that $\gamma \in L_{\infty,\alpha}(\Omega)$ is an anisotropic conductivity. Then there is a $C^{1,\alpha}$-smooth diffeomorphism $F : \overline{\Omega} \to \widetilde{\overline{\Omega}}$, $\widetilde{\Omega} = F(\Omega) \subset \mathbb{C}$ such that*

$$F_*\gamma = \beta, \tag{2.1}$$

where $F_*\gamma$ is defined by (1.3)*, and β is the identity matrix multiplied by a L^∞-smooth scalar function. Moreover,*

$$\beta = (\det \gamma \circ F^{-1})^{1/2} I.$$

The proof of this result is well known, but as smoothness of F is crucial later, we give the proof for the convenience of the reader.

Proof. The equation (2.1) is a priori a nonlinear system for the derivatives of F. However, in two dimensions this equation completely linearizes and is equivalent to the *Beltrami equation*

$$\overline{\partial} F = \mu \partial F, \tag{2.2}$$

where the complex derivatives are $\partial = \frac{1}{2}(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y})$, $\overline{\partial} = \frac{1}{2}(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y})$ and *the Beltrami coefficient* $\mu = \mu_F(z)$, called also the *complex dilatation*, is given by

$$\mu = \frac{-\gamma_{11} + \gamma_{22} - 2i\gamma_{12}}{\gamma_{11} + \gamma_{22} + 2\sqrt{\gamma_{11}\gamma_{22} - \gamma_{12}^2}}. \tag{2.3}$$

The function μ is invariant in multiplication of the conductivity with a positive scalar function. Thus we see that $\mu \in C^{0,\alpha}(\overline{\Omega})$. The function μ has the crucial property that it is strictly less than one in modulus:

$$\sup_{z\in\Omega} |\mu(z)| < 1. \tag{2.4}$$

Let us extend the conductivity matrix γ (a priori defined only in Ω) to the whole plane to be the identity matrix outside Ω. Similarly, μ is extended outside Ω by zero.

Next we consider how to solve the Beltrami equation, and for this we consider it in the whole plane. In order to have a unique solution we fix the behavior of F at infinity. Thus, consider

$$\begin{aligned} &\overline{\partial} F(z) = \mu(z)\partial F(z) \quad \text{in } \mathbb{C}, \\ &F(z) = z + h(z), \\ &\lim_{z\to\infty} h(z) = 0, \end{aligned} \tag{2.5}$$

where μ is a compactly supported L^∞-function satisfying (2.4). This problem has unique solution $F \in L^p_\delta$ when p is close enough to 2 and $-2/p < \delta < 1 - 1/p$. For the proof of this, see, for example, [2] or [17]. The proof is based on the fact that (2.5) can be written as an integral equation

$$F(z) + \frac{1}{\pi}\int_{\mathbb{C}} \frac{\mu(\zeta)\partial F(\zeta)}{z-\zeta}\, da(\zeta) = z, \tag{2.6}$$

where $da(\zeta)$ is Euclidean area in $\mathbb{C}$ (or $\mathbb{R}^2$). As $||\mu||_\infty < 1$, it turns out that the left-hand side of (2.6) is of the form of the identity plus a contractive operator in Sobolev space $W^{1,p}(\Omega)$, with appropriate p, and thus (2.6) can be solved by an application of the Neumann-series argument.

Using interior Schauder estimates for (2.5), we see that if γ and thus μ are $C^{0,\alpha}$-smooth, the solution F has to be locally $C^{1,\alpha}$-smooth in $\mathbb{C}$, particularly in $\overline{\Omega}$. Using formula (1.3), we see that $F_*\gamma$ is $C^{0,\alpha}$-smooth in closure of $\widetilde{\Omega}$. □

In general, any solution $F : \Omega \to \widetilde{\Omega}$ to the Beltrami equation for μ which satisfies (2.4) and for which $F \in H^1(\Omega)$ is called *quasi-regular*. If a quasi-regular map $F : \Omega \to \widetilde{\Omega}$ is a homeomorphism, it is said to be *quasi-conformal*. The quasi-conformality can be defined also in geometrical terms; see [2, 11].

Next we recall the recent results for inverse problems for anisotropic conductivities γ. Let us consider a class of conductivities in Ω, given by

$$\Sigma(\gamma) = \{F_*\gamma \mid F : \Omega \to \Omega \text{ is a homeomorphism, } F, F^{-1} \in H^1(\Omega; \mathbb{C}), F|_{\partial\Omega} = I\};$$

that is, $\Sigma(\gamma)$ is the equivalence class of the conductivity γ in push forwards with boundary preserving diffeomorphisms. Then $\Lambda_\sigma = \Lambda_\gamma$ for all $\sigma \in \Sigma(\gamma)$. By [3], the converse is true; that is, if σ is a strictly positive definite L^∞-conductivity and $\Lambda_\sigma = \Lambda_\gamma$, then $\sigma \in \Sigma(\gamma)$. In other words, Λ_γ determines the equivalence class $\Sigma(\gamma)$. Note that diffeomorphism $F : \Omega \to \Omega$ such that $F \in H^1(\Omega; \mathbb{C})$ and $F|_{\partial\Omega} = I$ is quasi-conformal.

3. Proof of main results. We start by proving Theorems 1.2 and 1.3. To prove them at the same time, we consider a (possibly anisotropic) conductivity γ in class $L_{\infty,\alpha}(\Omega)$.

First we show that we can assume that Ω_m is the unit disc $\mathbb{D} \subset \mathbb{C}$. To prove this, let $f_m : \partial\Omega \to \partial\Omega_m$ be the boundary modeling map.

Our first observation is that as Ω_m is a simply connected domain, it follows from the Riemann mapping theorem that it can be mapped to unit disc $\mathbb{D}$ conformally. Moreover, as Ω_m is a $C^{1,\alpha}$-smooth domain, it follows from the Kellog–Warschawski theorem [14, Theorem 3.6] that the Riemann map can be chosen to be be a $C^{1,\alpha}$-diffeomorphism $F_0 : \overline{\Omega}_m \to \overline{\mathbb{D}}$ such that $F_0 : \Omega_m \to \mathbb{D}$ is conformal. Thus, if σ is some conductivity in class $L_{\infty,\alpha}(\Omega_m)$, we have that $\sigma_0 = (F_0)_*\sigma$ is a conductivity in class $L_{\infty,\alpha}(\mathbb{D})$.

Second, we observe that the uniformly anisotropic conductivity $\widehat{\gamma}_{\lambda,\theta,\eta}$ of the form (1.5) in Ω_m changes under $(F_0)_*$ to a uniformly anisotropic conductivity $(F_0)_*\widehat{\gamma}_{\lambda,\theta,\eta} = \widehat{\gamma}_{\lambda,\theta_0,\eta_0}$ in $\mathbb{D}$ such that $\eta_0 = \eta \circ F_0^{-1}$.

Third, we see that as $F_0 : \Omega_m \to \mathbb{D}$ is conformal, the maximal anisotropy of $(F_0)_*\sigma$ and σ satisfies

$$K_{\mathbb{D}}((F_0)_*\sigma) = K_\Omega(\sigma);$$

that is, the maximal anisotropy is preserved in conformal transformations for any σ.

Fourth, if $f_0 = F_0|_{\partial\Omega_m}$, then $\Lambda_{\sigma_0} = (f_0)_*\Lambda_\sigma$. Also, we see that our data is invariant in the change of the model domain in the sense that $(\widetilde{f}_m)_*\Lambda_\gamma = (\widetilde{f}_0)_*((\widetilde{f}_m)_*\Lambda_\gamma)$, where $\widetilde{f}_m = f_0 \circ f_m : \partial\Omega \to \partial\mathbb{D}$.

These four observations yield that it is enough to prove the assertion in the case when $\Omega_m = \mathbb{D}$. Indeed, changing Ω_m to $\mathbb{D}$ with F_0 keeps the boundary measurements, the smoothness of objects, the maximal anisotropy, as well as class of uniformly anisotropic conductivities invariant. More precisely, we can replace the boundary modeling map f_m by the map $\widetilde{f}_m = f_0 \circ f_m$. Figure 3.1 will help clarify the argument that follows.

Thus, let us return to proving Theorem 1.2 in the case when $\Omega_m = \mathbb{D}$. Let $f_m : \partial\Omega \to \partial\mathbb{D}$ be the boundary modeling map that is a $C^{1,\alpha}$-smooth diffeomorphism and where γ is a conductivity in class $L_{\infty,\alpha}(\Omega)$, with Dirichlet–Neumann map Λ_γ.

Let F_m be some $C^{1,\alpha}(\overline{\Omega})$-diffeomorphism $F_m : \overline{\Omega} \to \overline{\mathbb{D}}$ such that $F_m|_{\partial\Omega} = f_m$. There are many ways to construct such a map, and for the convenience of the reader

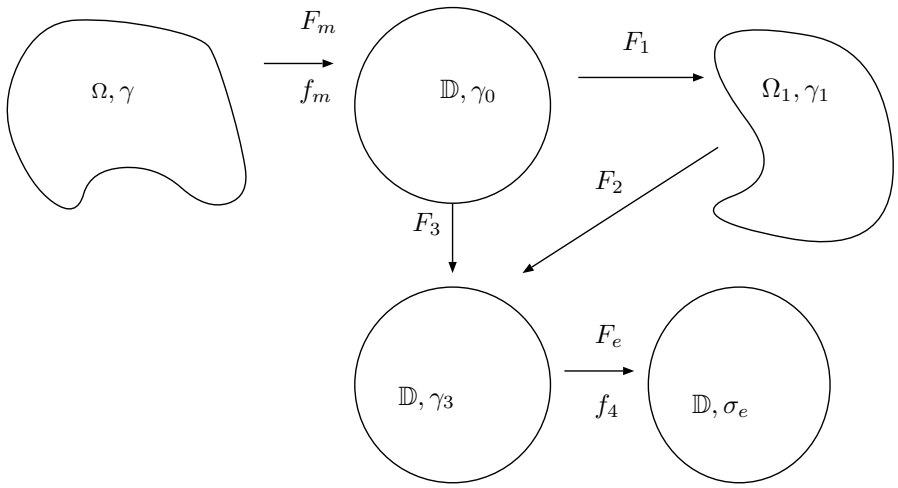


FIG. 3.1.

we present one simple way. Let $G : \Omega \to \mathbb{D}$ be a Riemann map. By [14, Theorem 3.7], G has $C^{1,\alpha}$-extension $G : \overline{\Omega} \to \overline{\mathbb{D}}$. Let $\phi = f_m \circ G^{-1} : \partial\mathbb{D} \to \partial\mathbb{D}$ and $\Phi(z) = |z| \exp(i(|z|^3 \arg(\phi(z/|z|)) + (1 - |z|^3)\arg(z/|z|)))$ be a $C^{1,\alpha}$-diffeomorphism $\mathbb{D} \to \mathbb{D}$ satisfying $\Phi|_{\partial\mathbb{D}} = \phi$. Then F_m can be chosen to be the map $\Phi \circ G$.

Let $\gamma_0 = (F_m)_*\gamma$ be an anisotropic conductivity in $\mathbb{D}$. By Lemma 2.1, there is a $C^{1,\alpha}$-diffeomorphism $F_1 : \overline{\Omega} \to \overline{\Omega}_1$ such that the conductivity $\gamma_1 = (F_1)_*\gamma_0$ is an isotropic $L^\infty(\Omega_1)$-conductivity.

As Ω_1 is a simply connected $C^{1,\alpha}$-smooth domain, by the Kellog–Warschawski theorem cited above there is a conformal map $F_2 : \Omega_1 \to \mathbb{D}$ such that $F_2 : \overline{\Omega}_1 \to \overline{\mathbb{D}}$ is a $C^{1,\alpha}$-diffeomorphism. Let $F_3 = F_2 \circ F_1 : \mathbb{D} \to \mathbb{D}$ and $f_3 = F_3|_{\partial D}$. Note that $(F_3)_*\gamma_0$ is the isotropic conductivity in $\mathbb{D}$, as F_2 is conformal.

The boundary values of quasi-conformal maps $\mathbb{D} \to \mathbb{D}$ are characterized as being the quasi-symmetric maps, that is, homeomorphic maps $f : \partial\mathbb{D} \to \partial\mathbb{D}$ such that $\theta(u) = \arg f(e^{iu})$ satisfy

$$k^{-1} \leq \frac{\theta(u+v) - \theta(u)}{\theta(u) - \theta(u-v)} \leq k \quad \text{for all} \quad u, v \in \mathbb{R}, \tag{3.1}$$

with some $k > 0$; see [11].

Let us consider next the map $f_4 = f_3^{-1} : \partial\mathbb{D} \to \partial\mathbb{D}$. Since f_3 and f_3^{-1} are $C^{1,\alpha}$-smooth, we see that f_4 satisfies

$$\lim_{v \to 0} \frac{\theta(u+v) - \theta(u)}{\theta(u) - \theta(u-v)} = 1 \quad \text{uniformly in } u \in \mathbb{R} \tag{3.2}$$

and is in particular quasi-symmetric. Thus f_4 is the boundary value of at least one quasi-conformal map. What is more, since f_4 satisfies condition (3.2), it follows from the results of Strebel [16] that among all quasi-conformal maps having f_4 as a boundary value there is a unique *extremal* map F_e in the sense that the L^∞-norm of the complex dilatation μ_{F_e} is minimal. More precisely, if $F : \mathbb{D} \to \mathbb{D}$ is a quasi-conformal map such that $F|_{\partial\mathbb{D}} = f_4$, then its Beltrami coefficient satisfies $||\mu_F||_{L^\infty} \geq ||\mu_{F_e}||_{L^\infty}$, and the equality holds only if $F = F_e$. Furthermore, the extremal F_e is a Teichmüller

mapping, i.e., its complex dilatation μ_{F_e} is of the form

$$\mu_{F_e}(z) = ||\mu_{F_e}||_\infty \frac{\overline{\phi(z)}}{|\phi(z)|}, \tag{3.3}$$

where $\phi : \mathbb{D} \to \mathbb{C}$ is holomorphic in $\mathbb{D}$, and thus has a discrete set of zeros. Note that F_e need not even be Lipschitz smooth near zeros of ϕ. Readers should also note that a certain assumption on the regularity of the boundary value f_m is necessary for the existence of extremal maps. This will be discussed after finishing the proof.

Let us now consider how a quasi-conformal map $F : \mathbb{D} \to \mathbb{D}$ with complex dilatation μ_F changes maximal anisotropy of conductivities. When σ is an isotropic conductivity in $\mathbb{D}$, that is, $K(\sigma) = 0$, one sees that for the anisotropic conductivity $\widetilde{\sigma} = F_*\sigma$ we have

$$K(x, \widetilde{\sigma}) = \mu_F(F^{-1}(x)) \quad \text{for } x \in \overline{\mathbb{D}},$$

and hence the maximal anisotropy satisfies $K(\widetilde{\sigma}) = ||\mu_F||_{L^\infty}$.

Let now $\gamma_3 = (F_3)_*\gamma_0$ be an isotropic conductivity in $\mathbb{D}$ and let $\sigma_e = (F_e)_*\gamma_3$ be an anisotropic conductivity in $\mathbb{D}$. Here, $F_e \circ F_3 \circ f_m|_{\partial\Omega} = f_m$. In particular, the above shows that

$$(f_m)_*\Lambda_\gamma = (f_4 \circ f_3 \circ f_m)_*\Lambda_\gamma = (f_4 \circ f_3)_*\Lambda_{\gamma_0} = (f_4)_*\Lambda_{\gamma_3} = \Lambda_{\sigma_e}.$$

In particular, this implies that the inverse problem of finding conductivities σ in $\mathbb{D}$ such that $(f_m)_*\Lambda_\gamma = \Lambda_\sigma$ has a solution $\sigma = \sigma_e$. By section 2, the knowledge of the boundary $\partial\Omega_m = \partial\mathbb{D}$ and the map $(f_m)_*\Lambda_\gamma$ determines the class $\Sigma(\sigma_e)$ of conductivities in $\mathbb{D}$. Now we can write the class $\Sigma(\sigma_e)$ also as

$$\Sigma(\sigma_e) = \{F_*\gamma_3 \ : \ F : \mathbb{D} \to \mathbb{D} \text{ is a homeomorphism, } F, F^{-1} \in H^1(\Omega; \mathbb{C}), F|_{\partial\mathbb{D}} = f_4\}.$$

Since

$$K(F_*\gamma_3) = ||\mu_F||_{L^\infty(\mathbb{D})},$$

we see that the conductivity $\sigma_e = (F_e)_*\gamma_3$ corresponding to the extremal map F_e is the unique conductivity σ in the class $\Sigma(\sigma_e)$ that has the smallest possible value of $K(\sigma)$.

Finally, since $|\mu_{F_e}(z)| = c_0$ is a constant function of $z \in \mathbb{D}$, and $\sigma_e = (F_e)_*\gamma_3$ with isotropic γ_3, we see that the ratio of the eigenvalues of the conductivity matrix $\sigma_e(z)$ is constant for $z \in \mathbb{D}$. Thus σ_e has the form $\sigma_e = \widehat{\gamma}_{\lambda,\theta,\eta}$ with $c_0 = (1-\lambda)/(1+\lambda)$, $\eta = \gamma_3 \circ (F_e)^{-1}$, and some θ. This proves Theorems 1.2 and 1.3 □

Next we prove Proposition 1.4.

Proof of Proposition 1.4. Consider isotropic conductivities γ_1 and γ_2 in Ω. In what follows, we use the notation of the proof of Theorem 1.2. By definition, f_m determines a map F_m. The construction of the map F_1 is based on the Beltrami coefficient of the conductivity. Clearly, the Beltrami coefficients for the conductivities $(F_m)_*\gamma_1$ and $(F_m)_*\gamma_2$ coincide, and thus F_1 and Ω_1 can be taken to be the same for both γ_1 and γ_2. The maps F_2, F_3, and F_e are constructed by using $\partial\Omega_1$ and F_1, and thus they coincide for γ_1 and γ_2. Since in general $\det(F_*\gamma)(x) = \det(\gamma(F^{-1}(x)))$, this proves Proposition 1.4. □

From our assumptions on $\partial\Omega$ and f_m (i.e., that they are in $C^{1,\alpha}$) it follows that the unique extremal exists. For general continuous f_m there are counterexamples,

for instance the so-called Strebel's chimney; see, e.g., [11]. For the sharpest known conditions giving the existence of the unique extremal, see [5].

We note also that if in formula (3.3) the function ϕ has zeros in Ω, then μ_{F_e} has a singularity of type $\overline{(z-z_0)^j}/(z-z_0)^j$, and this could affect the behavior of the reconstruction algorithm we propose in a way to be explained later. However, in all the numerical examples we have tested these difficulties do not appear, probably since our deformations are relatively small.

4. Electrode model. In the numerical simulations below we have used the so-called complete electrode model [15], which is a certain finite-dimensional approximation of Dirichlet-to-Neumann map. This model is chosen because it is an accurate model for the measurements made in practice. As noted before, in experimental measurements one places the measurement electrodes on the boundary, e.g., the skin of the patient, without knowing the exact parameterization of the boundary. Thus this model is a paradigm of the case when the boundary is unknown.

To define the electrode model, let $e_j \subset \partial\Omega$, $j = 1, \dots, J$, be disjoint open paths modelling the electrodes that are used for the measurements. Let u solve the equations

$$\nabla\cdot\gamma\nabla v = 0 \quad \text{in } \Omega, \tag{4.1}$$

$$z_j \nu\cdot\gamma\nabla v + v|_{e_j} = V_j, \tag{4.2}$$

$$\nu\cdot\gamma\nabla v|_{\partial\Omega\setminus\cup_{j=1}^J e_j} = 0, \tag{4.3}$$

where V_j are constants representing electric potentials on electrode e_j. This models the case where electrodes e_j having potentials V_j are attached to the boundary, z_j is the contact impedance between electrode e_j and the body surface, and the normal current outside the electrodes vanish. By [15], (4.1)–(4.3) has a solution $u \in H^1(\Omega)$. The measurements in this model are the currents observed on the electrodes, given by

$$I_j = \frac{1}{|e_j|}\int_{e_j} \nu\cdot\gamma\nabla v(x)\, ds(x), \quad j = 1, \dots, J.$$

Thus the electrode measurements are given by map $E : \mathbb{R}^J \to \mathbb{R}^J$, $E(V_1, \dots, V_J) = (I_1, \dots, I_J)$. We say that E is the electrode measurement matrix for $(\partial\Omega, \gamma, e_1, \dots, e_J, z_1, \dots, z_J)$. Let Ω and $\widetilde{\Omega}$ be $C^{1,\alpha}$-smooth domains. We say that $f : \partial\Omega \to \partial\widetilde{\Omega}$ is length preserving on $\cup_{j=1}^J e_j$ if $||Df(\tau)|| = 1$ for $x \in \cup_{j=1}^J e_j$, where τ is the unit tangent vector of $\partial\Omega$.

PROPOSITION 4.1. *Let Ω and $\widetilde{\Omega}$ be $C^{1,\alpha}$-smooth domains and $F : \overline{\Omega} \to \overline{\widetilde{\Omega}}$ be a $C^{1,\alpha}$-diffeomorphism, $e_j \subset \partial\Omega$ be disjoint open sets, and γ be a conductivity on Ω. Let $f = F|_{\partial\Omega}$, $\widetilde{e}_j = f(e_j) \subset \partial\widetilde{\Omega}$, and $\widetilde{\gamma} = (F)_*\gamma$. Assume that f is length preserving on $\cup_{j=1}^J e_j$. Then the electrode measurement matrices E for $(\partial\Omega, \gamma, e_1, \dots, e_J, z_1, \dots, z_J)$ and $\widetilde{E}$ for $(\partial\widetilde{\Omega}, \widetilde{\gamma}, \widetilde{e}_1, \dots, \widetilde{e}_J, z_1, \dots, z_J)$ coincide.*

Proof. We start with an invariant formulation of electrode measurements E. For this, let R be the Robin-to-Neumann map given by $Rf = \nu\cdot\gamma\nabla u|_{\partial\Omega}$, where u is the solution of

$$\begin{aligned} &\nabla\cdot\gamma\nabla u = 0 \quad \text{in } \Omega, \\ &z\nu\cdot\gamma\nabla v + \eta v|_{\partial\Omega} = h, \end{aligned} \tag{4.4}$$

where $z = z(x)$ is a $C^\infty(\partial\Omega)$ function such that $z|_{e_j} = z_j$ and $\eta = \sum_{j=1}^J \chi_{e_j}(x)$, where χ_{e_j} is the characteristic function of electrode e_j. Note that if the boundary and the

contact impedance are known, the Robin-to-Neumann and the Dirichlet-to-Neumann maps determine each other; that is, they represent equivalent information.

Consider now the bilinear form corresponding to liner maps $E : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}$ and $R : H^{-1/2}(\partial\Omega) \times H^{-1/2}(\partial\Omega) \to \mathbb{R}$ given by

$$E[V, \widetilde{V}] = \sum_{j=1}^{J} (EV)_j \widetilde{V}_j |e_j|, \quad R[h, \widetilde{h}] = \int_{\partial\Omega} (Rh)\, \widetilde{h}\, ds.$$

Let $S = \operatorname{span}(\chi_{e_j} : j = 1, \dots, J) \subset H^{-1/2}(\partial\Omega)$ and define $M : V = (V_j)_{j=1}^J \mapsto \sum_{j=1}^J V_j \chi_{e_j}$ to be a map $M : \mathbb{R}^N \to S$. Then

$$E[V, \widetilde{V}] = R[MV, M\widetilde{V}]. \tag{4.5}$$

Moreover, for $h = MV$ with some $V \in \mathbb{R}^J$, we have

$$R[h, h] = \int_{\partial\Omega} (u + z\nu\cdot\gamma\nabla u)\nu\cdot\gamma\nabla u\, ds = \int_\Omega \gamma\nabla u\cdot\nabla u\, dx + \int_{\partial\Omega} z\,|\nu\cdot\gamma\nabla u|^2\, ds, \tag{4.6}$$

where u solves (4.4). The integral over Ω in (4.6) is invariant in coordinate deformations. Note that in the above formula the integral over the boundary is not coordinate invariant.

Let $\widetilde{E}$ be the electrode measurement matrix for $\widetilde{\gamma}$ in $\widetilde{\Omega}$ with electrodes $\widetilde{e}_j = f(e_j)$, and let $\widetilde{R}$ be the Robin-to-Neumann map for $\widetilde{\gamma}$ defined analogously to (4.4). Since f is length preserving on the electrodes, we see using (1.3) that $\nu\cdot\gamma\nabla u(x) = \nu\cdot\widetilde{\gamma}\nabla\widetilde{u}(f(x))$ for $\widetilde{u} = u \circ F^{-1}$ and $x \in \partial\Omega$, and thus we can see from (4.6) that

$$R[h, \widetilde{h}] = \widetilde{R}[h \circ f^{-1}, \widetilde{h} \circ f^{-1}],$$

for $h, \widetilde{h} \in H^{-1/2}(\partial\Omega)$ supported in the closure of $\bigcup_{j=1}^J e_j$. Thus for the map $\widetilde{M} : V \mapsto \sum_{j=1}^J V_j \chi_{\widetilde{e}_j}$ we have by formula (4.5) that $\widetilde{E}[V, \widetilde{V}] = \widetilde{R}[\widetilde{M}V, \widetilde{M}\widetilde{V}]$. Combining this and (4.5), we obtain

$$E[V, V'] = \widetilde{E}[V, V'].$$

In particular, this implies that the matrices E and $\widetilde{E}$ coincide. ∎

In particular, in the case where $\widetilde{\Omega}$ is the model domain Ω_m and $f = f_m : \partial\Omega \to \partial\Omega_m$ is the model map for the boundary, the assumption that f is length preserving on electrodes means the very natural assumption that in electrode measurements the parameterization of the electrodes is known. Then by Proposition 4.1, the electrode model discretization E of Λ_γ equals the corresponding discretization $\widetilde{E}$ of $(f_m)_*\Lambda_\gamma$. Summarizing, the electrode measurements do not change if we have modeled the geometry of the boundary incorrectly but the electrodes are modeled correctly.

5. Numerical examples. The performance of the proposed method is evaluated by test cases with simulated EIT data. First, in section 5.1 we briefly discuss the discretization and the computational methods that are used, and the results are then given in section 5.2.

5.1. Discretization and notation. The numerical solution of the forward model is based on the finite element method (FEM). The variational formulation and the finite element discretization of the electrode model (4.1)–(4.3) in the case of isotropic conductivities have been previously discussed, e.g., in [10]. The extension of the FEM model to the anisotropic case is straightforward; the details will be given in a subsequent publication.

For the functions $\eta(x)$ and $\theta(x)$ in (1.5) we use piecewise constant approximations that are defined on a lattice of regular pixels. Thus, we have

$$\eta = \sum_{i=1}^{M} \eta_i \chi_i(x), \quad \theta = \sum_{i=1}^{M} \theta_i \chi_i(x), \tag{5.1}$$

where χ_i is the characteristic function of the ith pixel in the lattice. Within the discretization (5.1), the parameters η and θ are identified with the coefficient vectors

$$\eta = (\eta_1, \eta_2, \ldots, \eta_M)^{\mathrm{T}} \in \mathbb{R}^M,$$
$$\theta = (\theta_1, \theta_2, \ldots, \theta_M)^{\mathrm{T}} \in \mathbb{R}^M,$$

and λ is a scalar parameter. Note that as $\widehat{\gamma}_{\lambda,\eta,\theta} = \widehat{\gamma}_{\lambda',\eta,\theta'}$, where $\lambda' = 1/\lambda$ and $\theta'(x) = \theta(x) + \pi/2$, we can assume in looking at the minimizing uniformly anisotropic conductivity that λ gets values $\lambda > 0$.

In practical EIT devices, the measurements are made such that known currents are injected into the domain Ω through some of the electrodes at $\partial\Omega$, and the corresponding voltages needed to maintain these currents are measured on some of the electrodes. Often, voltages are measured only on those electrodes that are not used to inject current. Thus, measurements made give only partial information on the matrix E. To take this into account, we introduce the following notation for the discretized problem. We assume that the EIT experiment consists of a set of K partial voltage measurements, $V^{(j)}$, $j = 1, \ldots, K$. For each measurement, consider a current pattern $I^{(j)}$, $j = 1, \ldots, K$, such that $\sum_{\ell=1}^{J} I_\ell^{(j)} = 0$. Typically, the corresponding measurements are the voltages (potential differences) between pairs of neighboring electrodes. Let us assume that the measurement vector $V^{(j)}$ corresponding to the current pattern $I^{(j)}$ consists of L voltage measurements; i.e., we have $V^{(j)} \in \mathbb{R}^L$. Thus, we write $V^{(j)} = P_j E^{-1} I^{(j)} + \epsilon^{(j)}$, where E is the electrode measurement matrix, random vector $\epsilon^{(j)}$ models the observation errors, and $P_j : \mathbb{R}^J \to \mathbb{R}^L$ is a measurement operator that maps the electrode potentials to measured voltages.

In the inverse problem, the voltage measurements $V^{(1)}, V^{(2)}, \ldots, V^{(K)}$ are concatenated into a single vector

$$V = (V^{(1)}, V^{(2)}, \ldots, V^{(K)})^{\mathrm{T}} \in \mathbb{R}^N, \quad N = KL.$$

For the finite element–based discretization of the forward problem $U : \mathbb{R}^{2M+1} \mapsto \mathbb{R}^N$, we use the notation

$$U(\eta, \theta, \lambda) = (U^{(1)}(\eta, \theta, \lambda), U^{(2)}(\eta, \theta, \lambda), \ldots, U^{(K)}(\eta, \theta, \lambda))^{\mathrm{T}} \in \mathbb{R}^N,$$

respectively. Here, $U^{(j)}(\eta, \theta, \lambda) = P_j E^{-1}(\eta, \theta, \lambda) I^{(j)} \in \mathbb{R}^L$ corresponds to partial voltage measurement, with current pattern $I^{(j)}$ and conductivity $\widehat{\gamma}_{\eta,\theta,\lambda}$.

Using the above notation, we write the discretized and regularized version of our inverse problem as finding the minimizer of the functional

$$F(\eta, \theta, \lambda) = \|V - U(\eta, \theta, \lambda)\|^2 + W_\eta(\eta) + W_\theta(\theta) + W_\lambda(\lambda), \quad \eta > 0, \lambda > 0, \tag{5.2}$$

where the regularizing penalty functionals are of the form

$$W_\eta(\eta) = \alpha_0 \sum_{i=1}^{M} \eta_i^2 + \alpha_1 \sum_{i=1}^{M} \sum_{j \in \mathcal{N}_i} |\eta_i - \eta_j|^2, \tag{5.3}$$

$$W_\theta(\theta) = \beta_0 \sum_{i=1}^{M} \theta_i^2 + \beta_1 \sum_{i=1}^{M} \sum_{j \in \mathcal{N}_i} |e^{i\theta_i} - e^{i\theta_j}|^2, \tag{5.4}$$

$$W_\lambda(\lambda) = \beta_2 \left(\log(\lambda) + \nu^{-2} \log(\lambda)^2 \right) \tag{5.5}$$

and $\mathcal{N}_i$ denotes the usual four-point nearest neighborhood system for pixel i in the lattice.

Our objective is to minimize the functional (5.2) by gradient-based optimization methods. Here we face difficulty due to the positivity constraints. To take the positivity constraint into account we employ the interior point search method [7]. In the interior point search the original constrained problem (5.2) is replaced by a sequence of augmented unconstrained problems of the form

$$\widetilde{F}_j(\eta, \theta, \lambda) = F(\eta, \theta, \lambda) + W_+^{(j)}(\eta), \tag{5.6}$$

where $W_+^{(j)}(\eta)$ is a penalty functional of the form

$$W_+^{(j)}(\eta) = \xi_j \sum_{i=1}^{M} \frac{1}{\eta_i} \tag{5.7}$$

and $\{\xi_j\}$ is a sequence of decreasing positive parameters such that $\xi_j \to 0$ as $j \to \infty$. Using a suitably chosen sequence of penalty functionals $W_+^{(j)}$, the solutions of the unconstrained problems converge (asymptotically) to the solution of the original constrained problem. The positivity constraint for λ can be taken care of with similar techniques. However, it is our experience that the positivity constraint was not needed for λ.

For the minimization of the functionals (5.6) we employ the Gauss–Newton optimization method with an explicit line search algorithm.

5.2. Results. In this section, we evaluate the performance of the proposed method with three different test cases. The first test case is EIT data from an ellipse domain Ω, in the second test case we consider an ellipse domain with a sharp cut, and in the last test case the domain is a smooth Fourier domain which has some resemblance to the cross section of the human body. In all of these cases, we use the unit disk as the model domain Ω_m.

In the simulations, we assume an EIT system with $J = 16$ electrodes. In each of the test cases, the electrodes were located at approximately equally spaced positions at the exterior boundary $\partial\Omega$ of the target domain Ω. The size of the electrodes was chosen such that the electrodes covered approximately 50% of the boundary $\partial\Omega$.

The EIT measurements were simulated using the usual adjacent pair drive data acquisition method. In the adjacent drive method, currents $+1$ and -1 are injected through two neighboring electrodes, say electrodes e_n and e_{n+1}, and the current through other electrodes is zero. The voltages are measured between all J pairs of

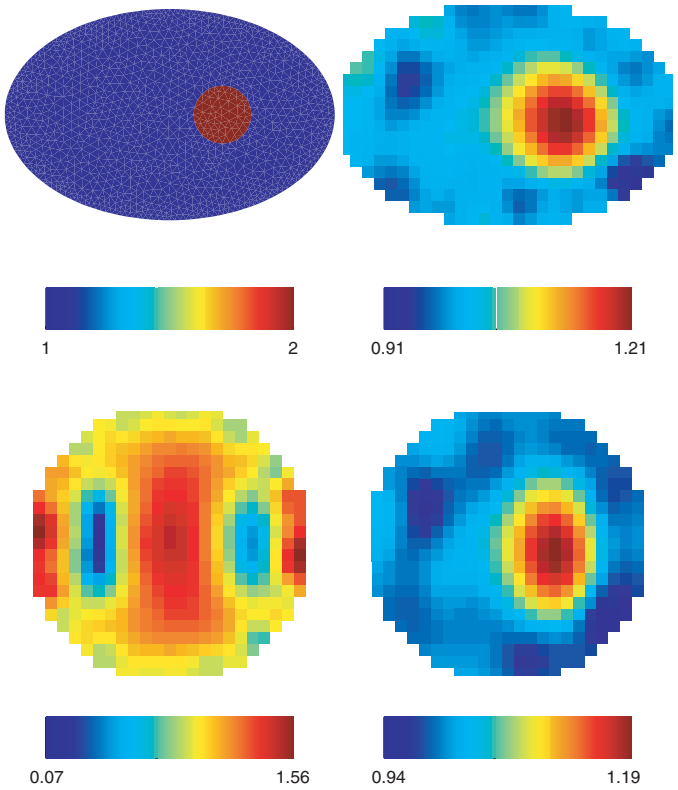


FIG. 5.1. *Test case with EIT data from an ellipse domain* Ω. *The main axes of the ellipse were* 1.25 *in horizontal direction and* 0.8 *in the vertical direction. Top left: Simulated conductivity distribution* γ. *Top right: Reconstruction of* γ *with the isotropic EIT model in the correct domain* Ω. *Bottom left: Reconstruction of* γ *with the isotropic model in incorrectly modeled geometry. The reconstruction domain* Ω_m *was the unit disk. Bottom right: Reconstruction of* η *with the uniformly anisotropic model in the same unit disk geometry.*

neighboring electrodes. However, three of these measurements are typically neglected since they include either one or both of the current feeding electrodes e_n or e_{n+1}. The rationale behind this is that the electrode contact impedances z_j are usually not known accurately. The possible errors in the contact impedance values cause a systematic error between the measured voltage and the forward model for the measurement made on the current feeding electrodes, and this error causes artifacts in the numerical reconstruction; see, e.g., [9]. Thus, with the adjacent pair drive method each partial measurement consists of $L = J - 3$ voltage measurements, and we have $V^{(j)} \in \mathbb{R}^{J-3}$. This data acquisition process is then repeated for all the J pairs of adjacent electrodes, leading to a total of $N = J(J-3)$ voltage measurements for one EIT experiment. Thus, with the $J = 16$ electrode system we have $V \in \mathbb{R}^{208}$.

The simulated EIT measurements were computed using the isotropic EIT model and the FEM. To simulate measurement noise, we added to the data Gaussian random noise with standard deviation of 1% of the maximum value of the simulated voltages. In all of the following test cases we used value $z_\ell = 1$ for the electrode contact impedances. These were assumed known in the inverse problem.

The results for the first test case are shown in Figures 5.1–5.2. The target con-

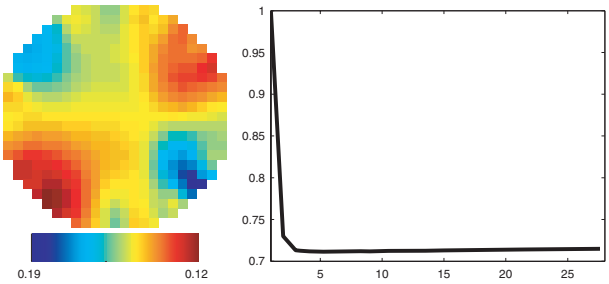


FIG. 5.2. *Test case with EIT data from an ellipse domain* Ω. *The main axes of the ellipse were* 1.25 *in horizontal direction and* 0.8 *in the vertical direction. Left: Reconstruction of the anisotropy angle parameter* θ *in the incorrectly modeled geometry. The computational domain* Ω_m *was the unit disk. Right: Evolution of the anisotropy parameter* λ *during the Gauss–Newton iteration.*

ductivity is shown in the top left image in Figure 5.1. The target domain Ω is an ellipse with main axes 1.25 in the horizontal direction and 0.8 in the vertical direction. For the simulation of the EIT measurements, the domain was discretized into a finite element mesh that consisted of 1256 nodal points and 2350 triangular elements.

The reconstruction of the conductivity γ with isotropic EIT model in the correct domain Ω is shown in the top right image in Figure 5.1. The reconstruction was obtained by using optimization techniques similar to those explained in the previous section. However, in the case of the isotropic model the unknown parameter vector is the conductivity vector $\gamma \in \mathbb{R}^M$, and the optimization functionals for the interior point search can be written as

$$H_j(\gamma) = \|V - U(\gamma)\|^2 + W_\gamma(\gamma) + W_+^{(j)}(\gamma), \tag{5.8}$$

where $U(\gamma)$ denotes the forward problem for the isotropic model and $W_\gamma(\gamma)$ and $W_+^{(j)}(\gamma)$ are defined by (5.3) and (5.7), respectively. To compute the reconstruction in the top right image in Figure 5.1, the domain Ω was triangulated to a finite element mesh that consisted of 2326 elements with 1244 nodal points. The conductivity was represented in a lattice of $M = 451$ pixels (i.e., $\gamma \in \mathbb{R}^{451}$). The regularization parameters for the penalty functional $W_\gamma(\gamma)$ in (5.8) were $\alpha_0 = 10^{-8}$ and $\alpha_1 = 10^{-4}$. When computing the reconstruction in the correctly modeled geometry, the interior point search was kept inactive (i.e., the sequence $\{\xi_j\}$ of interior point search parameters were all zeros). The conductivity vector was initialized to a constant value of one in the optimization process. The Gauss–Newton optimization algorithm was iterated until convergence was obtained.

The image in the bottom left in Figure 5.1 shows the reconstruction of the conductivity with the isotropic model in incorrectly modeled geometry Ω_m. In this case, the computational domain Ω_m was the unit disk, which was triangulated to 2190 elements with 1176 nodal points. The conductivity parameters were represented in a lattice of $M = 437$ pixels (i.e., $\gamma \in \mathbb{R}^{437}$). The regularization parameters for the penalty functional $W_\gamma(\gamma)$ in (5.8) were $\alpha_0 = 10^{-8}$ and $\alpha_1 = 2 \cdot 10^{-4}$. The sequence of interior point search parameters $\{\xi_j\}$ were from $2 \cdot 10^{-5}$ to $5 \cdot 10^{-6}$. The constant vector $\gamma = 1 \in \mathbb{R}^{437}$ was used as the initial guess in the Gauss–Newton optimization.

The image in the bottom right in Figure 5.1 shows the reconstruction of η with the uniformly anisotropic model in incorrectly modeled geometry Ω_m. Here, by the solution in *the uniformly anisotropic model* we mean the optimal solution of the form

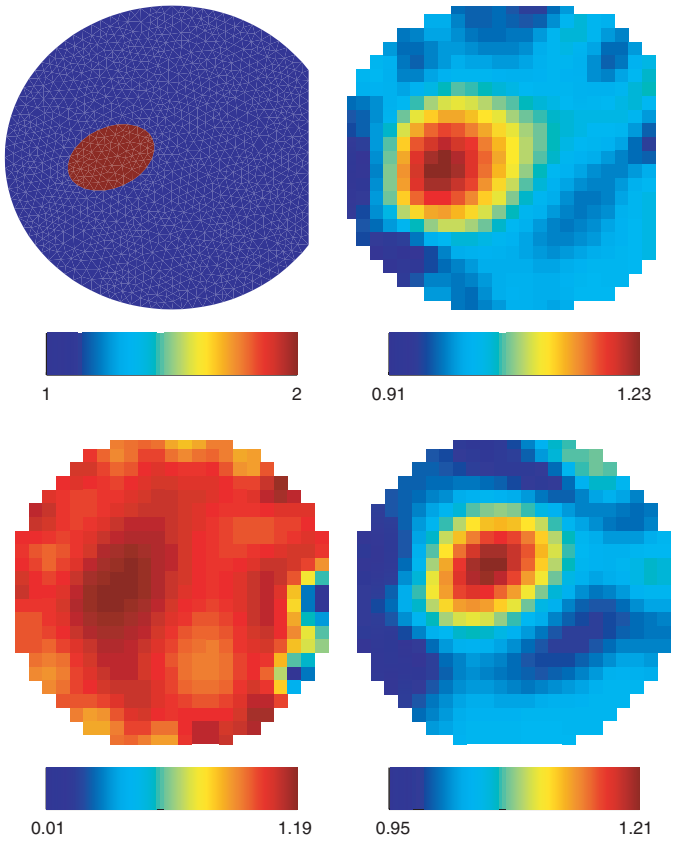


FIG. 5.3. *Test case with EIT data from a truncated ellipse domain Ω with main axes $a = 1.1$ and $b = 0.9$. Top left: Simulated conductivity distribution γ. Top right: Reconstruction of the conductivity γ with the isotropic model in the correct geometry Ω. Bottom left: Reconstruction of γ with the isotropic model in incorrectly modeled geometry. The reconstruction domain Ω_m was the unit disk. Bottom right: Reconstruction of the parameter η with the uniformly anisotropic model in the same unit disk geometry.*

(1.5) of the minimization problem. The reconstruction was obtained by minimizing a sequence of optimization functionals of the form (5.6). The reconstructed angle parameter θ is shown in the left image in Figure 5.2, and the evolution of the parameter λ during the iteration is shown in the right image in Figure 5.2. The computational domain Ω_m was the unit disk. The finite element triangularization and the number of the image pixels were the same as in the isotropic case in bottom left image in Figure 5.1. Thus, the unknowns in the inverse problem are $\eta \in \mathbb{R}^{437}$, $\theta \in \mathbb{R}^{437}$, and $\lambda \in \mathbb{R}$. The parameters for the regularizing penalty functionals $W_\eta(\eta)$ in (5.3) were $\alpha_0 = 10^{-8}$ and $\alpha_1 = 10^{-4}$. The parameters for the penalty functionals $W_\theta(\theta)$ and $W_\lambda(\lambda)$ in (5.4)–(5.5) were $\beta_0 = 10^{-8}$, $\beta_1 = 5 \cdot 10^{-6}$, and $\beta_2 = 0$, respectively. The sequence of interior point search parameters $\{\xi_j\}$ was from $1 \cdot 10^{-5}$ to $1 \cdot 10^{-12}$. The Gauss–Newton optimization was started from the constant values $\eta = 1 \in \mathbb{R}^{437}$, $\theta = 0 \in \mathbb{R}^{437}$, and $\lambda = 1$, which correspond to isotropic unit conductivity.

The results for the second test case are shown in Figure 5.3. The simulated conductivity distribution is shown in the top left image. In this case the domain Ω is a

truncated ellipse with main axes 1.1 in the horizontal direction and 0.9 in the vertical direction, respectively. For the simulation of the EIT measurements, the domain was divided into a finite element mesh of 2383 triangular elements with 1240 nodes.

The top right image in Figure 5.3 shows the reconstruction of the conductivity with the isotropic model in the correct geometry. For the reconstruction, the domain Ω was divided into a finite element mesh of 2337 triangular elements with 1217 nodes, and the conductivity was represented on a lattice of $M = 455$ pixels. Thus, the unknown parameter vector was $\gamma \in \mathbb{R}^{455}$. The regularization parameters for the penalty functional $W_\gamma(\gamma)$ in (5.8) were $\alpha_0 = 10^{-8}$ and $\alpha_1 = 10^{-4}$. The sequence of interior point search parameters $\{\xi_j\}$ was all zeros. The Gauss–Newton optimization was started from the constant unit conductivity.

The bottom left image in Figure 5.3 shows the reconstructed conductivity with the isotropic model in the incorrectly modeled geometry. The reconstruction domain Ω_m was the unit disk. The finite element mesh and pixel lattice were the same as those used for the unit disk in Figure 5.1. Thus, the unknown conductivity vector was $\gamma \in \mathbb{R}^{437}$. The parameters for the regularizing penalty functional $W_\gamma(\gamma)$ were $\alpha_0 = 10^{-8}$ and $\alpha_1 = 10^{-4}$, and the sequence of interior point search parameters $\{\xi_j\}$ was from 10^{-5} to 10^{-8}. The constant unit conductivity was used as the initial guess in the optimization.

The bottom right image in Figure 5.3 shows the reconstruction of η with the uniformly anisotropic model in the incorrectly modeled geometry. The computational domain Ω_m was the unit disk with the same discretization that was used in Figure 5.1. Thus, the unknowns were $\eta \in \mathbb{R}^{437}$, $\theta \in \mathbb{R}^{437}$, and $\lambda \in \mathbb{R}$. The parameters for the regularizing penalty functionals $W_\eta(\eta)$ in (5.3) were $\alpha_0 = 10^{-8}$ and $\alpha_1 = 10^{-4}$. The parameters for the penalty functionals $W_\theta(\theta)$ and $W_\lambda(\lambda)$ in (5.4)–(5.5) were $\beta_0 = 10^{-8}$, $\beta_1 = 5 \cdot 10^{-6}$, and $\beta_2 = 0$, respectively. The sequence of parameters $\{\xi_j\}$ was from 10^{-5} to 10^{-12}. The initializations for the parameters in the Gauss–Newton optimization were the constant values $\eta = 1 \in \mathbb{R}^{437}$, $\theta = 0 \in \mathbb{R}^{437}$, and $\lambda = 1$.

The results for the last test case are shown in Figure 5.4. In this case, the target domain Ω is bounded by a smooth Fourier boundary $\partial\Omega$. The true isotropic conductivity distribution within the domain Ω is shown in the top left image in Figure 5.4. For the simulation of the EIT measurements, the domain Ω was divided into a mesh of 2316 triangular elements with 1239 nodes.

The reconstruction of the conductivity γ with the isotropic model in the correct geometry Ω is shown in the top right image in Figure 5.4. The domain was divided into a mesh of 2200 triangular elements with 1181 nodes for the image reconstruction process. The number of pixels was $M = 446$ for the representation of the conductivity image (i.e., $\gamma \in \mathbb{R}^{446}$). The regularization parameters for the penalty functional $W_\gamma(\gamma)$ were $\alpha_0 = 10^{-8}$ and $\alpha_1 = 10^{-5}$, and the sequence of parameters $\{\xi_j\}$ was all zeros. The constant unit conductivity was used as the initial guess in the Gauss–Newton optimization algorithm.

The reconstruction of the conductivity γ with the isotropic model in the incorrectly modeled geometry is shown in the bottom left image in Figure 5.4. The reconstruction domain Ω_m was the unit disk. The finite element mesh and the pixel lattice were the same as those used in Figures 5.1–5.3. Thus, the parameter vector in the inverse problem was $\gamma \in \mathbb{R}^{437}$. The parameters in the penalty functional $W_\gamma(\gamma)$ were $\alpha_0 = 10^{-8}$ and $\alpha_1 = 2{\cdot}10^{-4}$, and the sequence of parameters $\{\xi_j\}$ was from $2{\cdot}10^{-5}$ to $5{\cdot}10^{-6}$. The constant unit conductivity was used as the initial guess in the optimization.

The reconstruction of η with the uniformly anisotropic model in the incorrectly

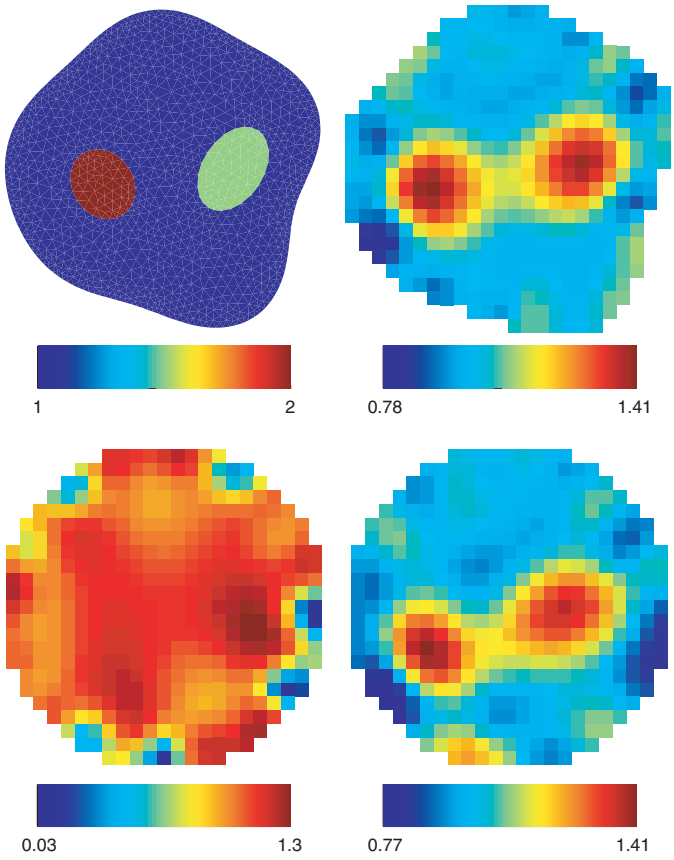


FIG. 5.4. *Test case with EIT data from an arbitrary domain Ω. Top left: True conductivity distribution γ. Top right: Reconstruction of the conductivity γ with isotropic model in the correct geometry Ω. Bottom left: Reconstruction of γ with the isotropic model in incorrectly modeled geometry. The reconstruction domain Ω_m was the unit disk. Bottom right: Reconstruction of the parameter η with the uniformly anisotropic model in the same unit disk geometry.*

modeled geometry is shown in the bottom right image in Figure 5.4. The reconstruction domain Ω_m was the unit disk with the same discretization as in Figures 5.1–5.3. Thus, the unknown parameter vectors were $\eta \in \mathbb{R}^{437}$, $\theta \in \mathbb{R}^{437}$, and $\lambda \in \mathbb{R}$. The parameters for the regularizing penalty functionals $W_\eta(\eta)$ in (5.3) were $\alpha_0 = 10^{-8}$ and $\alpha_1 = 10^{-5}$. The parameters for the penalty functionals $W_\theta(\theta)$ and $W_\lambda(\lambda)$ in (5.4)–(5.5) were $\beta_0 = 10^{-8}$, $\beta_1 = 5 \cdot 10^{-6}$, and $\beta_2 = 0$, respectively. The sequence of parameters $\{\xi_j\}$ was from 10^{-5} to 10^{-12}. The initializations for the image parameters were the constant values $\eta = 1 \in \mathbb{R}^{437}$, $\theta = 0 \in \mathbb{R}^{437}$, and $\lambda = 1$.

6. Discussion. As can be seen from Figures 5.1–5.4, the proposed approach gives good results. In all test cases, the traditional reconstructions with the isotropic model are erroneous when the imaging geometry is modeled incorrectly. The effects of erroneous geometry are seen in the reconstructions as distortions and severe artifacts, especially near the boundary. On the other hand, the reconstructions of η with the uniformly anisotropic model in the same erroneous geometry are clear of these artifacts and represent a deformed picture of the original isotropic conductivity. These results

indicate that the proposed method offers an efficient tool to eliminate the difficulties that arise from inaccurately known geometry in practical EIT experiments.

Acknowledgments. The authors are thankful for Prof. Kari Astala and Prof. Seppo Rickman for discussions on quasi-conformal maps that were crucial for the obtained results. Also, thanks to the anonymous referees for valuable comments.

REFERENCES

[1] A. Adler, R. Guardo, and Y. Berthiaume, *Impedance imaging of lung ventilation: Do we need to account for chest expansion?* IEEE Trans. Biomedical Engineering, 43 (1996), pp. 414–420.

[2] L. V. Ahlfors, *Lectures on Quasiconformal Maps*, Van Nostrand, New York, 1966.

[3] K. Astala, M. Lassas, and L. Päivärinta, *Calderón's inverse problem for anisotropic conductivity in the plane*, Comm. Partial Differential Equations, 30 (2005), pp. 207–224.

[4] K. Astala and L. Päivärinta, *Calderon's inverse conductivity problem in the plane*, Ann. Math., (2005), to appear.

[5] V. Bozin,, N. Lakic, V. Markovic, and M. Mateljevic, *Unique extremality*, J. Anal. Math., 75 (1998), pp. 299–338.

[6] A.-P. Calderón, *On an inverse boundary value problem*, in Proceedings of the Seminar on Numerical Analysis and Its Applications to Continuum Physics (Rio de Janeiro, 1980), Soc. Brasil Mat., Rio de Janeiro, 1980, pp. 65–73.

[7] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Classics in Applied Mathematics 4, SIAM, Philadelphia, 1990.

[8] E. Gersing, B. Hoffman, and M. Osypka, *Influence of changing peripheral geometry on electrical impedance tomography measurements*, Medical & Biological Engineering & Computing, 34 (1996), pp. 359–361.

[9] V. Kolehmainen, M. Vauhkonen, P. A. Karjalainen, and J. P. Kaipio, *Assessment of errors in static electrical impedance tomography with adjacent and trigonometric current patterns*, Physiological Measurement, 18 (1997), pp. 289–303.

[10] J. Kaipio, V. Kolehmainen, E. Somersalo, and M. Vauhkonen, *Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography*, Inverse Problems, 16 (2000), pp. 1487–1522.

[11] O. Lehto, *Univalent Functions and Teichmüller Mappings*, Graduate Texts in Math. 109, Springer, New York, 1986.

[12] A. Nachman, *Reconstructions from boundary measurements*, Ann. Math., 128 (1988), pp. 531–576.

[13] A. Nachman, *Global uniqueness for a two-dimensional inverse boundary value problem*, Ann. Math., 143 (1996), pp. 71–96.

[14] Ch. Pommerenke, *Boundary Behaviour of Conformal Maps*, Grundlehren Math. Wiss. 299, Springer-Verlag, Berlin, 1992.

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

[16] K. Strebel, *On the existence of extremal Teichmüller mappings*, J. Anal. Math., 30 (1976), pp. 464–480.

[17] J. Sylvester, *An anisotropic inverse boundary value problem*, Comm. Pure Appl. Math., 43 (1990), pp. 201–232.

[18] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. Math., 125 (1987), pp. 153–169.

[19] M. Taylor, *Partial Differential Equations.* I. *Basic Theory*, Appl. Math. Sci. 115, Springer-Verlag, New York, 1996.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 384–407

THEORY OF DETONATION WITH AN EMBEDDED SONIC LOCUS*

D. SCOTT STEWART† AND ASLAN R. KASIMOV‡

Abstract. A steady planar self-sustained detonation has a sonic surface in the reaction zone that resides behind the lead shock. In this work we address the problem of generalizing sonic conditions for a three-dimensional unsteady self-sustained detonation wave. The conditions are proposed to be the characteristic compatibility conditions on the exceptional surface of the governing hyperbolic system of reactive Euler equations. Two equations are derived that are necessary to determine the motion of both the lead shock and the sonic surface. Detonation with an embedded sonic locus is thus treated as a two-front phenomenon: a reaction zone whose domain of influence is bounded by two surfaces, the lead shock surface and the trailing characteristic surface. The geometry of the two surfaces plays an important role in the underlying dynamics. We also discuss how the sonic conditions of detonation stability theory and detonation shock dynamics can be obtained as special cases of the general sonic conditions.

Key words. chemically reacting flows, supersonic flows, transonic flows, shocks and singularities

AMS subject classifications. 35L40, 76H05, 76N99

DOI. 10.1137/040616930

1. Introduction.

A detonation wave is a shock wave that triggers exothermic reactions in an explosive as it propagates so that the energy released in the reactions sustains the shock propagation. Modern theories of detonation originate from the theory first developed independently by Zel'dovich, von Neumann, and Doering in the 1940s (ZND theory; see Fickett and Davis [5] for details) that describes the dynamics of a steady one-dimensional planar detonation in a gaseous explosive. The ZND theory is applicable to both self-sustained detonations, that is, autonomous waves whose motion is sustained entirely by the energy released in their reaction zone, and overdriven detonations which require an additional external support to maintain their motion at a nominal speed. In self-sustained steady one-dimensional planar detonations, which are also called Chapman−Jouguet (CJ) detonations, there exists an embedded sonic locus within or at the end of the reaction zone, such that at that point the flow speed is sonic relative to the shock. As a consequence, the lead-shock dynamics is influenced only by the flow between the shock and the sonic locus. In contrast, the lead-shock dynamics of overdriven detonations is influenced by the entire region between the shock and the support (e.g., a piston); no sonic locus exists in such detonations. Without the condition of sonicity, the equations governing the CJ detonation (the mass, momentum, and energy equations) are not closed, since the detonation speed is unknown; the sonicity condition provides the necessary closure. Understanding the nature of the sonic conditions in detonations more general than

*Received by the editors October 13, 2004; accepted for publication (in revised form) May 26, 2005; published electronically November 22, 2005. This work was supported by the U.S. Air Force Office of Scientific Research under contracts F49620-00-1-0005 and F49620-03-1-0048 (Program Manager Dr. Arje Nachman).

http://www.siam.org/journals/siap/66-2/61693.html

†Department of Mechanical and Industrial Engineering, University of Illinois, Urbana-Champaign, Urbana, IL 61801 (dss@uiuc.edu). This author was supported by the U.S. Air Force Research Laboratory Munitions Directorate, Eglin AFB, under contract F8630-00-1-0002, and by the U.S. Department of Energy, Los Alamos National Laboratory, under contract DOE/LANL 3223501019Z.

‡Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139 (aslan@mit.edu).

planar, one-dimensional, steady detonations of the ZND theory has been difficult to achieve. It is precisely this task of deriving the general sonic conditions and clarifying their nature that is central to our present investigation.

Research that began in the late 1950s and early 1960s (see, e.g., [5, 3]) has shown that most detonation waves, especially in gases, have a multidimensional cellular structure with transversely propagating shock waves in the reaction zone and significant unsteady dynamics. In condensed explosives, the detonation is more often observed to be steady, but importantly it has been known for a long time that high-explosive detonation shocks are almost always curved. Clearly, the ZND theory is too simple to account for the observed structure and must be appropriately modified. There exist conceptual problems that cannot be addressed within the framework of the ZND theory if unsteady and multidimensional detonations are considered. The principal problem has to do with the nature of the sonic condition whose generalization to include unsteady and multidimensional effects has been limited so far to linearized problems and quasi-steady detonations.

In the linear stability theory of detonation, the far-field conditions are commonly referred to as "radiation conditions" or "boundedness conditions" depending on specific circumstances (see [4, 7, 9, 13]). The radiation condition is imposed to filter out incoming acoustic perturbations by considering the far-field acoustic solutions of the governing linearized system and to eliminate the incoming waves by setting their amplitude equal to zero. It follows then that the far-field solutions are linearly dependent and their linear combination forms a far-field constraint on the general solution of the linearized problem. Such a constraint serves as a dispersion relation that allows one to determine the eigenvalues. It turns out (see section 5) that the CJ limit (self-sustained wave) of the radiation condition coincides with the linearized governing equation on the forward characteristic. One can also show that the radiation condition in that case is also a boundedness condition for the solutions of the linearized system at the sonic locus. Thus in the linear stability problem, the general nature of the radiation conditions that provide the dispersion relation is such that they serve as filters of the incoming perturbations and are thus conditions on the forward characteristic surface that acts as an information boundary.

In the theory of detonation shock dynamics (DSD; see the topical review by Stewart [14] for a general discussion and history of the problem), one treats a quasi-steady *curved* detonation and derives sonic conditions (called generalized sonic conditions) that include effects of multidimensionality through the shock curvature term, which is assumed small on the scales of the reaction zone. Originally, the effect of curvature in the sonic conditions was considered by Wood and Kirkwood [17] and later was derived rationally in the works of Bdzil [1] and Stewart and Bdzil [15, 16]. Yao and Stewart [18] considered an extension of the sonic conditions to include asymptotically small unsteady corrections, but their analysis relies partially on the steady concept of a sonic locus by assuming that the flow is sonic relative to the lead shock, which constrains the sonic locus to always be parallel to the shock. The quasi-steady generalized CJ conditions reflect the fact that in a curved detonation, the flow divergence or convergence acts as a sink or source, respectively, of the energy of the lead shock. Thus, for example, in a diverging steady detonation, the sonic condition expresses an exact balance of the heat release and flow divergence, as shown by the equation given in Stewart and Bdzil [16]:

$$(\gamma - 1)\, Q\omega - c^2\, (D + U_n)\, \kappa = 0, \tag{1.1}$$

where Q is the heat release, γ is the adiabatic exponent, ω is the reaction rate at

the sonic point, c is the sound speed, D and U_n are the normal detonation speed and particle velocity at the sonic point relative to the lead shock, and κ is the shock curvature. Equation (1.1) is obtained from the equation (called Master equation)

$$\frac{\mathrm{d}U_n^2}{\mathrm{d}\lambda} = \frac{2U_n^2\left[(\gamma-1)\,Q\omega - c^2\,(D_n+U_n)\,\kappa\right]}{\omega\,(c^2-U_n^2)} \tag{1.2}$$

(λ is the reaction progress variable) that follows directly from the governing equations by a regularity argument, namely, that for the left-hand side of (1.2) to remain finite, the numerator of the right-hand side has to vanish at the sonic point because the denominator vanishes there: $c^2 = U_n^2$.

For unsteady weakly curved detonations, the Master equation can again be written in a form similar to (1.2), but the numerator contains more terms (see [18]):

$$\frac{\partial U_n}{\partial n} = \frac{1}{c^2-U_n^2}\left[(\gamma-1)\,Q\omega - c^2\,(D+U_n)\,\kappa + U_n\left(\frac{\partial U_n}{\partial t}+\frac{\partial D}{\partial t}\right) - v\frac{\partial p}{\partial t}\right], \tag{1.3}$$

where t is time, n is the normal distance from the shock ($n<0$ in the reaction zone), v is the specific volume, and p is pressure. A regularity argument is again invoked that requires that the numerator of (1.3) vanish at the sonic point, *assuming* that the denominator vanishes there as well: $c^2-U_n^2=0$. The latter assumption is one of the key elements that distinguishes the present theory from that of Yao and Stewart [18]—we do not define the sonic locus in the shock-attached frame, so that in our theory, $c^2-U_n^2$ does not necessarily vanish at the sonic locus. In fact, from the characteristic analysis, we find that $c + U_n = \partial n_*/\partial t = \mathcal{D} - D$, where n_* is the distance between the shock and the sonic locus, and $\mathcal{D}$ and D are the speeds of the sonic locus and of the shock, respectively. Thus $c+U_n$ at the sonic locus is equal to the relative speed of the sonic locus and the shock. Therefore, the theory of Yao and Stewart contains an implicit assumption that the sonic locus and the shock are parallel in the characteristic (n,t)-plane. In unsteady detonations, a possible imbalance of the heat release and flow divergence is reflected in the unsteadiness of the curved detonation.

Our generalization of the sonic conditions stems from the following observations. In a general unsteady flow that is sufficiently smooth, with a lead detonation shock, one considers all forward propagating characteristic surfaces, which are the envelopes of the forward propagating acoustic wavefronts. For initial conditions that admit smooth evolution, there may exist a limiting forward characteristic surface that never intersects the shock or intersects the shock only at times that are very long compared to the passage time of particles through the detonation reaction zone. This limiting characteristic is thus identified as a separatrix of the family of forward characteristic surfaces whose motion is toward the shock. On the upstream side of the separatrix, the forward characteristic surfaces flow into the shock in a finite time, while on the downstream side, they flow away from the shock. The region that affects the lead-shock dynamics (the domain of influence) is the region between the shock surface and the limiting characteristic surface so that the evolution of the detonation wave depends only on the data in that region. The limiting sonic surface is then specifically embedded in the reaction zone, usually at a finite distance behind the shock. In Kasimov and Stewart [8], we illustrated the behavior of the sonic locus as a limiting characteristic in one-dimensional detonations by means of a numerical simulation.

Thus a general sonic locus is proposed to be a characteristic surface of the governing hyperbolic equations such that the surface acts as an information boundary that precludes incoming acoustic perturbations from influencing the lead-shock dynamics.

Such a definition is in agreement with the limiting cases of the steady detonation, the unsteady linearized theory, and the weakly curved slowly varying detonation theories that have been derived previously. The new concept clarifies the meaning of the sonic locus by emphasizing its nature as a characteristic surface. In particular, since the sonic locus is a boundary of the domain of influence of the reaction zone, it follows immediately that the detonation problem is, in general, a *two-front problem* with both fronts (the shock and sonic loci) as free boundaries. Therefore, the sonic conditions must be given by *two* equations, a situation that has not been explicitly emphasized but is nevertheless a part of all previous theories of detonation. For example, in the planar CJ detonation, the two equations are (1) the well-known CJ condition, $M_{\rm CJ} = 1$, where $M_{CJ} = -U_n/c$ is the local Mach number relative to the shock and (2) the condition that the sonic point coincides with the end of the reaction zone (for single-step exothermic reaction), $\lambda = 1$. We propose that the sonic conditions for general multidimensional detonations are (1) the condition of local sonicity, that is, for an observer moving with the sonic surface, the particle speed normal to that surface, $\mathcal{U}_n$, is locally sonic,

$$\mathcal{U}_n = -c\,, \tag{1.4}$$

and (2) the compatibility condition in the sonic surface defined as a characteristic surface of the governing reactive Euler equations,

$$\rho c \mathbf{n}_* \cdot \left(\frac{D\mathbf{u}}{Dt} + \frac{1}{\rho}\nabla p \right) + \rho c^2 \nabla \cdot \mathbf{u} + \frac{Dp}{Dt} = \rho c^2 \sigma \omega, \tag{1.5}$$

where $\mathbf{n}_*$ is the unit normal to the sonic surface, $\mathbf{u}$ is the lab-frame particle velocity, $D/Dt = \partial/\partial t + \mathbf{u}\cdot\nabla$ is the material derivative, and σ is the thermicity coefficient. These two conditions are direct consequences of the governing hyperbolic equations and hold therefore under quite general circumstances; no asymptotic ideas are involved.

In section 2 we work out the theory of the characteristic surfaces for general systems of quasi-linear hyperbolic PDEs and derive compatibility conditions in the exceptional surface. The conditions are specialized to reactive Euler equations in section 2.2. In section 3 we discuss the simplest version of the sonic conditions in one spatial dimension to emphasize the connection with the standard theory of characteristics. Section 4 is devoted to two-dimensional detonations where we specialize the sonic conditions to local frames in order to exhibit the connection with the older theories of DSD. The connection of the present work with the theories of detonation stability is a subject of section 5. We conclude in section 6.

2. General theory. This section is divided into two subsections. The first is a general discussion and review of properties of characteristic surfaces defined for systems of hyperbolic PDEs. We quickly specialize to the reactive, compressible flow equations, but the presentation is not restricted to compressible Euler equations and has applications to other hyperbolic systems. The second subsection derives conditions that must be satisfied on a characteristic (sonic) surface, specifically for the reactive Euler equations that are relevant for application to detonation.

2.1. Characteristic surfaces of hyperbolic PDEs and compatibility conditions. The analysis given next closely follows that given in von Mises' treatise [10]. This presentation was developed by G. S. S. Ludford (along with von Mises' wife Hilda Geiringer) to complete the von Mises monograph after his death. Its teaching

was a regular feature of Ludford's famous courses on applied mathematics given at Cornell University. The von Mises reference is one of the few places one can find the general theory of characteristic surfaces written in a succinct and concise manner, and while classical in its form, it is seldom referenced and not widely known. This powerful presentation in fact becomes the basis for our developments and extensions to generate useful and new three-dimensional results for application to detonations in particular. A useful discussion of characteristic surfaces can also be found in Chapman [2]. Another useful reference is Ovsiannikov [12], where one can find a general characteristic form of equations of inert gas dynamics; the conditions on the acoustic characteristic surfaces are found to be similar to ours (see (2.28)), when no chemical reactions take place.

Consider a general system of quasi-linear hyperbolic equations written in the form

$$a_{ij}^k \frac{\partial u_j}{\partial x_k} = b_i \,, \tag{2.1}$$

where the coefficients a_{ij}^k are functions of the state variables u_j, $j = 1, 2, \ldots, J$, index i represents the individual equations of motion, x_k are the independent variables, and b_i are the source terms. Form a linear combination of the equations by multiplying the equations by arbitrary α_i and summing over all equations,

$$\alpha_i a_{ij}^k \frac{\partial u_j}{\partial x_k} \equiv m^k \frac{\partial}{\partial x_k} \, (u_j) = \alpha_i b_i . \tag{2.2}$$

Each term on the left-hand side of (2.2), $\alpha_i a_{ij}^k \, (\partial u_j / \partial x_k)$, is a directional derivative in space with direction tangents, $\mathbf{m}$, whose components, labeled by k, are given by $m^k = \alpha_i a_{ij}^k$. An *exceptional surface* [10] (or more commonly referred to as a characteristic surface) is defined as a surface such that the linear combination (2.2) of directional derivatives expresses changes only in that surface. Then all direction tangents must lie in that surface, and therefore the linear combination (2.2) contains no derivatives normal to the surface. If such an exceptional surface exists, then the unit normal vector $\boldsymbol{\beta}$ to the surface must be orthogonal to all tangent vectors, $\mathbf{m}$ (see Figure 2.1), that is,

$$m^k \beta_k = \alpha_i \beta_k a_{ij}^k = 0 \,. \tag{2.3}$$

This is a system of J homogeneous linear algebraic equations for α_i, with a nontrivial solution if and only if

$$\det |\beta_k a_{ij}^k| = 0 \,, \tag{2.4}$$

which is a Jth order polynomial that determines a constraint on the direction vector $\boldsymbol{\beta}$. Note that only directions in the space of the independent variables are solved for. If one of the independent variables is time, then the constraint on the direction in space time defines the velocity of the characteristic, which we later denote as the *speed relation*.

The *compatibility condition* is simply the differential relation, (2.2), found on the characteristic surface. The first step solves for β_k by solving the characteristic polynomial. The second step is, with a chosen direction, one that expresses the compatibility relation in the characteristic surface. Since the system of equations for α_i is singular, then the solution for α_i is determined up to an arbitrary constant; i.e.,

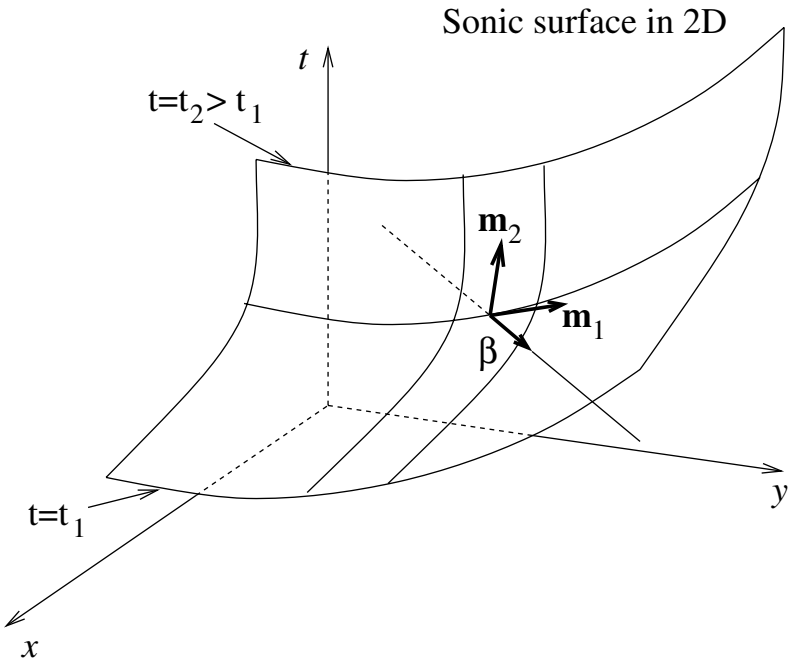


FIG. 2.1. *The sonic surface in* $2+1$ *dimensions, which is generated by time evolution of the two-dimensional sonic locus (a line in the xy-plane) along the third (time t) axis from* $t = t_1$ *to* $t = t_2$.

the ratio between the α_i is determined in terms of the β_k. Say such a direction β_k^* with a corresponding α_i^* is found. Then the *compatibility condition* is specifically

$$\alpha_i^* a_{ij}^k \frac{\partial u_j}{\partial x_k} = \alpha_i^* b_i \,. \tag{2.5}$$

2.2. Compatibility conditions for reactive Euler equations. We now start with reactive Euler equations with a single chemical reaction and closely follow the derivation given in von Mises [10] for the general case of fluid motion for inert flow. Further generalization to a multiple-step chemistry is straightforward. The general equation of state is used in its incomplete form, $e = e(p, \rho, \lambda)$.

Note that a simple device is in use. To simplify the algebraic presentation, the equations of motion are assumed to be analyzed at a point instantaneously aligned with the x-axis, which is taken in the direction of the velocity vector $\mathbf{u} = u\mathbf{i} + v\mathbf{j} + w\mathbf{k}$. Therefore, without loss of generality, the material derivative is $d/dt = \partial/\partial t + u\partial/\partial x$. The general condition for the exceptional surfaces is expressed for this special system and subsequently rewritten in a frame-invariant notation so that any coordinate system can be used. The notion of an exceptional (characteristic) surface is the one that is based on the physical equations and not the coordinates, and it is simply a matter of expressing the equations and directions indicated in those coordinates.

The equations of motion are written as

$$u\frac{\partial u}{\partial x} + \frac{\partial u}{\partial t} + \frac{1}{\rho}\frac{\partial p}{\partial x} = 0, \tag{2.6}$$

$$u\frac{\partial v}{\partial x}+\frac{\partial v}{\partial t}+\frac{1}{\rho}\frac{\partial p}{\partial y}=0, \tag{2.7}$$

$$u\frac{\partial w}{\partial x}+\frac{\partial w}{\partial t}+\frac{1}{\rho}\frac{\partial p}{\partial z}=0, \tag{2.8}$$

$$\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}+\frac{u}{\rho}\frac{\partial \rho}{\partial x}+\frac{1}{\rho}\frac{\partial \rho}{\partial t}=0, \tag{2.9}$$

$$u\frac{\partial p}{\partial x}+\frac{\partial p}{\partial t}-c^2\left(u\frac{\partial \rho}{\partial x}+\frac{\partial \rho}{\partial t}\right)=\rho c^2\sigma\omega, \tag{2.10}$$

$$u\frac{\partial \lambda}{\partial x}+\frac{\partial \lambda}{\partial t}=\omega, \tag{2.11}$$

where p is pressure, ρ is density, λ is the reaction-progress variable, ω is the reaction rate, and c is the frozen sound speed. We have used the definition of the thermicity coefficient given by Fickett and Davis [5],

$$\sigma=-\frac{1}{\rho c^2}\frac{e_\lambda}{e_p}\,, \tag{2.12}$$

and the general expression for the sound speed,

$$c^2=\frac{p-\rho^2 e_\rho}{\rho^2 e_p}, \tag{2.13}$$

where the subscripts of e denote partial differentiation with respect to the arguments.

The state vector u_j is given by $(u_j)=(u,v,w,p,\rho,\lambda)$, with $j=1,\ldots,6$. For the purpose of assigning the a^k_{ij}, we number (2.6) through (2.11) by $j=1,\ldots,6$. The generalized independent coordinates are given by the list $(x_k)=(x,y,z,t)$ with $k=1,\ldots,4$. The equations of motion written in the form (2.1) subsequently identify a^k_{ij} as

$$\begin{aligned}
&[a^k_{11}]=[u,0,0,1]\,,\ a^k_{12}=0\,,\ a^k_{13}=0\,,\ [a^k_{14}]=\left[\frac{1}{\rho},0,0,0\right]\,,\ a^k_{15}=0\,,\ a^k_{16}=0\,,\\
&a^k_{21}=0\,,\ [a^k_{22}]=[u,0,0,1]\,,\ a^k_{23}=0\,,\ [a^k_{24}]=\left[0,\frac{1}{\rho},0,0\right]\,,\ a^k_{25}=0\,,\ a^k_{26}=0\,,\\
&a^k_{31}=0\,,\ a^k_{32}=0\,,\ [a^k_{33}]=[u,0,0,1]\,,\ [a^k_{34}]=\left[0,0,\frac{1}{\rho},0\right]\,,\ a^k_{35}=0\,,\ a^k_{36}=0\,,\\
&[a^k_{41}]=[1,0,0,0]\,,\ [a^k_{42}]=[0,1,0,0]\,,\ [a^k_{43}]=[0,0,1,0]\,,\\
&a^k_{44}=0\,,\ [a^k_{45}]=\left[\frac{u}{\rho},0,0,\frac{1}{\rho}\right]\,,\ a^k_{46}=0\,,\\
&a^k_{51}=0\,,\ a^k_{52}=0\,,\ a^k_{53}=0\,,\ [a^k_{54}]=[u,0,0,1]\,,\ [a^k_{55}]=[-c^2u,0,0,-c^2]\,,\ a^k_{56}=0\,,
\end{aligned}$$

$$a^k_{61}=0\,,\ a^k_{62}=0\,,\ a^k_{63}=0\,,\ a^k_{64}=0\,,\ a^k_{65}=0\,,\ [a^k_{66}]=[u,0,0,1]\,. \tag{2.14}$$

The 6×6 characteristic matrix, $\beta_k a^k_{ij}$, becomes

$$\begin{bmatrix} \beta_0 & 0 & 0 & \beta_1/\rho & 0 & 0 \\ 0 & \beta_0 & 0 & \beta_2/\rho & 0 & 0 \\ 0 & 0 & \beta_0 & \beta_3/\rho & 0 & 0 \\ \beta_1 & \beta_2 & \beta_3 & 0 & \beta_0/\rho & 0 \\ 0 & 0 & 0 & \beta_0 & -c^2\beta_0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \beta_0 \end{bmatrix}, \tag{2.15}$$

where $\beta_0 \equiv u\beta_1 + \beta_4$. Setting its determinant equal to zero results in the characteristic equation

$$-\frac{\beta_0^4}{\rho}\left[\beta_0^2 - c^2(\beta_1^2 + \beta_2^2 + \beta_3^2)\right] = 0. \tag{2.16}$$

A fourfold repeated root is associated with the stream surfaces that form the characteristic surface described by setting $\beta_0 = u\beta_1 + \beta_4 = 0$. In addition, there are two other surfaces associated with the roots of the other factor,

$$\beta_0 = \pm c\sqrt{\beta_1^2 + \beta_2^2 + \beta_3^2}. \tag{2.17}$$

Our focus is on these directions since in a nominally one-dimensional, unsteady flow they would correspond to the forward and backward facing acoustic characteristics (i.e, C_+ and C_-) that are called the "Mach lines." We specifically work out the compatibility relation for both of them, as they occur in a pair, and later we will use the results for the characteristic surface that would correspond to the forward characteristic, as we will explain subsequently.

To display the compatibility relation we need to solve the equations for α_i, namely (2.3). Using the previous definitions, one obtains the six equations

$$\alpha_1\beta_0 + \alpha_4\beta_1 = 0\,, \qquad \alpha_2\beta_0 + \alpha_4\beta_2 = 0\,,$$

$$\alpha_3\beta_0 + \alpha_4\beta_3 = 0\,, \qquad \frac{1}{\rho}\left(\alpha_1\beta_1 + \alpha_2\beta_2 + \alpha_3\beta_3\right) + \alpha_5\beta_0 = 0\,,$$

$$\frac{\alpha_4}{\rho}\beta_0 - c^2\alpha_5\beta_0 = 0\,, \qquad \alpha_6\beta_0 = 0\,. \tag{2.18}$$

The solution of this system is, in terms of α_4 (note that $\beta_0 = u\beta_1 + \beta_4 \neq 0$),

$$\alpha_1 = -\frac{\alpha_4\beta_1}{\beta_0}\,,\ \alpha_2 = -\frac{\alpha_4\beta_2}{\beta_0}\,,\ \alpha_3 = -\frac{\alpha_4\beta_3}{\beta_0}\,,\ \alpha_5 = \frac{\alpha_4}{\rho c^2}\,,\ \alpha_6 = 0\,. \tag{2.19}$$

The compatibility condition (2.2) written out long becomes

$$\alpha_1 a^k_{1j}\frac{\partial u_j}{\partial x_k} + \alpha_2 a^k_{2j}\frac{\partial u_j}{\partial x_k} + \alpha_3 a^k_{3j}\frac{\partial u_j}{\partial x_k} + \alpha_4 a^k_{4j}\frac{\partial u_j}{\partial x_k} + \alpha_5 a^k_{5j}\frac{\partial u_j}{\partial x_k} = \alpha_5 b_5. \tag{2.20}$$

Substituting for the α_i in terms of α_4 leads to

$$\frac{-\alpha_4}{\beta_0}\left[\beta_1 a^k_{1j}\frac{\partial u_j}{\partial x_k} + \beta_2 a^k_{2j}\frac{\partial u_j}{\partial x_k} + \beta_3 a^k_{3j}\frac{\partial u_j}{\partial x_k}\right] + \alpha_4\left[a^k_{4j}\frac{\partial u_j}{\partial x_k} + \frac{1}{\rho c^2}a^k_{5j}\frac{\partial u_j}{\partial x_k}\right] = \frac{\alpha_4}{\rho c^2}b_5. \tag{2.21}$$

The reader is reminded that each of the terms in the equation represents one of the governing equations. Let us introduce the unit vector

$$\mathbf{n} = \frac{\beta_1 \mathbf{i} + \beta_2 \mathbf{j} + \beta_3 \mathbf{k}}{\sqrt{\beta_1^2 + \beta_2^2 + \beta_2^2}}. \tag{2.22}$$

This unit vector is normal to the tangent plane of the Mach cones, and hence normal to the instantaneous realization of the characteristic surface in the physical space.

We also notice that the first three terms in (2.21) represent the first three components of the momentum equation and can be rewritten as

$$\frac{-\alpha_4}{\beta_0}\left(\sqrt{\beta_1^2 + \beta_2^2 + \beta_2^2}\right)\mathbf{n} \cdot \left[\frac{D\mathbf{u}}{Dt} + \frac{1}{\rho}\nabla p\right]. \tag{2.23}$$

The second collection of terms in (2.21) can be rewritten as

$$\alpha_4 \left[\frac{1}{\rho}\frac{D\rho}{Dt} + \nabla \cdot \mathbf{u} + \frac{1}{\rho c^2}\left(\frac{Dp}{Dt} - c^2 \frac{D\rho}{Dt}\right)\right], \tag{2.24}$$

and the right-hand side of (2.21) is

$$\frac{\alpha_4}{\rho c^2} b_5 = \alpha_4 \sigma \omega. \tag{2.25}$$

Putting it all together leads to the frame-invariant expression of the compatibility condition on the characteristic surface (canceling out the common α_4 and the material derivatives of density, and multiplying through by ρc^2),

$$-\frac{\sqrt{\beta_1^2 + \beta_2^2 + \beta_2^2}}{\beta_0}(\rho c^2)\mathbf{n} \cdot \left[\frac{D\mathbf{u}}{Dt} + \frac{1}{\rho}\nabla p\right] + \left[\rho c^2 \nabla \cdot \mathbf{u} + \frac{Dp}{Dt}\right] = \rho c^2 \sigma \omega. \tag{2.26}$$

The characteristic equations (2.17) for the directions show that

$$\frac{\sqrt{\beta_1^2 + \beta_2^2 + \beta_2^2}}{\beta_0} = \pm \frac{1}{c}, \tag{2.27}$$

so that it can be used to write the compatibility condition in the form

$$\mp(\rho c)\mathbf{n} \cdot \left[\frac{D\mathbf{u}}{Dt} + \frac{1}{\rho}\nabla p\right] + \left[\rho c^2 (\nabla \cdot \mathbf{u}) + \frac{Dp}{Dt}\right] = \rho c^2 \sigma \omega. \tag{2.28}$$

The compatibility condition is a differential relation that holds on the characteristic surface. But the other condition is that the motion is confined to be along the space-time characteristic direction defined by *speed relation*

$$u\beta_1 + \beta_4 = \pm c\sqrt{\beta_1^2 + \beta_2^2 + \beta_3^2}\,. \tag{2.29}$$

It is important to interpret (2.29) as well as a frame-invariant relation. The components $(\beta_1, \beta_2, \beta_3)$ can be chosen to be those of a unit normal to the surface, and hence $\sqrt{\beta_1^2 + \beta_2^2 + \beta_3^2} = 1$. Also the term $u\beta_1$ has the meaning $\mathbf{u} \cdot \mathbf{n}$. Finally, β_4 is the velocity of the characteristic surface normal to itself, $\beta_4 = V_n$ (say). Rewriting the expression above leads to

$$V_n = \mathbf{u} \cdot \mathbf{n} \pm c\,. \tag{2.30}$$

In one dimension, this reduces to the familiar equation for the slope of the characteristics $V_n \equiv dx/dt = u \pm c$.

Consider the forward propagating surface that corresponds to the choice of the plus sign in the previous relation (2.30). Note that the particle velocity in the frame of an observer traveling in the forward surface is $u_n - V_n$ and the speed relation can be written as

$$\frac{u_n - V_n}{c} = -1\,. \tag{2.31}$$

This means that on this characteristic surface the local normal Mach number is always unity, which is the conventional definition of sonic.

The compatibility and the speed relation, taken together, are two pieces of information, namely a differential condition in the sonic surface and a scalar speed relation, that determine the motion of the surface. If we include additional reactions and replace λ by λ_q, $q = 1, 2, \ldots, N$, where N is the number of reactions, then in the subsequent derivations only the right-hand side of (2.28) will change since additional reactions generate only additional roots that are multiples of the root associated with the streamline characteristic but not to the acoustics. The right-hand side of the compatibility condition becomes the sum, $\rho c^2 \sigma_q \omega_q$, over $q = 1, \ldots, N$, where

$$\sigma_q = -\frac{1}{\rho c^2}\frac{e_{\lambda_q}}{e_p} \tag{2.32}$$

is the thermicity coefficient and ω_q is the rate of qth reaction. The sound speed in the governing equations is the frozen sound speed and is still given by (2.13).

If we specify the result to a detonation wave that is propagating from left to right in the positive x-direction, then the normal to the characteristic surface embedded in the reaction zone, which can possibly intersect the shock, points forward. Therefore, we select the plus sign in (2.28). Let us denote the unit normal to the characteristic surface $\mathbf{n}_*$ (in general, the subscript $*$ will refer to a quantity evaluated at the sonic surface). The *compatibility condition* for this surface is then

$$\rho c \mathbf{n}_* \cdot \left(\frac{D\mathbf{u}}{Dt} + \frac{1}{\rho}\nabla p\right) + \rho c^2 \nabla \cdot \mathbf{u} + \frac{Dp}{Dt} = \rho c^2 \sigma \omega, \tag{2.33}$$

where it is understood that all terms are evaluated at the sonic surface, although we drop the subscript $*$ in most of the terms for the sake of clarity. The compatibility condition (2.33) holds on the exceptional surface at which the flow is locally sonic; that is, an observer moving with the surface observes that the flow speed normal to the surface is locally sonic:

$$\mathcal{U}_{n*} = \mathbf{u}_* \cdot \mathbf{n}_* - \mathcal{D} = -c_*\,, \tag{2.34}$$

where $\mathcal{D}$ is the normal speed of the sonic surface in the lab frame.

3. One-dimensional sonic conditions. Equation (2.33) simplifies now to

$$\rho c\left(\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{1}{\rho}\frac{\partial p}{\partial x}\right) + \rho c^2 \frac{\partial u}{\partial x} + \frac{\partial p}{\partial t} + u\frac{\partial p}{\partial x} = \rho c^2 \sigma \omega, \tag{3.1}$$

which can be rewritten as

$$\frac{dp_*}{dt} + \rho_* c_* \frac{du_*}{dt} = \rho_* c_*^2 \sigma_* \omega_*, \tag{3.2}$$

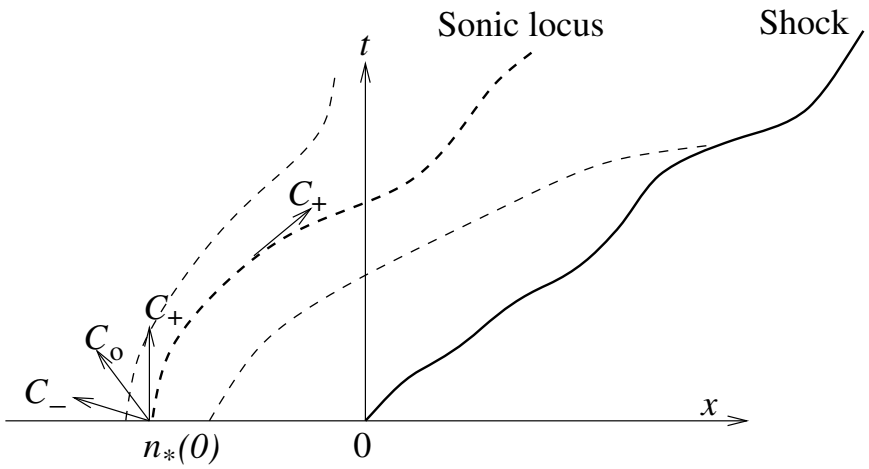


FIG. 3.1. *One-dimensional sonic locus as the C_+ characteristic emanating from the initial steady sonic locus.*

where the spatial and temporal derivatives in (3.1) are combined to form a time derivative along the forward characteristic direction,

$$\frac{d}{dt} = \frac{\partial}{\partial t} + (c_* + u_*)\frac{\partial}{\partial x} = \frac{\partial}{\partial t} + \frac{dx_*}{dt}\frac{\partial}{\partial x},$$

$$\frac{dx_*}{dt} = c_* + u_*. \tag{3.3}$$

As we have mentioned before, the sonic locus is a special characteristic that is a separatrix of two families of characteristic lines, namely those that reach the shock front in finite time and those that do not. It is assumed that the sonic locus exists initially as, for example, in a steady detonation and continues to exist during unsteady evolution. Then the initial condition selects the separatrix from the entire family of forward characteristics for all of which (3.3) and (3.2) hold. It must be pointed out that it is not, in general, possible to identify the separatrix in an arbitrary initial condition.

One can look at (3.2) as a differential equation that does not involve derivatives normal to the characteristic surface. The sonic locus is an (x, t)-curve along a limiting C_+ characteristic (see Figure 3.1), and the derivative $\partial/\partial x$ does not appear. Indeed, the time derivatives in (3.2) are the derivatives along the characteristics; that is, the derivatives lie in the tangent plane of the characteristic surface.

For one-dimensional detonation with point symmetry ($j = 0, 1, 2$ correspond to planar, cylindrical, and spherical symmetry, respectively), one easily finds that the compatibility condition is

$$\frac{dp_*}{dt} + \rho_* c_* \frac{du_*}{dt} + \frac{j}{r}\rho_* c_*^2 u_* = \rho_* c_*^2 \sigma_* \omega_*, \tag{3.4}$$

where r is the radial coordinate, while the speed relation is

$$\frac{dr_*}{dt} = c_* + u_*. \tag{3.5}$$

For a steady one-dimensional planar detonation wave in a mixture with complex reaction network, the compatibility condition reduces to the equation

$$\sigma_q \omega_q = 0 \tag{3.6}$$

that, together with $c_* + U_* = 0$, defines the sonic locus. For a discussion of the condition in applications to multiple-step reactions in detonation waves, see [5].

4. Sonic conditions of detonation shock dynamics. We call (2.33) and (2.34) the *sonic conditions* on the limiting forward characteristic surface, and their application to detonation theory is a main result of this paper. Specifically, we consider initial-value problems where there is an initially prescribed detonation shock locus with states behind it that lead subsequently to smooth evolution in the reaction zone for a self-sustained detonation. In this section, we specialize sonic conditions to one- and two-dimensional detonations. We show that when linearized, the compatibility condition reduces to the *radiation condition* of detonation stability theory (see, e.g., [7, 9, 13]). For the two-dimensional, slowly varying, and weakly curved detonations, the compatibility condition reduces to the *thermicity condition* of detonation shock dynamics (DSD theory; see, e.g., [18]). In both detonation stability theory and DSD, the governing equations are usually written in a frame of reference attached to the shock front since one is often interested in the shock-front dynamics rather than anything else. For the purpose of comparison with the known sonic conditions, we write our sonic conditions in the shock-attached frame. But before doing that, it is instructive to look at the sonic conditions written in the frame of the sonic locus.

4.1. Sonic conditions in the sonic-frame Bertrand coordinates. We express the sonic conditions in two-dimensional surface-attached Bertrand coordinates which use the normal distance to a prescribed front and the arclength to a reference point along the front as the intrinsic surface-based coordinates (see, e.g., [11, 18]). Since the Bertrand coordinates are developed by the sonic surface, they are perfectly suited to simplify the conditions since only derivatives in the surface and normal to that surface appear. Let (η, ζ) be the normal signed distance to the surface and transverse distance measured along the surface (see Figure 4.1). Let $(\mathbf{n}, \mathbf{t})$ be the corresponding unit normal and tangent vectors to the sonic surface. The coordinate transformation from the laboratory frame to the Bertrand frame is defined by

$$\mathbf{r} = \mathbf{r}_s + \eta \mathbf{n}, \tag{4.1}$$

where $\mathbf{r}$ is the lab-frame position of a point in space and $\mathbf{r}_s(\zeta, t)$ is the position of the sonic surface. Then various differential operators in the Bertrand frame are written as follows:

$$\nabla = \mathbf{n}\frac{\partial}{\partial \eta} + \frac{\mathbf{t}}{1 + \eta\kappa_*}\frac{\partial}{\partial \zeta}, \tag{4.2}$$

$$\nabla \cdot \mathbf{u} = \frac{\partial u_\eta}{\partial \eta} + \frac{1}{1 + \eta\kappa_*}\left(\kappa_* u_\eta + \frac{\partial u_\zeta}{\partial \zeta}\right), \qquad \mathbf{u} \cdot \nabla = u_\eta \frac{\partial}{\partial \eta} + \frac{u_\zeta}{1 + \eta\kappa_*}\frac{\partial}{\partial \zeta}, \tag{4.3}$$

$$\frac{\partial}{\partial t} = \frac{\partial}{\partial t} - \mathcal{D}\frac{\partial}{\partial \eta} + \mathcal{S}\frac{\partial}{\partial \zeta}, \tag{4.4}$$

and

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{u} \cdot \nabla = \frac{\partial}{\partial t} + (u_\eta - \mathcal{D})\frac{\partial}{\partial \eta} + \left(\mathcal{S} + \frac{u_\zeta}{1 + \eta\kappa_*}\right)\frac{\partial}{\partial \zeta}, \tag{4.5}$$

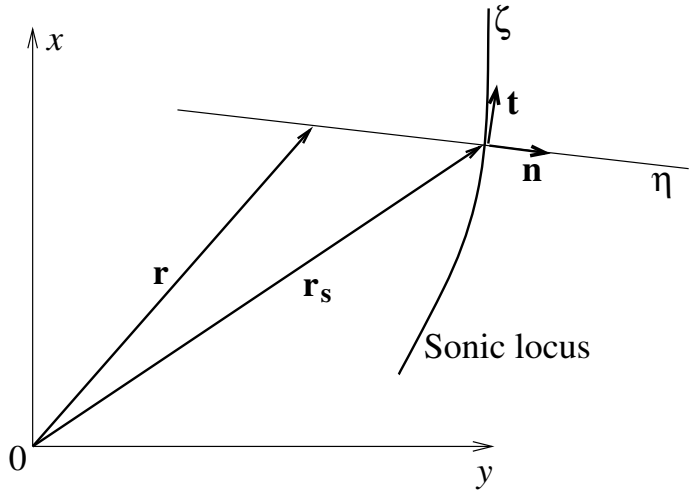


FIG. 4.1. *Bertrand frame attached to the sonic locus.*

where the lab-frame particle speed is $\mathbf{u} = u_\eta \mathbf{n} + u_\zeta \mathbf{t}$, and κ_* is the curvature of the sonic surface. Note that $\eta = 0$ in the sonic surface and that $u_\eta - \mathcal{D} = \mathcal{U}_\eta$ is the normal particle velocity relative to the sonic frame. We introduced the rate of strain of the arclength,

$$\mathcal{S} = \frac{\partial \zeta}{\partial t}, \tag{4.6}$$

and used the fact that

$$\frac{\partial \eta}{\partial t} = -\mathcal{D}. \tag{4.7}$$

Next we calculate the compatibility condition (2.33) in terms of the new coordinates. Clearly, $\mathbf{n} \cdot \nabla p = \partial p / \partial \eta$, and all other terms are also straightforward, except for $\mathbf{n} \cdot D\mathbf{u}/Dt$. To calculate the latter, we write

$$\begin{aligned} \mathbf{n} \cdot \frac{D\mathbf{u}}{Dt} &= \mathbf{n} \cdot \frac{D}{Dt}\left(u_\eta \mathbf{n} + u_\zeta \mathbf{t}\right) = \mathbf{n} \cdot \left(\frac{Du_\eta}{Dt}\mathbf{n} + u_\eta \frac{D\mathbf{n}}{Dt} + \frac{Du_\zeta}{Dt}\mathbf{t} + u_\zeta \frac{D\mathbf{t}}{Dt} \right) \\ &= \frac{Du_\eta}{Dt} + u_\zeta \mathbf{n} \cdot \left[\frac{\partial \mathbf{t}}{\partial t} + (u_\eta - \mathcal{D}) \frac{\partial \mathbf{t}}{\partial \eta} + (\mathcal{S} + u_\zeta) \frac{\partial \mathbf{t}}{\partial \zeta} \right], \end{aligned} \tag{4.8}$$

where we have used (4.5) and $\mathbf{n} \cdot \mathbf{t} = 0, \quad \mathbf{n} \cdot D\mathbf{n}/Dt = 0$. To determine $\mathbf{n} \cdot \partial \mathbf{t}/\partial t$, we differentiate the coordinate transformation, $\mathbf{r} = \mathbf{r}_s + \eta \mathbf{n}$, with respect to time and find

$$0 = \frac{d\mathbf{r}}{dt} = \frac{\partial \mathbf{r}_s}{\partial t} + \frac{\partial \zeta}{\partial t}\frac{\partial \mathbf{r}_s}{\partial \zeta} + \frac{\partial \eta}{\partial t}\mathbf{n} + \eta \left(\frac{\partial \mathbf{n}}{\partial t} + \frac{\partial \zeta}{\partial t}\frac{\partial \mathbf{n}}{\partial \zeta} \right). \tag{4.9}$$

We evaluate the last result in the sonic surface, at $\eta = 0$, to obtain

$$\frac{\partial \mathbf{r}_s}{\partial t} + \mathcal{S}\mathbf{t} - \mathcal{D}\mathbf{n} = 0, \tag{4.10}$$

and differentiate the latter with respect to ζ, and noting that $\mathbf{t} = \partial \mathbf{r}_s/\partial\zeta$, we find, using the Frenet formulas,

$$\frac{\partial \mathbf{n}}{\partial \zeta} = \kappa \mathbf{t}, \qquad \frac{\partial \mathbf{t}}{\partial \zeta} = -\kappa \mathbf{n}, \tag{4.11}$$

that

$$\mathbf{n} \cdot \frac{\partial \mathbf{t}}{\partial t} = \frac{\partial \mathcal{D}}{\partial \zeta} + \kappa \mathcal{S}. \tag{4.12}$$

Then, collecting all terms in (4.8), we find that

$$\mathbf{n} \cdot \frac{D\mathbf{u}}{Dt} = \frac{Du_\eta}{Dt} + u_\zeta \frac{\partial \mathcal{D}}{\partial \zeta} - \kappa u_\zeta^2. \tag{4.13}$$

What is left is to collect terms in (2.33), which results in the following equation:

$$\begin{aligned} &\rho c \left(\frac{Du_\eta}{Dt} + u_\zeta \frac{\partial \mathcal{D}}{\partial \zeta} - \kappa u_\zeta^2 + \frac{1}{\rho}\frac{\partial p}{\partial \eta} \right) \\ &+\rho c^2 \left(\frac{\partial u_\eta}{\partial \eta} + \kappa_* u_\eta + \frac{\partial u_\zeta}{\partial \zeta} \right) + \frac{Dp}{Dt} = \rho c^2 \sigma \omega. \end{aligned} \tag{4.14}$$

Expanding the material derivative according to (4.5) and rearranging derivatives along the same directions, we obtain

$$\begin{aligned} &\frac{\partial p}{\partial t} + \rho c \frac{\partial u_\eta}{\partial t} + \kappa_* \rho c^2 u_\eta + (c + u_\eta - \mathcal{D}) \left(\frac{\partial p}{\partial \eta} + \rho c \frac{\partial u_\eta}{\partial \eta} \right) \\ &+\rho c^2 \frac{\partial u_\zeta}{\partial \zeta} + \rho c u_\zeta \left(\frac{\partial \mathcal{D}}{\partial \zeta} - \kappa_* u_\zeta \right) + (\mathcal{S} + u_\zeta) \left(\frac{\partial p}{\partial \zeta} + \rho c \frac{\partial u_\eta}{\partial \zeta} \right) = \rho c^2 \sigma \omega. \end{aligned} \tag{4.15}$$

An important observation now is that in the sonic surface the flow is locally sonic with

$$c + u_\eta - \mathcal{D} = 0, \tag{4.16}$$

which is the speed relation. Therefore, all normal-derivative terms in the compatibility condition (4.15) drop out, resulting in

$$\frac{\partial p}{\partial t} + \rho c \frac{\partial u_\eta}{\partial t} + \kappa_* \rho c^2 u_\eta = \rho c^2 \sigma \omega - R_*, \tag{4.17}$$

where the terms that explicitly depend on the transverse variation are lumped into R_*, given by

$$R_* = \rho c^2 \frac{\partial u_\zeta}{\partial \zeta} + \rho c u_\zeta \left(\frac{\partial \mathcal{D}}{\partial \zeta} - \kappa_* u_\zeta \right) + (\mathcal{S} + u_\zeta) \left(\frac{\partial p}{\partial \zeta} + \rho c \frac{\partial u_\eta}{\partial \zeta} \right).$$

The reader is reminded that everything in (4.17) is evaluated in the sonic surface.

By definition, the compatibility condition must not contain derivatives along the normal to the characteristic surface in (ζ, η, t)-space. Since our coordinate frame is local, that is, attached to the characteristic surface, then the time derivative in (4.17) does indeed lie in the surface, similar to the time derivative along the C_+

characteristic in one dimension. Furthermore, the ζ-derivative is also in the surface, as ζ is the arclength. The only derivative that is *off* the characteristic surface in (ζ, η, t)-space is $\partial/\partial\eta$, and that derivative is indeed absent in (4.17). If R_* can be neglected, (4.17) is similar to the thermicity condition of the old DSD theories with an important difference that here $\mathcal{U}_\eta$ and $\mathcal{D}$ are the particle velocity in the *sonic* frame and normal speed of the *sonic* surface, respectively; in the older theories of DSD, the same variables are calculated in the *shock-attached* frame. The approximate form that neglects R_* is valid only in the limit of weak curvature, slow time, and small transverse variation. Equation (4.17) is an exact relation that is valid for general two-dimensional detonations with an embedded sonic surface, provided only that the Bertrand coordinates are invertible, which is true if the radius of curvature of the sonic locus is large compared to the length of the reaction zone.

4.2. Sonic conditions of DSD theory: Formulation in the shock-attached frame. The linear stability problem and the DSD problem were originally formulated in shock-attached coordinates: in the first case, this dates back to the first rigorous analysis given by Erpenbeck [4]; in the second case, the shock-attached coordinates were used because the goal of DSD theory is to determine the dynamics of the shock front [16, 18].

Here we revisit the formulation of DSD in the shock-attached coordinates and use Bertrand coordinates attached to the shock. Let (n, ξ) be the normal and transverse coordinates, and let $(\mathbf{n}, \mathbf{t})$ represent the corresponding unit normal and tangent vectors in the shock frame; then the coordinate transformation is given by

$$\mathbf{r} = \mathbf{r}_s(\xi, t) + n\mathbf{n}(\xi, t). \tag{4.18}$$

The time derivative in the shock-attached frame is represented as

$$\frac{\partial}{\partial t} = \frac{\partial}{\partial t} - D\frac{\partial}{\partial n} + S\frac{\partial}{\partial \xi},$$

the velocity in the lab frame is $\mathbf{u} = u_n\mathbf{n} + u_\xi\mathbf{t}$, D is the normal shock speed, $S = \partial\xi/\partial t$ is the stretch rate of the arclength along the shock, and $U_n = u_n - D$ is the normal particle speed relative to the shock. The material derivative is then

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + U_n\frac{\partial}{\partial n} + \left(S + \frac{u_\xi}{1 + n_*\kappa}\right)\frac{\partial}{\partial \xi}. \tag{4.19}$$

Differential operators involving ∇ are similar to those in the sonic frame, (4.2)–(4.3), only now the velocity is expressed in the shock frame. A slight complication arises from the fact that $\mathbf{n}_*$ in (2.33) is the unit normal to the sonic surface, which in general is different from $\mathbf{n}$, the unit normal to the shock. Therefore, the shock-frame compatibility condition will contain terms, proportional to $\mathbf{n}_* \cdot \mathbf{n}$, which need to be evaluated.

Let

$$\mathbf{n}_* = a_n\mathbf{n} + a_\xi\mathbf{t}, \tag{4.20}$$

where the components, $a_n = \mathbf{n}_* \cdot \mathbf{n}$ and $a_\xi = \mathbf{n}_* \cdot \mathbf{t}$, will be determined below (see equations (4.32)). Then, $\mathbf{n}_* \cdot \nabla p = a_n\partial p/\partial n + a_\xi\partial p/\partial\xi$, and

$$\mathbf{n}_* \cdot \frac{D\mathbf{u}}{Dt} = \mathbf{n}_* \cdot \frac{D}{Dt}(u_n\mathbf{n} + u_\xi\mathbf{t}) = \frac{Du_n}{Dt}\mathbf{n}_* \cdot \mathbf{n} + u_n\mathbf{n}_* \cdot \frac{D\mathbf{n}}{Dt} + \frac{Du_\xi}{Dt}\mathbf{n}_* \cdot \mathbf{t} + u_\xi\mathbf{n}_* \cdot \frac{D\mathbf{t}}{Dt}. \tag{4.21}$$

We now calculate each term on the right-hand side of this equation. Consider

$$\mathbf{n}_* \cdot \frac{D\mathbf{n}}{Dt} = a_\xi \mathbf{t} \cdot \frac{D\mathbf{n}}{Dt} = a_\xi \mathbf{t} \cdot \left[\frac{\partial \mathbf{n}}{\partial t} + \left(S + \frac{u_\xi}{1 + n_* \kappa} \right) \frac{\partial \mathbf{n}}{\partial \xi} \right]. \tag{4.22}$$

By time-differentiating the coordinate transformation (4.18) and evaluating the result at the shock, we find that

$$\frac{\partial \mathbf{r}_s}{\partial t} + S\mathbf{t} - D\mathbf{n} = 0. \tag{4.23}$$

Differentiating this result with respect to ξ and using $\mathbf{t} = \partial \mathbf{r}_s / \partial \xi$, we find

$$\frac{\partial \mathbf{t}}{\partial t} + \left(\frac{\partial S}{\partial \xi} - \kappa D \right) \mathbf{t} - \left(\frac{\partial D}{\partial \xi} + \kappa S \right) \mathbf{n} = 0, \tag{4.24}$$

from which it follows that

$$\mathbf{t} \cdot \frac{\partial \mathbf{n}}{\partial t} = -\mathbf{n} \cdot \frac{\partial \mathbf{t}}{\partial t} = -\frac{\partial D}{\partial \xi} - \kappa S \tag{4.25}$$

and

$$\frac{\partial S}{\partial \xi} - \kappa D = 0. \tag{4.26}$$

Using (4.25) and the Frenet formula, $\partial \mathbf{n} / \partial \xi = \kappa \mathbf{t}$, we find that (4.22) results in

$$\mathbf{n}_* \cdot \frac{D\mathbf{n}}{Dt} = a_\xi \left(-\frac{\partial D}{\partial \xi} + \frac{\kappa u_\xi}{1 + n_* \kappa} \right). \tag{4.27}$$

Similarly, we find

$$\mathbf{n}_* \cdot \frac{D\mathbf{t}}{Dt} = a_n \mathbf{n} \cdot \left[\frac{\partial \mathbf{t}}{\partial t} + \left(S + \frac{u_\xi}{1 + n_* \kappa} \right) \frac{\partial \mathbf{t}}{\partial \xi} \right] = a_n \left(\frac{\partial D}{\partial \xi} - \frac{\kappa u_\xi}{1 + n_* \kappa} \right). \tag{4.28}$$

Equation (4.21) becomes

$$\mathbf{n}_* \cdot \frac{D\mathbf{u}}{Dt} = a_n \frac{Du_n}{Dt} + a_\xi \frac{Du_\xi}{Dt} + (u_n a_\xi - u_\xi a_n) \left(-\frac{\partial D}{\partial \xi} + \frac{\kappa u_\xi}{1 + n_* \kappa} \right). \tag{4.29}$$

Collecting all terms, we obtain that the shock-frame compatibility condition is

$$\frac{\partial p}{\partial t} + (c + U_n) \frac{\partial p}{\partial n} + \rho c \left[\frac{\partial u_n}{\partial t} + (c + U_n) \frac{\partial u_n}{\partial n} \right] + \frac{\kappa}{1 + n_* \kappa} \rho c^2 u_n = \rho c^2 \sigma \omega - R, \tag{4.30}$$

where κ (without the * subscript) is the local curvature of the shock, n_* is the normal distance from the shock to the sonic surface, and all terms are evaluated in the sonic surface. By R in the right-hand side of (4.30) we denote the following collection of terms:

$$\begin{aligned} R = & \left(S + \frac{u_\xi}{1 + n_* \kappa} \right) \left(\frac{\partial p}{\partial \xi} + \rho c \frac{\partial u_n}{\partial \xi} \right) + \frac{\rho c^2}{1 + n_* \kappa} \frac{\partial u_\xi}{\partial \xi} \\ & + c (a_n - 1) \left(\frac{\partial p}{\partial n} + \rho \frac{Du_n}{Dt} \right) + c a_\xi \left(\frac{\partial p}{\partial \xi} + \rho \frac{Du_\xi}{Dt} \right) \\ & + \rho c (u_n a_\xi - u_\xi a_n) \left(-\frac{\partial D}{\partial \xi} + \frac{\kappa u_\xi}{1 + n_* \kappa} \right). \end{aligned} \tag{4.31}$$

From the derivations below (see (4.39)), the coefficients a_n and a_ξ in (4.31) are given by

$$a_n = \left[1 + \left(\frac{1}{1+n_*\kappa}\frac{\partial n_*}{\partial \xi}\right)^2\right]^{-1/2}, \qquad a_\xi = -\frac{a_n}{1+n_*\kappa}\frac{\partial n_*}{\partial \xi}, \tag{4.32}$$

so that small transverse variation implies smallness of $a_n - 1$ and a_ξ.

Note that the operator $\partial/\partial t + (c + U_n)\,\partial/\partial n$ in (4.30) in general is *not* the time derivative along the sonic locus, unlike the one in (4.17). In the sonic frame, we had $\mathcal{U}_{n*} = -c_*$ exactly as a speed relation. But now it is no longer true that $c_* + U_{n*} = 0$! In one dimension, we could write $c_* + U_{n*} = dn_*/dt$, in which case the operator $\partial/\partial t + (c + U_n)\,\partial/\partial n$ does indeed become a total derivative along the sonic locus. But in general two-dimensional detonation waves, the derivative $\partial/\partial t + (c + U_n)\,\partial/\partial n$ does not lie in the tangent plane of the sonic locus; only if the transverse variations can be neglected is the derivative in the sonic surface.

The speed relation expressed in the shock-attached coordinates is derived next. Let the equation

$$\psi(x, y, t) = 0 \tag{4.33}$$

represent the level set of the sonic surface in the laboratory frame. Then its unit normal and normal speed are given by

$$\mathbf{n}_* = \frac{\nabla\psi}{|\nabla\psi|} \quad \text{and} \quad \mathcal{D} = -\frac{1}{|\nabla\psi|}\frac{\partial\psi}{\partial t}, \tag{4.34}$$

respectively, so that the general speed relation (2.34) can be rewritten as

$$\frac{\partial\psi}{\partial t} + c\,|\nabla\psi| + \mathbf{u}\cdot\nabla\psi = 0. \tag{4.35}$$

An interesting form of the speed relation is obtained from (4.35) by noting that $|\nabla\psi| = \mathbf{n}_*\cdot\nabla\psi$,

$$\frac{\partial\psi}{\partial t} + (\mathbf{u} + c\,\mathbf{n}_*)\cdot\nabla\psi = 0, \tag{4.36}$$

a transport equation that underscores propagation of the sonic surface in the direction of $\mathbf{u} + c\mathbf{n}_*$ with the normal speed $c + \mathbf{u}\cdot\mathbf{n}_*$. The derivative $\mathcal{L} = \partial/\partial t + (\mathbf{u} + c\mathbf{n}_*)\cdot\nabla$ is a directional time derivative normal to the sonic surface so that (4.36) is an expression of constancy of ψ in the sonic surface.

In the shock-attached frame, (n, ξ, t), the level-set equation can be written as

$$\psi \equiv n - n_*(\xi, t) = 0, \tag{4.37}$$

where n_* is the normal distance from the shock to the sonic surface. Then we obtain that

$$\nabla\psi = \mathbf{n} - \frac{1}{1+n_*\kappa}\frac{\partial n_*}{\partial \xi}\mathbf{t}, \qquad \frac{\partial\psi}{\partial t} = -D - \frac{\partial n_*}{\partial t} - S\frac{\partial n_*}{\partial \xi}, \tag{4.38}$$

and

$$\mathbf{n}_* = \frac{1}{|\nabla\psi|}\left(\mathbf{n} - \frac{1}{1+n_*\kappa}\frac{\partial n_*}{\partial \xi}\mathbf{t}\right). \tag{4.39}$$

Be reminded that in these expressions κ is the curvature of the shock. Substituting these formulas into (4.35), we obtain the speed relation in the shock-attached frame,

$$\frac{\partial n_*}{\partial t} + \left(S + \frac{u_\xi}{1+n_*\kappa}\right)\frac{\partial n_*}{\partial \xi} = U_n + c\sqrt{1+\left(\frac{1}{1+n_*\kappa}\frac{\partial n_*}{\partial \xi}\right)^2}. \tag{4.40}$$

Again, this is an exact relation that expresses the speed relation for the sonic surface written in the shock-attached Bertrand coordinates in terms of the shock properties, that is, the curvature κ and the stretch S, and the flow state in the sonic surface, $n_*(\xi,t)$, U_n, u_ξ, and c. Thus we have two equations, (4.30) and (4.40), that represent the sonic conditions in the shock-attached Bertrand frame.

Equation (4.40) can be rewritten as

$$c + U_n = \frac{\partial n_*}{\partial t} + \left(S + \frac{u_\xi}{1+n_*\kappa}\right)\frac{\partial n_*}{\partial \xi} + c\left[1 - \sqrt{1+\left(\frac{1}{1+n_*\kappa}\frac{\partial n_*}{\partial \xi}\right)^2}\right], \tag{4.41}$$

from which one can see that the speed relation is similar to the equation of the forward characteristic in one dimension, (3.3), which in the shock-attached frame is $c + U_n = dn_*/dt$ but involves more terms, all due to the transverse variation.

Next we make certain approximations in order to simplify the sonic conditions (4.30) and (4.40) and to see their connection with the older formulations of DSD. Let us assume that the shock curvature is small, $\kappa = o(1)$, and the transverse flow speed and transverse variations are also small, $u_\xi = o(1)$, $\partial/\partial\xi = o(1)$. Then retaining only the leading-order terms, from (4.40), we obtain that

$$\frac{\partial n_*}{\partial t} = U_n + c_*. \tag{4.42}$$

Retaining only leading-order curvature terms in (4.30), we obtain that

$$\frac{\partial p}{\partial t} + \rho c\frac{\partial u_n}{\partial t} + \kappa\rho c^2 u_n - \rho c^2\sigma\omega = 0, \tag{4.43}$$

where the time derivative is now

$$\frac{\partial}{\partial t} = \frac{\partial}{\partial t} + (c+U_n)\frac{\partial}{\partial n} = \frac{\partial}{\partial t} + \frac{\partial n_*}{\partial t}\frac{\partial}{\partial n}.$$

The time derivative in (4.43) must be taken along the sonic locus; that is, the state variables, p and u_n, must first be evaluated at the sonic locus, and only then should their derivatives be taken.

5. On the sonic conditions of detonation stability theory. In this section, we show that the linearized version of the compatibility condition reduces to the radiation conditions of detonation stability theory (see, e.g., [7, 9, 13]). Here we derive the one- and two-dimensional radiation conditions.

A one-dimensional radiation condition follows directly from (3.2) by straightforward linearization. Let us denote the steady base state by an overbar and perturbations about the base state by a prime, e.g., $p = \bar{p}(n) + p'(n,t)$, etc. Then the perturbed sonic state is given by

$$p_* = \bar{p}_*(n_*) + p'(n_*,t), \qquad u_* = \bar{u}_*(n_*) + u'(n_*,t), \tag{5.1}$$

$$\rho_* = \bar{\rho}_*(n_*) + \rho'(n_*,t), \qquad \lambda_* = \bar{\lambda}_*(n_*) + \lambda'(n_*,t), \tag{5.2}$$

where we can take $n_* = \bar{n}_*$ in the primed quantities since the correction to the sonic locus, $n'_* = n_* - \bar{n}_*$, that results from the use of the speed relation,

$$\dot{n}'_* = c'_* + U'_{n*}, \tag{5.3}$$

contributes only higher-order terms. But we also expand the leading-order terms about the exact sonic locus to obtain, for example, that

$$\bar{p}_* (n_*) = \bar{p}_* (\bar{n}_*) + \frac{\mathrm{d}\bar{p}_* (\bar{n}_*)}{\mathrm{d}n} n'_*. \tag{5.4}$$

The perturbations such as in the last expression will be absent if the steady-state gradients vanish at the steady sonic locus, which is often the case.

Finally, the linearized compatibility condition is

$$\frac{dp'}{dt} + \bar{\rho}_* \bar{c}_* \frac{du'}{dt} + \left(\frac{\mathrm{d}\bar{p}_*}{\mathrm{d}n} + \bar{\rho}_* \bar{c}_* \frac{\mathrm{d}\bar{p}_*}{\mathrm{d}n} \right) \dot{n}'_* = \bar{\rho}_* \bar{c}_*^2 \bar{\sigma}_* \omega', \tag{5.5}$$

where everything with an overbar is evaluated at $n = \bar{n}_*$. We have also taken into account that $\bar{\omega}_* = 0$; ω' is the perturbation of the reaction rate.

In the special case of an ideal gas, the equation of state is $p = \rho RT$, $e = pv/(\gamma - 1) - \lambda Q$, so that $\rho c^2 \sigma = (\gamma - 1) Q\rho$. For simple-depletion kinetics with $\nu = 1$, the gradients of the steady-state pressure and velocity vanish at the sonic locus, and therefore the term proportional to $\dot{n}'_*$ in (5.5) will drop out. Assuming normal-mode perturbations, $p' = \bar{p}'(n) \exp(\alpha t)$, etc., the radiation condition (5.5) reduces to

$$\alpha (\bar{p}'_* + \bar{\rho}_* \bar{c}_* \bar{u}'_*) + (\gamma - 1) Q \bar{\rho}_* k \exp(-E/\bar{p}_* \bar{v}_*) \bar{\lambda}'_* = 0, \tag{5.6}$$

which is exactly the CJ limit of the radiation condition derived by Lee and Stewart [9].

If the depletion factor is less than unity, that is, $\nu < 1$ in $\omega = k(1 - \lambda)^\nu \exp(-E/pv)$, then the reaction-rate perturbation away from the sonic locus is

$$\omega' = \left(\frac{\partial \bar{\omega}}{\partial \bar{p}} \right) p' + \left(\frac{\partial \bar{\omega}}{\partial \bar{v}} \right) v' + \left(\frac{\partial \bar{\omega}}{\partial \bar{\lambda}} \right) \lambda'. \tag{5.7}$$

As $\bar{\lambda} \to 1$, one finds that $(\partial \bar{\omega}/\partial \bar{\lambda}) \sim (1 - \bar{\lambda})^{\nu - 1} \to \infty$, so the last term in the previous expansion is nonuniform as the sonic locus is approached, clearly a result of the base-state reaction rate vanishing at the sonic locus. Near the sonic locus the reaction rate perturbation is

$$\omega' = \omega (\lambda_*) - \omega \left(\bar{\lambda}_* \right) = k (-\lambda')^\nu \exp(-E/\bar{p}_* \bar{v}_*), \tag{5.8}$$

which is a *nonlinear* function of λ', another indication of the nonuniformity of solutions of the original linearized system of Euler equations. If all perturbations in expansions (5.1) and (5.2) are assumed to be $O(\epsilon)$ with $\epsilon \to 0$, then the left-hand side of (5.5) is also $O(\epsilon)$, while the right-hand side is $O(\epsilon^\nu)$. It follows then that although in the main-reaction layer (i.e., the region behind the shock but away from the sonic locus) the perturbations are $O(\epsilon)$, they are no longer $O(\epsilon)$ as the sonic locus is approached (that is, in the transonic layer). This potential nonuniformity has to be dealt with by considering the linear stability problem separately in the main-reaction layer and the transonic layer, a problem that is beyond the scope of the present paper. Here we indicate only the possibility of essentially nonlinear dynamics in the transonic

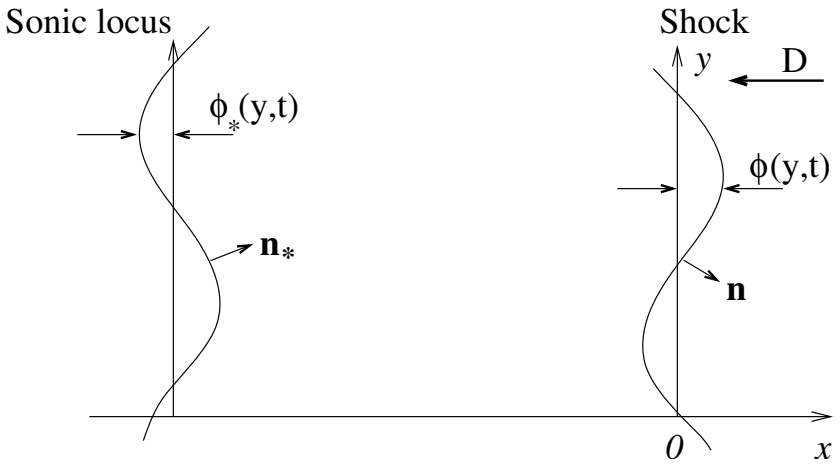


FIG. 5.1. *Perturbation of a two-dimensional steady detonation with an embedded sonic locus.*

layer, a situation common in transonic-flow problems. The linear stability problem of detonation has to be formulated so that this nonlinear character is carefully accounted for, and the solutions in the main reaction layer and transonic layer should be properly matched.

Consider now a two-dimensional detonation wave with an embedded sonic locus subject to a small perturbation of the shock locus, $\phi(y,t)$, as shown in Figure 5.1. Most treatments of detonation stability employ a Cartesian frame of reference attached to the perturbed shock so that the coordinate transformation from the lab frame is

$$x = x^l - Dt - \phi\left(y^l, t\right), \qquad y = y^l. \tag{5.9}$$

Here x^l and y^l are the lab-frame coordinates, D is the steady-state detonation speed, and ϕ is the small shock displacement in the x-direction. Thus the shock is always fixed at $x = 0$ and the reaction zone is at $x < 0$, while the unperturbed medium is at $x > 0$. The differential operators in the moving frame are now

$$\nabla = \frac{\partial}{\partial x}\mathbf{i} + \left(\frac{\partial}{\partial y} - \phi_y \frac{\partial}{\partial x}\right)\mathbf{j} \quad \text{and} \quad \frac{D}{Dt} = \frac{\partial}{\partial t} + U_1\frac{\partial}{\partial x} + u_2\frac{\partial}{\partial y} - u_2\phi_y\frac{\partial}{\partial x}, \tag{5.10}$$

where $U_1 = u_1 - D - \partial\phi/\partial t$ and u_2 are the x- and y- components of the particle speed relative to the perturbed shock, respectively.

Notice that the displacement of the sonic locus, $\phi_*(y,t)$, is not the same as ϕ and therefore, the unit normal, $\mathbf{n}_*$, to the sonic locus differs from $\mathbf{n}$, the unit normal to the shock. To the leading order in the displacements, the unit normals are given by

$$\mathbf{n} = \mathbf{i} - \frac{\partial\phi}{\partial y}\mathbf{j}, \qquad \mathbf{n}_* = \mathbf{i} - \frac{\partial\phi_*}{\partial y}\mathbf{j}. \tag{5.11}$$

One can show that the small transverse component of $\mathbf{n}_*$ contributes only second-order terms to the compatibility condition. Indeed, let $\phi = \phi' = o(1)$, $\phi_* = \phi'_* = o(1)$ and and linearize the state variables about the steady state, as, e.g., $p = \bar{p}(x) + p'(x,t)$, $\mathbf{u} = (\bar{u}_1 + u'_1)\,\mathbf{i} + u'_2\mathbf{j}$, etc., similar to the one-dimensional case; the primed quantities are small corrections to the base state. We have assumed that the gradients of the

steady-state variables vanish at the steady sonic locus. Otherwise one needs to retain terms such as $d\bar{p}/dx\,(\bar{x})\,x'_*$; see earlier in this section. Retaining only linear terms in perturbations, we find that

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \bar{U}_1 \frac{\partial}{\partial x} + \left(U'_1 \frac{\partial}{\partial x} + u'_2 \frac{\partial}{\partial y} \right), \tag{5.12}$$

$$\begin{aligned} \mathbf{n}_* \cdot \frac{D\mathbf{u}}{Dt} &= \left(\mathbf{i} - \frac{\partial \phi'_*}{\partial y} \mathbf{j} \right) \cdot \left[\frac{\partial}{\partial t} + \bar{U}_1 \frac{\partial}{\partial x} + \left(U'_1 \frac{\partial}{\partial x} + u'_2 \frac{\partial}{\partial y} \right) \right] (\bar{u}_1 \mathbf{i} + \mathbf{u}') \\ &= \bar{U}_1 \frac{\partial \bar{u}_1}{\partial x} + \left(\frac{\partial u'_1}{\partial t} + \bar{U}_1 \frac{\partial u'_1}{\partial x} + U'_1 \frac{\partial \bar{u}_1}{\partial x} \right), \end{aligned} \tag{5.13}$$

$$\mathbf{n}_* \cdot \nabla p = \frac{\partial \bar{p}}{\partial x} + \frac{\partial p'}{\partial x}, \qquad \nabla \cdot \mathbf{u} = \frac{\partial \bar{u}_1}{\partial x} + \left(\frac{\partial u'_1}{\partial x} + \frac{\partial u'_2}{\partial y} \right), \tag{5.14}$$

and

$$\frac{Dp}{Dt} = \bar{U}_1 \frac{\partial \bar{p}}{\partial x} + \left(\frac{\partial p'}{\partial t} + U'_1 \frac{\partial \bar{p}}{\partial x} + \bar{U}_1 \frac{\partial p'}{\partial x} \right). \tag{5.15}$$

Before linearization of the compatibility condition, it is convenient to rewrite it as

$$\rho c \left(\mathbf{n}_* \cdot \frac{D\mathbf{u}}{Dt} + c \nabla \cdot \mathbf{u} \right) + \frac{Dp}{Dt} + c \mathbf{n}_* \cdot \nabla p = \rho c^2 \sigma \omega. \tag{5.16}$$

We then find that

$$\mathbf{n}_* \cdot \frac{D\mathbf{u}}{Dt} + c \nabla \cdot \mathbf{u} = \left(\frac{\partial u'_1}{\partial t} + \bar{c} \frac{\partial u'_2}{\partial y} \right) + (c' + U'_1) \frac{\partial \bar{u}_1}{\partial x} \tag{5.17}$$

and

$$\frac{Dp}{Dt} + c \mathbf{n}_* \cdot \nabla p = \frac{\partial p'}{\partial t} + (c' + U'_1) \frac{\partial \bar{p}}{\partial x}, \tag{5.18}$$

so that the linearized compatibility condition becomes

$$\frac{\partial p'}{\partial t} + \bar{\rho}_* \bar{c}_* \frac{\partial u'_1}{\partial t} + \bar{\rho}_* \bar{c}_*^2 \frac{\partial u'_2}{\partial y} = \bar{\rho}_* \bar{c}_*^2 \bar{\sigma}_* \omega', \tag{5.19}$$

or, in terms of the normal modes ($p' \to p' \exp(\alpha t + iky)$, etc.),

$$\alpha \left(p' + \bar{\rho}_* \bar{c}_* u'_1 \right) + ik \bar{\rho}_* \bar{c}_*^2 u'_2 = \bar{\rho}_* \bar{c}_*^2 \bar{\sigma}_* \omega'. \tag{5.20}$$

If one sets the right-hand side of (5.20) to zero, then one obtains the CJ limit of the radiation condition of Short and Stewart [13]. But (5.20) is more general, as it includes a general rate term and holds for a general equation of state. Still the discussion above concerning possible nonuniformities in the transonic layer is obviously important here as well.

5.1. The compatibility condition as a boundedness condition. We now show that for detonations with depletion factor $\nu > 1/2$, the linearized compatibility condition

$$\frac{dp'}{dt} + \bar{\rho}_*\bar{c}_* \left(\frac{dU'}{dt} + \frac{dD'}{dt} \right) - (\gamma - 1)\, Q\bar{\rho}_*\omega' = 0 \tag{5.21}$$

is necessary for the linear stability problem to have solutions bounded at $n \to \bar{n}_*$. Indeed, the one-dimensional Euler equations written in the shock-attached frame

$$v_t + Uv_n - vU_n = 0, \tag{5.22}$$

$$U_t + UU_n + vp_n = -D_t, \tag{5.23}$$

$$p_t + Up_n + \gamma pU_n = (\gamma - 1)Q\rho\omega, \tag{5.24}$$

$$\lambda_t + U\lambda_n = \omega \tag{5.25}$$

can be linearized so that the following set of linear equations is obtained:

$$v'_t + \bar{U}v'_n + \bar{v}_nU' - \bar{v}U'_n - \bar{U}_nv' = 0, \tag{5.26}$$

$$U'_t + \bar{U}U'_n + \bar{U}_nU' + \bar{v}p'_n + \bar{p}_nv' = -D'_t, \tag{5.27}$$

$$p'_t + \bar{U}p'_n + \bar{p}_nU' + \gamma\bar{p}U'_n + \gamma\bar{U}_np' - (\gamma - 1)\, Q\left(\bar{\rho}\omega' + \bar{\omega}\rho'\right) = 0, \tag{5.28}$$

$$\lambda'_t + \bar{U}\lambda'_n + \bar{\lambda}_nU' = \omega', \tag{5.29}$$

where the perturbations are assumed to be small deviations from the corresponding steady-state values. Adding (5.28) and (5.27) multiplied by $\bar{\rho}\bar{c}$, one obtains

$$\begin{aligned} &\left[\frac{\partial}{\partial t} + \left(\bar{U} + \bar{c}\right)\frac{\partial}{\partial n}\right] p' + \bar{\rho}\bar{c}\left[\frac{\partial}{\partial t} + \left(\bar{U} + \bar{c}\right)\frac{\partial}{\partial n}\right](U' + D') - (\gamma - 1)\, Q\bar{\rho}\omega' \\ &\quad + \left(\bar{p}_n + \bar{\rho}\bar{c}\bar{U}_n\right)U' + \gamma\bar{U}_np' + \bar{\rho}\bar{c}\bar{p}_nv' - (\gamma - 1)\, Q\bar{\omega}\rho' = 0. \end{aligned} \tag{5.30}$$

The first two terms are seen to form time derivatives along the steady C_+ characteristic direction, $\partial/\partial t + \left(\bar{U} + \bar{c}\right)\partial/\partial n$, so that the first line of (5.30) tends to the compatibility condition in the limit $n \to \bar{n}_*$ (so that $\bar{U} + \bar{c} \to 0$). All terms in the second line vanish as $n \to \bar{n}_*$, provided that $\nu > 1/2$ (so that the spatial derivatives of the base state vanish at the sonic locus) and that all perturbations remain uniformly bounded. Thus the compatibility condition is necessarily satisfied if perturbations are bounded and $\nu > 1/2$.

6. Conclusions. In this work we have introduced a general definition of a sonic locus for multidimensional unsteady self-sustained detonation waves and discussed its properties under limiting conditions that are relevant to detonation stability theories and asymptotic theories of slowly evolving weakly curved detonations. We have shown that previously known sonic conditions of steady detonation theory, linear stability theory, and DSD are limiting cases of our generalized conditions. Self-sustained detonations are introduced as two-front phenomena with the lead shock and the limiting characteristic surface (as the sonic locus) as free boundaries. The sonic conditions that we have derived can be considered as closure equations that together with the Euler equations and Rankine–Hugoniot conditions complete the set of governing equations for self-sustained detonations.

An important ingredient of the present theory is that the sonic surface is assumed to exist initially; we simply take it as given by the initial conditions. The initial condition could be, for example, a steady detonation wave in which a sonic surface can

be defined unambiguously, and a clear exact case is that of the steady CJ detonation or a weakly perturbed detonation that corresponds to theories relevant to detonation instability or DSDs, both of which are perturbation theories that assume either deviations from a plane CJ state or weak spatial and temporal variation from plane states. Many important initial conditions, for example in initiation problems, will not have an initial sonic locus. But as the detonation forms and becomes a self-sustained wave, the sonic locus will appear somewhere in the flow. From that point on, the detonation dynamics is described by our theory, provided only that the sonic locus persists in the flow, which is the case if the flow evolution is smooth.

Appearance of strong discontinuities within the reaction zone, such as shock waves, can destroy a sonic surface, in which case the present theory may not be applicable. It is indeed the case in gas-phase detonations that strong transverse shock fronts almost always exist which can interact with the sonic surface. Yet, the situation is quite different in condensed explosives, in which smooth reaction zones are more common. In any case, the range of phenomena that the present theory can address is considerable, and even in the case of cellular detonations, the onset of cellular dynamics and propagation of weakly unstable detonations may be phenomena that the present theory is applicable to. Some applications of the theory to weakly curved and slowly evolving detonations can be found in [6] and in forthcoming papers.

Acknowledgments. We thank A. Kapila of RPI and J. Bdzil of LANL for their discussion of this work.

REFERENCES

[1] J. B. Bdzil, *Steady-state two-dimensional detonation*, J. Fluid Mech., 108 (1981), pp. 185–286.
[2] C. J. Chapman, *High Speed Flow*, Cambridge University Press, Cambridge, UK, 2000.
[3] A. N. Dremin, *Toward Detonation Theory,* Springer-Verlag, New York, 1999.
[4] J. J. Erpenbeck, *Stability of idealized one-reaction detonations*, Phys. Fluids, 7 (1964), pp. 684–696.
[5] W. Fickett and W. C. Davis, *Detonation*, University of California Press, Berkeley, CA, 1979.
[6] A. R. Kasimov, *Theory of Instability and Nonlinear Evolution of Self-Sustained Detonation Waves*, Ph.D. thesis, University of Illinois at Urbana-Champaign, Urbana, IL, 2004.
[7] A. R. Kasimov and D. S. Stewart, *Spinning instability of gaseous detonations*, J. Fluid Mech., 466 (2002), pp. 179–203.
[8] A. R. Kasimov and D. S. Stewart, *On the dynamics of self-sustained one-dimensional detonations: A numerical study in the shock-attached frame,* Phys. Fluids, 16 (2004), pp. 3566–3578.
[9] H. I. Lee and D. S. Stewart, *Calculation of linear detonation instability: One-dimensional instability of plane detonation,* J. Fluid Mech., 212 (1990), pp. 103–132.
[10] R. von Mises, *Mathematical Theory of Compressible Fluid Flow*, Academic Press, New York, 1958.
[11] M. Matalon, C. Cui, and J. K. Bechtold, *Hydrodynamic theory of premixed flames: Effect of stoichiometry, variable transport coefficients, and arbitrary reaction orders,* J. Fluid Mech., 487 (2003), pp. 179–210.
[12] L. V. Ovsiannikov, *Lektsii po Osnovam Gazovoi Dinamiki*, Institut Komputernykh Issledovanii, Moskva-Izhevsk, Russian, 2003.
[13] M. Short and D. S. Stewart, *Cellular detonation stability. Part* 1. *A normal-mode linear analysis,* J. Fluid Mech., 368 (1998), pp. 229–262.
[14] D. S. Stewart, *The shock dynamics of multi-dimensional condensed and gas-phase detonations*, Proceedings of the Combustion Institute, 27 (1998), pp. 2189–2205.
[15] D. S. Stewart and J. B. Bdzil, *The shock dynamics of stable multi-dimensional detonation,* Comb. Flame, 72 (1988), pp. 311–323.
[16] D. S. Stewart and J. B. Bdzil, *A lecture on detonation shock dynamics,* in Mathematical Modeling in Combustion Science, Lecture Notes in Phys. 249, Springer-Verlag, New York,

1988, pp. 17–30.

[17] W. W. Wood and J. G. Kirkwood, *Diameter effect in condensed explosives. The relation between velocity and radius of curvature in the detonation wave,* J. Chem. Phys., 22 (1954), pp. 1920–1924.

[18] J. Yao and D. S. Stewart, *On the dynamics of multi-dimensional detonation waves,* J. Fluid Mech., 309 (1996), pp. 225–275.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 408–432

THE SPEED LAW FOR HIGHLY RADIATIVE FLAMES IN A GASEOUS MIXTURE WITH LARGE ACTIVATION ENERGY*

JAN BOUWE VAN DEN BERG†, CLAUDE-MICHEL BRAUNER‡, JOSEPHUS HULSHOF†, AND ALESSANDRA LUNARDI§

Abstract. We study a thermodiffusive combustion model for premixed flames propagating in reactive gaseous mixtures which contain inert dust. As observed by Joulin, radiative transfer of heat may significantly enhance the flame temperature and its propagation speed. The Joulin effect is at its most pronounced in the parameter regime where the medium is very transparent while radiative flux dominates convection. In this asymptotic regime, where in the limit the flame temperature achieves its upper bound, we determine the law that describes the relation between the propagation speed of the flame and the control parameters. Finally, we present strong numerical evidence for the validity of the asymptotic analysis.

Key words. travelling wave, singular limit, asymptotic analysis, combustion, radiation, premixed flame, Eddington equation

AMS subject classifications. 35K55, 35B25, 80A25

DOI. 10.1137/04062031X

1. Introduction.

Combustion is one of the important phenomena in our world. It occurs in controlled applications such as rocket engines, energy plants, and cooking on natural gas, as well as in forest fires and mine and tunnel accidents. Experiments in combustion research are both difficult and expensive, which underlines the need for good mathematical models and their analysis.

Combustion models are based on the incorporation of different physical and chemical principles, expressed in the language of mathematics. Simplifying, sometimes heuristic, assumptions are unavoidable to make mathematical treatment possible, be it by numerical, formal asymptotic, or analytical methods. In the latter the modern theory of infinite-dimensional dynamical systems and its application to free boundary problems (FBPs) plays an important role. Such FBPs occur as various flame front models. Asymptotic arguments are strongly intertwined with the derivation of such FBPs from physical and chemical principles.

In this paper we study a thermodiffusive combustion model for premixed flames propagating in reactive gaseous mixtures which contain inert dust that radiates thermal energy. Radiative transfer of heat involves both emission and absorption of radiation and may significantly influence the flame temperature, its propagation speed, and the flammability of the medium itself. This is the so-called Joulin effect [13, 5]: the propagation speed increases compared to a similar flame without radiation, and

*Received by the editors December 6, 2004; accepted for publication (in revised form) February 26, 2005; published electronically November 22, 2005. This work was supported by a CNRS/NWO grant and the RTN network Fronts-Singularities, HPRN-CT-2002-00274.
http://www.siam.org/journals/siap/66-2/62031.html
†Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, the Netherlands (janbouwe@few.vu.nl, jhulshof@few.vu.nl). The first author was supported by a VENI grant from NWO. The third author was supported by the CWI in Amsterdam.
‡Mathématiques Appliquées de Bordeaux, Université Bordeaux I, 33405 Talence cedex, France (brauner@math.u-bordeaux1.fr).
§Dipartimento di Matematica, Università di Parma Via D'Azeglio 85/A, 43100 Parma, Italy (lunardi@prmat.math.unipr.it).

there is a temperature overshoot at the flame front. A radiative flame can be ignited at a lower external temperature than a nonradiative flame.

The Joulin effect is at its most pronounced in the parameter regime where the medium is very transparent while radiative flux dominates convection. In this asymptotic regime our goal is to determine the law that describes the relation between the propagation speed of the flame and the control parameters.

In section 2 we will discuss the model in more detail. For now, we just highlight the most important features. Following Buckmaster and Ludford [6, p. 218], we formulate the thermodiffusive model with the thin reactive flame zone replaced by a free boundary. At the free boundary the normal derivative of the (normalized) temperature T is related to the reaction rate ω, which is given by an Arrhenius-type law:

$$\omega = A \exp\left(-\frac{N}{T^*}\right). \tag{1}$$

Here N is a (dimensionless) activation energy, T^* is the (dimensionless) temperature at the free boundary, and A is a so-called preexponential constant, which will be specified and discussed in detail later.

As a model for the radiative field, we take the Eddington equation, which contains two important radiative parameters: the (dimensionless) opacity α and the Boltzmann number β, a measure of the radiative energy flux compared to the convective flux.

Flames will be modelled as travelling waves propagating into the fresh region where the fuel mass fraction and the temperature are constant, Y_- for the fuel mass fraction and T_- for the temperature. A conservation law implies that the temperature T_+ far behind the flame front is given by

$$T_+ = T_- + Y_-.$$

Depending on the opacity of the medium, radiation may significantly influence the flame profile; see [12, 3, 5]. *Radiative flames* are characterized by an overshoot of the flame temperature T^* as well as an enhancement in the burning rate and flame speed μ, which is given by

$$\mu = \frac{\omega}{Y_-}. \tag{2}$$

In [4] it was proved that the flame temperature T^* is bounded by

$$T_- + Y_- < T^* < T_- + 2Y_-.$$

These bounds, which were already conjectured in [5], are achieved in certain limits. The lower bound is in fact the flame temperature in the absence of radiation (the "adiabatic" case), and it is approached as $\alpha \to \infty$ or $\beta \to 0$ (see [4]). In the present paper, however, we focus on the combined asymptotic regime

$$\alpha \to 0, \tag{3a}$$

$$\beta \to \infty, \tag{3b}$$

$$\alpha\beta \to 0, \tag{3c}$$

because in this regime the flame temperature approaches the upper bound, i.e.,

$$\text{in the limit (3):} \qquad T^* \to T_- + 2Y_-, \tag{4}$$

and the radiative effects are most pronounced.

We are going to combine this asymptotic regime of the radiative parameters with the high activation limit; i.e., we take

$$\varepsilon = \frac{1}{N}$$

as the main small parameter. This is very much in the same spirit as the near-equidiffusional flame (NEF) approximation that is frequently used in the absence of radiative effects; see [15]. There the reciprocal ε of the activation energy is coupled with the Lewis number. Here we couple ε with the radiative parameters.

Of primary physical interest are flames that, in this asymptotic regime, propagate with a finite velocity μ. In view of (2), μ is proportional to the reaction rate ω given in (1). Hence, in the high activation limit $N = \varepsilon^{-1} \to \infty$, finite speeds of propagation can be obtained, provided A is of the order $\exp(N/T_c)$, where T_c is a characteristic temperature, to be fixed shortly. Indeed, since in this notation

$$\mu = \frac{1}{Y_-} \exp\left(\frac{1}{\varepsilon}\left(\frac{1}{T_c} - \frac{1}{T^*}\right)\right), \tag{5}$$

the characteristic temperature T_c should equal the asymptotic value of the flame temperature T^*, and hence in view of (4) the only possibility is

$$T_c = T_- + 2Y_-.$$

This is the upper extreme of T^*, and it stands in sharp contrast with the NEF approach, where the suitable choice for T_c is the lower extreme, namely $T_- + Y_-$.

Since we want to look at the asymptotic regime where simultaneously the reciprocal ε of the activation energy tends to zero and the radiative parameters α and β behave as given in (3), we have to couple α and β with ε. Limit condition (3c) suggests it is convenient to introduce the combined parameter

$$\chi \stackrel{\text{def}}{=} \alpha\beta.$$

Our results show that an asymptotically *finite* propagation speed requires

$$\chi = O(\varepsilon) \qquad \text{and} \qquad \beta^{-1} = O(\varepsilon^{1/2}). \tag{6}$$

Since α has a more direct physical meaning than χ, let us give an alternative formulation of these conditions. For simplicity we assume that both α and β are (asymptotically) powers of ε:

$$\alpha \sim \alpha_0 \varepsilon^a \qquad \text{and} \qquad \beta \sim \beta_0 \varepsilon^{-b}.$$

The connection with (6) is made through

$$\chi = \alpha\beta = \alpha_0\beta_0\varepsilon^{a-b} \sim \chi_0 \varepsilon^{a-b}.$$

To obtain a finite flame velocity, one of the following four possibilities must hold (see also Figure 1):

$$\begin{array}{rlll} \text{I:} & a = \frac{3}{2} & \text{and} & b = \frac{1}{2}; \\ \text{II:} & a > \frac{3}{2} & \text{and} & b = \frac{1}{2}; \\ \text{III:} & a = b + 1 & \text{and} & b > \frac{1}{2}; \\ \text{IV:} & a > b + 1 & \text{and} & b > \frac{1}{2}. \end{array}$$

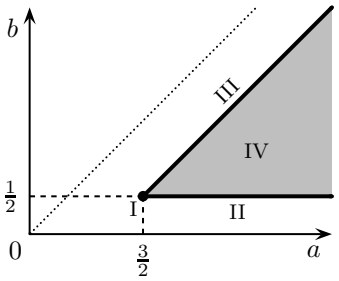


FIG. 1. *The asymptotic regime under consideration in terms of the exponents a and b. The area below the dotted line corresponds to radiation-dominated flames* (3). *Finite wave speeds are found in the shaded region and, more significantly, on its boundary.*

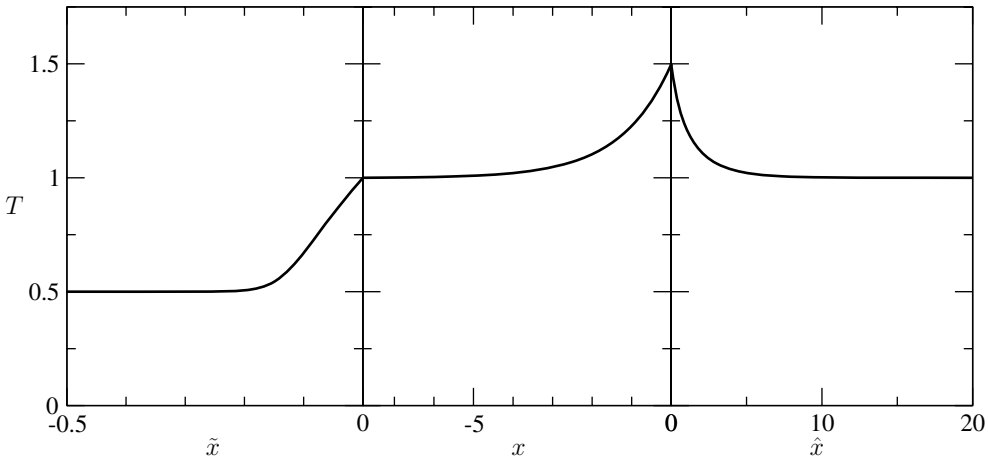


FIG. 2. *In the combined asymptotic limit of high activation energy* $\varepsilon \to 0$ *coupled with the radiative parameters* $\alpha = \alpha_0 \varepsilon^a$ *and* $\beta = \beta_0 \varepsilon^{-b}$ *with* $b \geq \frac{1}{2}$ *and* $a \geq b + 1$, *the solution profile separates into three spatial scales. The numerical solution shown is for* $\varepsilon = 0.001$, $\alpha = 0.3\varepsilon^{3/2}$, $\beta = 0.3\varepsilon^{-1/2}$, *and* $T_- = Y_- = 0.5$. *Notice that the scales are very different in the three regions since on the left the variable is* $\tilde{x} = \alpha\beta^{-1}x$, *in the middle it is* x, *and on the right it is* $\hat{x} = \alpha\beta x$.

We remark that the special scaling $\alpha \sim \alpha_0 \varepsilon^{3/2}$ and $\beta \sim \beta_0 \varepsilon^{1/2}$ in case I was first observed by Joulin and Eudier [13].

In the limit (3) the flame profile naturally separates into three spatial scales (see also Figure 2):

$$x, \qquad \hat{x} = \alpha\beta\, x, \qquad \tilde{x} = \frac{\alpha}{\beta}\, x, \tag{7}$$

where x is the spatial variable in a comoving frame (with speed μ). In section 3 we will perform a matching analysis of these three scales. This enables us to derive a law for the asymptotic speed μ of the front. In the four cases identified above, the speed

law reads (with $T_+ = T_- + Y_-$)

$$\text{(8a)} \qquad \text{I:} \quad \ln(\mu Y_-) = -\frac{\alpha_0\beta_0 T_+^2}{\mu^2} E_1\Big(\tfrac{Y_-}{T_+}\Big) - \frac{\mu^2}{\beta_0^2 T_+^7} E_2\Big(\tfrac{Y_-}{T_+}\Big);$$

$$\text{(8b)} \qquad \text{II:} \quad \ln(\mu Y_-) = - \frac{\mu^2}{\beta_0^2 T_+^7} E_2\Big(\tfrac{Y_-}{T_+}\Big);$$

$$\text{(8c)} \qquad \text{III:} \quad \ln(\mu Y_-) = -\frac{\alpha_0\beta_0 T_+^2}{\mu^2} E_1\Big(\tfrac{Y_-}{T_+}\Big);$$

$$\text{(8d)} \qquad \text{IV:} \quad \ln(\mu Y_-) = 0.$$

Here

$$E_1(s) = \frac{8s + 9s^2 + \frac{16}{3}s^3 + \frac{5}{4}s^4}{(1+s)^2},$$

$$E_2(s) = \frac{3s}{16(1+s)^2} + \frac{3}{4(1+s)^2}\int_0^s \frac{t\,dt}{(1+t)^4 - 1}.$$

It is clear that case I is central to the whole analysis and the other cases are fairly straightforward reductions. On the other hand, in the asymptotic analysis presented in section 3 we will in some sense compute cases II and III and then combine them to obtain case I. The last case, IV, is rather boring, since there is just one finite velocity, namely $\mu = 1/Y_-$. Notice that this corresponds with the maximal flame speed for any of the asymptotic regimes, i.e., $\mu \le 1/Y_-$.

When we compare the four cases we conclude that case I is by far the most interesting, and we will explore it in section 4. Case II leads to a unique flame speed for any set of parameter values, as does case IV (trivially). Case III represents the classical bell-shaped curve of the flame speed versus a heat-loss parameter in nonadiabatic flames (cf. [6, p. 44]).

We remark that in the whole asymptotic regime (3) the profile of any travelling wave with finite speed of propagation (in the asymptotic limit $\varepsilon \to 0$) decomposes into the three different spatial scales (7). The asymptotic analysis in section 3 is thus valid for the whole parameter regime of radiation-dominated flames. In Figure 1 this corresponds to the area below the dotted line. The fact that finite wave speeds occur in only part of this parameter regime is merely a consequence of the way the wave speed is related to T^* and ε (i.e., via (5)).

The organization of the paper is as follows. In section 2 we introduce the mathematical model and make the reduction to a travelling wave problem. In section 3.1 we explain how the matched asymptotic analysis works, while in sections 3.2–3.4 the calculations are performed; i.e., we analyze the profile in three different spatial scales and match these to obtain the full asymptotic picture. This also leads us to the formula for the speed law presented above. Finally, in section 4 we look in more detail at the speed law, we compare with numerical computations, and we draw conclusions about the bifurcation diagrams.

2. Models and equations.

2.1. Premixed flame propagation with constant opacity. We introduce the thermodiffusive combustion model with constant density, simple chemistry, and large activation energy for a premixed flame propagating in a reactive gaseous mixture. We incorporate the flux of the thermal radiative field generated by the radiation of

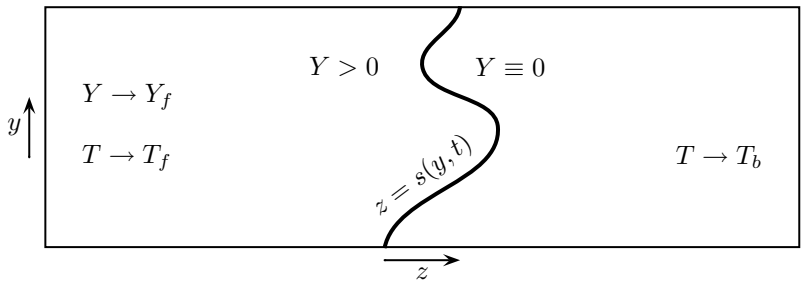


FIG. 3. *The geometric setting of the propagating flame.*

dust particles. The geometric setting is the following (see also Figure 3): the flame propagates into the *fresh* region, where, far ahead of the flame front ($z \to -\infty$), the fuel mass fraction Y and the temperature T are constant:

$$\lim_{z\to-\infty} Y(z) = Y_f \qquad \text{and} \qquad \lim_{z\to-\infty} T(z) = T_f.$$

The region of the flame where the reaction occurs is infinitesimally thin and is located at $z = s(y,t)$, the free boundary of the problem, y being the lateral two-dimensional variable. To the right of the free boundary all fuel has been burnt ($Y(z) = 0$ for $z \geq s(y,t)$), and far behind the flame front the temperature approaches the *burnt* temperature $T_b = \lim_{z\to\infty} T(z)$. The time-dependent system of equations for mass fraction Y and temperature T reads

$$\frac{\partial}{\partial t}(\rho Y) - \nabla(\rho D \nabla Y) = 0, \quad z < s(y,t); \qquad Y = 0, \quad z \geq s(y,t); \tag{9a}$$

$$\frac{\partial}{\partial t}(\rho C_p T) - \lambda \Delta T + \nabla \cdot \mathbf{F_R} = 0, \quad z \neq s(y,t). \tag{9b}$$

The physical parameters are the diffusion constant D, the heat conduction coefficient λ, the specific heat C_p, and the density ρ of the gas (part of which is fuel). The divergence of the radiative energy flux $\mathbf{F_R}$ appears as a loss term in the temperature equation (9b). At the flame front, the jump conditions for the normal derivatives

$$\rho D \left[\frac{\partial Y}{\partial n}\right] = \omega(T); \quad \lambda \left[\frac{\partial T}{\partial n}\right] = -Q\omega(T) \qquad \text{at } z = s(y,t), \tag{9c}$$

are imposed to balance the heat flux coming out of the flame with the mass flux going into the flame, the reaction heat Q being the proportionality constant between the two. These fluxes are also coupled with the chemical reaction rate ω, for which we take a simple Arrhenius law. In the free boundary approximation it reads

$$\omega = A \exp\left(-\frac{E}{2RT^*}\right). \tag{9d}$$

Here T^* denotes the temperature at the flame front, and the other constants are the gas constant R, the activation energy E, and the "preexponential" factor A. Note that the factor 2 in the reaction rate is a consequence of the derivation of the free boundary jump conditions from the reaction-diffusion formulation; see [8]. The appearance of this factor follows from a detailed analysis of the flame in the thin reaction zone,

which leads to (9c), where ω is the square root of the Arrhenius factor in the reaction rate (cf. [8]).

As a law for the radiative flux $\mathbf{F_R}$ we take the Eddington equation

$$-L^2\nabla(\nabla\cdot\mathbf{F_R})+3\mathbf{F_R}+4\sigma_{\rm sb}L\nabla T^4=0, \tag{9e}$$

where $\sigma_{\rm sb}$ is the Stefan–Boltzmann constant and L is the mean free path length of the photons. In astrophysics the Eddington equation is a well-known approximation to the radiative field [11, 17, 16]. It is a good approximation when scattering is nearly isotropic, particularly in a one-dimensional setting. The travelling waves that we use as a model for the propagating flames have indeed a one-dimensional structure. Since the radiative transfer plays a central role in our model, we give some insight in the derivation of the Eddington equation in Appendix A.

We emphasize that the Eddington equation models radiative transfer rather than radiative heat losses. There is, however, an asymptotic limit, discussed in [4, 1], where the radiative flux is given by $\nabla\cdot\mathbf{F_R}=\frac{4\sigma_{\rm sb}}{L}(T^4-T_b^4)$. This asymptotic limit thus looks like heat loss to a reservoir held at $T=T_b$. It can be compared to the usual radiative heat loss models (see [18, sect. 8.2], [6, p. 43]) that are based on the law $\nabla\cdot\mathbf{F_R}=\frac{4\sigma_{\rm sb}}{L}(T^4-T_f^4)$, which differs only in the temperature of the reservoir (T_f instead of T_b).

2.2. Dimensionless variables. We now make the system of equations (9) dimensionless and scale out many of the parameters. We define nondimensional temperature $\hat{T}$, radiative flux $\hat{\mathbf{F}}_\mathbf{R}$, time $\hat{t}$, and spatial coordinate $\hat{\mathbf{r}}$ by comparison with suitable chosen reference quantities indexed by s:

$$\hat{t}=\frac{t}{t_s},\quad \hat{\mathbf{r}}=\frac{\mathbf{r}}{r_s},\quad \hat{T}=\frac{T}{T_s},\quad \hat{\mathbf{F}}_\mathbf{R}=\frac{\mathbf{F_R}}{F_s}.$$

We choose the reference quantities such that they satisfy the following set of equations:

$$\frac{\lambda t_s}{\rho C_p r_s^2}=1,\quad \frac{4\sigma_{\rm sb}T_s^4}{F_s}=1,\quad \frac{\rho D}{Br_s}=1,\quad \frac{\lambda T_s}{QBr_s}=1,$$

that is,

$$F_s=\frac{4\sigma_{\rm sb}Q^4D^4\rho^4}{\lambda^4},\quad T_s=\frac{QD\rho}{\lambda},\quad t_s=\frac{\rho^3C_pD^2}{\lambda B^2},\quad r_s=\frac{D\rho}{B}.$$

Here B is defined by

$$A=B\exp\left(\frac{E}{2RT_C}\right),$$

where T_C is the characteristic temperature. The necessity of this splitting of the preexponential factor A has already been discussed in the introduction. In the high activation energy asymptotics that we are going to employ it is widely used; see, for example, [6, p. 17]. Note that the factor 2 in the reaction rate accounts for B^2, rather than B appearing in t_s.

We have chosen not to rescale the mass fraction Y (which was already dimensionless) because the above choices already simplify the equations as much as we want. Although the additional scaling of Y that we have at our disposal is welcome from a

mathematical point of view, using it obscures the physical role of the control parameters Y_f and/or T_f. Our motivation for the above choices is that we have at hand two important radiative parameters, namely

$$\alpha = \frac{r_s}{L} = \frac{D\rho}{BL},$$

which is a dimensionless opacity, and

$$\beta = \frac{F_s t_s}{\rho C_p T_s r_s} = \frac{4\sigma_{\rm sb} Q^3 D^4 \rho^4}{\lambda^4 B},$$

which is a measure of the radiative flux compared to the convective flux. Furthermore, there is the Lewis number

$$\mathrm{Le} = \frac{r_s^2}{D t_s} = \frac{\lambda}{\rho C_p D},$$

a diffusion parameter. In the new variables the system becomes (where we drop the hats from the notation)

$$\begin{aligned} Y_t - \frac{1}{\mathrm{Le}}\Delta Y = 0, \quad z < s(y,t); \qquad Y \equiv 0, \quad z \geq s(y,t);\\ T_t - \Delta T + \beta\,\nabla\cdot\mathbf{F_R} = 0, \quad z \neq s(y,t);\\ -\nabla(\nabla\cdot\mathbf{F_R}) + 3\alpha^2\mathbf{F_R} + \alpha\nabla T^4 = 0. \end{aligned}$$

The jump conditions at the free boundary $z = s(y,t)$ are

$$\left[\frac{\partial Y}{\partial n}\right] = \mathrm{Le}\,\omega(T) \quad \text{and} \quad \left[\frac{\partial T}{\partial n}\right] = -\omega(T) \qquad \text{at } z = s(y,t),$$

with nondimensional chemical reaction rate (still denoted by ω)

$$\omega(T) = \exp\left(N\left(\frac{1}{T_c} - \frac{1}{T}\right)\right),$$

where

$$N = \frac{E}{2RT_s}$$

is the dimensionless activation energy, and $T_c = T_C/T_s$ is the dimensionless characteristic temperature, the significance of which was already discussed in the introduction.

2.3. Planar travelling waves. We consider flames modelled by planar (one-dimensional) travelling wave solutions, and we thus introduce the travelling wave coordinate $x = z + \mu t$, describing waves travelling at speed μ to the left (into the fresh region). In such a travelling wave the radiative flux has only one component, which we rescale by β for convenience:

$$\mathbf{F_R} = (q/\beta, 0, 0).$$

Also, we introduce the new combined parameter

$$\chi = \alpha\beta.$$

Finally, we may reposition the free boundary at the origin. This leads to the system

$$\mu Y' - \frac{1}{\mathrm{Le}} Y'' = 0, \quad x < 0; \qquad Y \equiv 0, \quad x \geq 0; \tag{10a}$$

$$\mu T' - T'' + q' = 0, \quad x \neq 0; \tag{10b}$$

$$-q'' + 3\alpha^2 q + \chi (T^4)' = 0, \quad x \in \mathbb{R}. \tag{10c}$$

The jump conditions at $x = 0$ are

$$[Y] = [T] = [q] = [q'] = 0, \tag{10d}$$

$$[T'] = -\omega(T), \quad [Y'] = \mathrm{Le}\, \omega(T), \tag{10e}$$

and $\omega(T)$ is still given by

$$\omega(T) = \exp\left(N \left(\frac{1}{T_c} - \frac{1}{T} \right) \right). \tag{10f}$$

Note that the equation (10c) for q implies the continuity of q and q'. The conditions at infinity are

$$T(-\infty) = T_-, \quad Y(-\infty) = Y_-, \quad T(+\infty) = T_+, \quad q(\pm\infty) = 0, \tag{10g}$$

in which $Y_- = Y_f$ is the (dimensionless) "fresh" mass fraction, T_- is the dimensionless fresh temperature, and T_+ is the dimensionless burnt temperature. In fact, direct integration of the equations (see section 3) shows that

$$T_+ = T_- + Y_-. \tag{11}$$

This conservation law (cf. [3, p.221]) reflects the fact that physically only the conditions in the fresh region can be controlled, whereas the temperature in the burnt region is determined by the reaction. The conservation law relating the asymptotic temperatures and fuel mass fraction is independent of the radiation parameters, so that the temperature far behind the flame front is nothing but the adiabatic flame temperature. (In absence of radiation effects, the temperature behind the flame is uniform: $T \equiv T_{ad} = T_- + Y_-$.) The limiting behavior for q at infinity means that radiative equilibrium is achieved at infinity. This follows naturally from (10c); in fact, q may be expressed in terms of T^4 by a convolution formula with a Green's function.

From [4] we know the existence of a travelling wave solution

$$(Y(x), T(x), q(x), \mu)$$

of the system (10) for all (positive) values of the parameters, provided the conservation law (11) is satisfied. Every solution satisfies

$$T_- \leq T(x) \leq T_- + 2Y_-.$$

It is remarkable that this bound is independent of the other parameters.

3. Matched asymptotic analysis.

3.1. Setting the stage. In this section we evaluate the simultaneous asymptotic regime of high activation energy and highly radiative flames. We thus introduce *three* small parameters:

$$\varepsilon = N^{-1}, \tag{12a}$$

$$\delta_1 = \chi = \alpha\beta, \tag{12b}$$

$$\delta_2 = 3\beta^{-2}. \tag{12c}$$

We will couple δ_1 and δ_2 with ε in a moment.

First, we remark that the equation for Y decouples and can be solved explicitly:

$$Y(x) = Y_-(1 - e^{\mathrm{Le}\mu x}), \quad x < 0; \qquad Y(x) = 0, \quad x \geq 0. \tag{13}$$

The jump condition for Y' leads to an expression for the flame velocity:

$$\mu = \frac{1}{Y_-} \exp\left(N\left(\frac{1}{T_c} - \frac{1}{T^*}\right)\right). \tag{14}$$

Since the remaining problem for T, q, and μ is independent of the Lewis number Le, it does not appear in the subsequent asymptotic analysis. However, it plays an important role in the stability analysis, which we discuss in a forthcoming paper [2].

Since Y is given by (13), the system (10a)–(10c) reduces to a set of two equations

$$T'' = \mu T' + q',$$
$$q'' = 3\alpha^2 q + \chi(T^4)'.$$

The first equation can be integrated once, but since T' is discontinuous at $x = 0$, the integration cannot be across the origin. We therefore integrate starting from $x = \infty$ for positive x, while starting from $x = -\infty$ for negative x. This leads to

$$T' = \mu(T - T_\pm) + q \qquad \text{for } x \neq 0.$$

Here and throughout the paper $T_\pm$ stands for T_+ on the right ($x > 0$) and T_- on the left ($x < 0$). We note that we will frequently have to treat the equations separately on the right and on the left.

Using the notation (12) the system reads

$$T' = \mu(T - T_\pm) + q, \tag{15a}$$

$$q'' = \delta_1^2\delta_2 q + \delta_1(T^4)', \tag{15b}$$

with "boundary conditions" at infinity

$$T(-\infty) = T_-, \quad T(+\infty) = T_+, \quad q(\pm\infty) = 0,$$

and at the origin T, q and q' are continuous, while T' satisfies the jump condition

$$\mu(T_- - T_+) = [T'] = -\mu Y_-. \tag{16}$$

The first equality stems from (15a), while the second equality is a consequence of the jump conditions (10e) and the explicit expression (13) for Y. The two equalities in (16) reflect the conservation law

$$T_+ = T_- + Y_-. \tag{17}$$

Here and in what follows we assume that μ is order 1, so that we are dealing with asymptotically finite speeds of propagation. The system (15) now naturally leads to the expansion

$$T \sim T_0 + \delta_1 T_1 + \delta_2 T_2 + \delta_1^2 T_3 + \delta_1\delta_2 T_4 + \delta_2^2 T_5,$$
$$q \sim q_0 + \delta_1 q_1 + \delta_2 q_2 + \delta_1^2 q_3 + \delta_1\delta_2 q_4 + \delta_2^2 q_5.$$

It turns out to be unnecessary to compute the terms of order δ_2^2 to completely determine the leading-order speed law. We will therefore not include those terms in the asymptotic expressions.

In view of (14) our overriding interest is in the temperature at the free boundary. Therefore we introduce the notation $T_i^* = T_i(0)$ for $i = 0, 1, 2$, and hence

$$T(0) = T^* \sim T_0^* + \delta_1 T_1^* + \delta_2 T_2^*.$$

In this new notation the relation (14) between the flame velocity μ and the flame temperature T^* reads

$$\ln(\mu Y_-) \sim \frac{1}{\varepsilon}\left(\frac{1}{T_c} - \frac{1}{T_0^* + \delta_1 T_1^* + \delta_2 T_2^*}\right). \tag{18}$$

There are now several straightforward remarks to make. For the terms on the right and left to balance (i.e., for a finite propagation speed), one needs

$$T_c = T_0^*.$$

This reduces (18) to

$$\ln(\mu Y_-) \sim -\frac{\delta_1}{\varepsilon}\frac{T_1^*}{(T_0^*)^2} - \frac{\delta_2}{\varepsilon}\frac{T_2^*}{(T_0^*)^2}. \tag{19}$$

We anticipate (see below) that $T_1^* < 0$ and $T_2^* < 0$, so the right-hand side of (19) is always nonzero. It is now immediate that we need

$$\delta_1 = O(\varepsilon) \qquad \text{and} \qquad \delta_2 = O(\varepsilon).$$

If both $\delta_1 \ll \varepsilon$ and $\delta_2 \ll \varepsilon$, then we just have $\mu = 1/Y_-$. If δ_1 and/or δ_2 are of order ε, then the left- and right-hand sides balance and the results announced in the introduction follow. Of course, they follow only after we have found the expressions for T_0^*, T_1^* and T_2^*, which are (we will spend the rest of this section establishing this; see (39), (44) and (45))

$$T_0^* = T_+ + Y_-;$$
$$T_1^* = -\mu^{-2}\left(8T_+^3 Y_- + 9T_+^2 Y_-^2 + \frac{16}{3}T_+ Y_-^3 + \frac{5}{4}Y_-^4\right);$$
$$T_2^* = -\frac{\mu^2}{4T_+^5}\int_0^{Y_-/T_+} \frac{t}{(t+1)^4 - 1}\,dt - \frac{\mu^2 Y_-}{16T_+^6}.$$

To calculate T_0^*, T_1^*, and T_2^* we have to match the profile that we obtain on the scale $x = O(1)$ to *two* larger scales. The scale at order $x = O(1)$ we call the inner region, $x = O(\delta_1^{-1})$ is the intermediate region, and $x = O(\delta_1^{-1}\delta_2^{-1})$ is the outer region (see also Figure 4). One may wonder why we do not have a scale $x = O(\delta_2^{-1})$. The

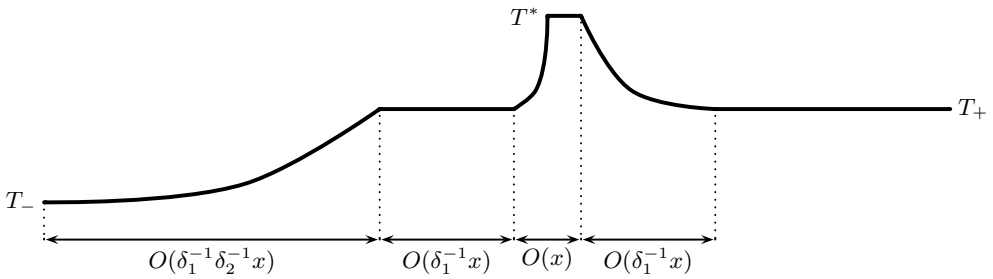


FIG. 4. *The three different scales of the asymptotic problem. The shape of the profiles shown of course uses some a posteriori knowledge which will be collected in the matched asymptotic analysis in sections* 3.2–3.4.

reason is that the profile turns out to be flat at this scale, and therefore no useful information can be extracted. Throughout we calculate with δ_1 and δ_2 as independent quantities. That they are possibly of the same order in ε does not matter whatsoever for the calculations.

In the intermediate and remote regions introduced above, the variables are

$$\begin{aligned} \text{intermediate: } \hat{x} &= \delta_1 x, \qquad \hat{T}(\hat{x}) = T(x) \text{ and } \hat{q}(\hat{x}) = q(x); \\ \text{outer: } \tilde{x} &= \delta_1\delta_2 x, \qquad \tilde{T}(\tilde{x}) = T(x) \text{ and } \tilde{q}(\tilde{x}) = q(x). \end{aligned}$$

(This means $\tilde{x}$ is a factor 3 larger than announced in the introduction, alas.) Although we are eventually interested in the value of T at the origin in the inner region, we start our analysis in the outer region, since there we know the boundary conditions:

$$\lim_{\hat{x}\to\pm\infty} \hat{T}(\hat{x}) = T_\pm \qquad \text{and} \qquad \lim_{\hat{x}\to\pm\infty} \hat{q}(\hat{x}) = 0.$$

We are thus going to work from the outside inward.

3.2. The outer region. The problem in the outer region is

$$\begin{aligned} \delta_1\delta_2\tilde{T}' &= \mu(\tilde{T} - T_\pm) + \tilde{q}, \\ \delta_2\tilde{q}'' &= \tilde{q} + (\tilde{T}^4)', \end{aligned}$$

with boundary conditions

$$\tilde{T}(\pm\infty) = T_\pm \qquad \text{and} \qquad \tilde{q}(\pm\infty) = 0.$$

Of course these equations must be solved on the right and on the left separately, because there are an intermediate as well as an inner region in between. At this outer scale the expansion for $\tilde{T}$ is

$$\tilde{T} = \tilde{T}_0 + \delta_1\tilde{T}_1 + \delta_2\tilde{T}_2,$$

with an analogous expansion for $\tilde{q}$.

The problem for $\tilde{T}_0$ and $\tilde{q}_0$ is

$$\begin{cases} 0 = \mu(\tilde{T}_0 - T_\pm) + \tilde{q}_0, \\ 0 = \tilde{q}_0 + (\tilde{T}_0^4)', \\ \tilde{T}_0(\pm\infty) = T_\pm,\ \tilde{q}_0(\pm\infty) = 0. \end{cases}$$

Combining the equations we get

$$\tilde{T}_0' = \frac{\mu(\tilde{T}_0 - T_\pm)}{4\tilde{T}_0^3}, \qquad \text{with } \tilde{T}_0(\pm\infty) = T_\pm. \tag{20}$$

On the right, the solution is constant:

$$\tilde{T}_0(\tilde{x}) = T_+ \quad \text{and} \quad \tilde{q}_0(\tilde{x}) = 0 \qquad \text{for } \tilde{x} > 0. \tag{21}$$

This follows from the fact that no exponentially growing terms can be present, since it is impossible to match those to the next scale. This argument is silently used several times in what follows.

On the left, we have a choice of the constant solution, an increasing solution and a decreasing solution. As it turns out, the increasing solution will be the one we need. It starts from T_- at $\tilde{x} = -\infty$, and since equation (20) is autonomous, the solution can be translated, and hence the value at the origin is a priori unknown. It has to be determined by matching with the intermediate region. For now, we just introduce the undetermined constant

$$\tilde{T}_0^* \stackrel{\text{def}}{=} \lim_{\tilde{x}\uparrow 0} \tilde{T}_0(\tilde{x}).$$

In order to match with the intermediate region we will need the asymptotic behavior near the origin, which in terms of $\tilde{T}_0^*$ is given by

$$\tilde{T}_0(\tilde{x}) \sim \tilde{T}_0^* + \frac{\mu(\tilde{T}_0^* - T_-)}{4\tilde{T}_0^{*3}}\,\tilde{x} \qquad \text{as } \tilde{x} \uparrow 0, \tag{22}$$

and $\tilde{q}_0(\tilde{x}) = \mu(\tilde{T}_0(\tilde{x}) - T_-)$ for $\tilde{x} < 0$.

Next, the problem for $\tilde{T}_1$ and $\tilde{q}_1$ (at order δ_1) is

$$\begin{cases} 0 = \mu\tilde{T}_1 + \tilde{q}_1, \\ 0 = \tilde{q}_1 + (4\tilde{T}_0^3\tilde{T}_1)', \\ \tilde{T}_1(\pm\infty) = 0,\ \tilde{q}_1(\pm\infty) = 0. \end{cases}$$

The limit behavior as $\tilde{x} \to \pm\infty$ is trivial since $T_\pm$ are independent of δ_1 (and δ_2). As it turns out, we need only the solution on the right. There $\tilde{T}_0 = T_+$, so the two equations can be reduced to

$$\tilde{T}_1' = \frac{\mu\tilde{T}_1}{4T_+^3}, \qquad \text{with } \tilde{T}_1(+\infty) = 0,$$

and hence the solution is simply

$$\tilde{T}_1 = \tilde{q}_1 = 0 \qquad \text{for } \tilde{x} > 0. \tag{23}$$

Similarly, the problem for $\tilde{T}_2$ and $\tilde{q}_2$ is

$$\begin{cases} 0 = \mu\tilde{T}_2 + \tilde{q}_2, \\ \tilde{q}_0 = \tilde{q}_2 + (4\tilde{T}_0^3\tilde{T}_2)', \\ \tilde{T}_2(\pm\infty) = 0,\ \tilde{q}_2(\pm\infty) = 0. \end{cases}$$

Again, we need only the solution on the right, which is simply

$$\tilde{T}_2 = \tilde{q}_2 = 0 \qquad \text{for } \tilde{x} > 0. \tag{24}$$

3.3. The intermediate problem. At the intermediate scale the problem reads

$$\delta_1 \hat{T}' = \mu(\hat{T} - T_\pm) + \hat{q},$$
$$\hat{q}'' = \delta_2 \hat{q} + (\hat{T}^4)',$$

with boundary conditions for $\hat{x} \to -\infty$

$$\hat{T}(\hat{x}) = \tilde{T}_0^* + \delta_1 O(\hat{x}^0) + \delta_2 \left[\tfrac{\mu(\tilde{T}_0^* - T_-)}{4\tilde{T}_0^{*3}} \hat{x} + O(\hat{x}^0) \right] + o(\delta_1, \delta_2); \tag{25}$$
$$\hat{q}(\hat{x}) = -\mu(\tilde{T}_0^* - T_-) + \delta_1 O(\hat{x}^0) + \delta_2 \left[-\tfrac{\mu^2(\tilde{T}_0^* - T_-)}{4\tilde{T}_0^{*3}} \hat{x} + O(\hat{x}^0) \right] + o(\delta_1, \delta_2);$$

and for $\hat{x} \to \infty$

$$\hat{T}(\hat{x}) = T_+ + o(\delta_1, \delta_2); \tag{26}$$
$$\hat{q}(\hat{x}) = o(\delta_1, \delta_2).$$

These boundary conditions are determined by the outer solution; namely, (22) leads to (25), while (21), (23), and (24) imply (26). At the intermediate scale the expansion for $\hat{T}$ is

$$\hat{T} = \hat{T}_0 + \delta_1 \hat{T}_1 + \delta_2 \hat{T}_2,$$

with an analogous expansion for $\hat{q}$.

The problem for $\hat{T}_0$ and $\hat{q}_0$ is

$$\begin{cases} 0 = \mu(\hat{T}_0 - T_\pm) + \hat{q}_0, \\ \hat{q}_0'' = (\hat{T}_0^4)', \\ \hat{T}_0(-\infty) = \tilde{T}_0^*,\ \hat{q}_0(-\infty) = -\mu(\tilde{T}_0^* - T_-), \\ \hat{T}_0(+\infty) = T_+,\ \hat{q}_0(+\infty) = 0. \end{cases}$$

The two equations can be combined into

$$\mu \hat{T}_0'' + (\hat{T}_0^4)' = 0. \tag{27}$$

On the left, we integrate from $\hat{x} = -\infty$ and obtain $\mu \hat{T}_0' + \hat{T}_0^4 - \tilde{T}_0^{*4} = 0$. Since $\hat{T}_0(-\infty) = \tilde{T}_0^*$ the solution on the left is

$$\hat{T}_0(\hat{x}) = \tilde{T}_0^* \quad \text{and} \quad \hat{q}_0(\hat{x}) = -\mu(\tilde{T}_0^* - T_-) \qquad \text{for } \hat{x} < 0. \tag{28}$$

On the right, integration of (27) from $\hat{x} = \infty$ gives

$$\mu \hat{T}_0' = -\hat{T}_0^4 + T_+^4, \qquad \text{with } \hat{T}_0(+\infty) = T_+. \tag{29}$$

This situation is very similar to that in the left outer region. We have a choice of the constant solution, an increasing solution and a decreasing solution, and it is the latter that we need. Since the equation is autonomous, the solution can be translated, and hence the value at the origin is a priori unknown. It has to be determined by matching with the inner region. Again, we introduce an undetermined constant

$$\hat{T}_0^* \overset{\text{def}}{=} \lim_{\hat{x} \downarrow 0} \hat{T}_0(\hat{x}).$$

For the asymptotic behavior near $\hat{x} = 0$ we get, using (27) and (29),

$$\hat{T}_0(\hat{x}) \sim \hat{T}_0^* - \mu^{-1}(\hat{T}_0^{*4} - T_+^4)\,\hat{x} + 2\mu^{-2}\hat{T}_0^{*3}(\hat{T}_0^{*4} - T_+^4)\,\hat{x}^2 \qquad \text{as } \hat{x} \downarrow 0, \tag{30}$$

and of course $\hat{q}_0(\hat{x}) = -\mu(\hat{T}_0(\hat{x}) - T_+)$ for $\hat{x} > 0$.

The problem for $\hat{T}_1$ and $\hat{q}_1$ is

$$\begin{cases} \hat{T}_0' = \mu\hat{T}_1 + \hat{q}_1, \\ \hat{q}_1'' = (4\hat{T}_0^3\hat{T}_1)', \\ \hat{T}_1(\pm\infty) = 0,\ \hat{q}_1(\pm\infty) = 0. \end{cases}$$

On the left, the equation reduces to $\mu\hat{T}_1'' + 4\tilde{T}_0^{*3}\hat{T}_1' = 0$, and hence the solution is constant. The value of the constant is unknown at this point. Since we shortly have to match with the inner region, we use the undetermined limit value at the origin $\hat{T}_1^- \stackrel{\text{def}}{=} \hat{T}_1(0^-)$ to denote the constant:

$$\hat{T}_1(\hat{x}) = \hat{T}_1^- \quad \text{and} \quad \hat{q}_1(\hat{x}) = -\mu\hat{T}_1^- \qquad \text{for } \hat{x} < 0. \tag{31}$$

On the right, one obtains $\mu\hat{T}_1'' = -4(\hat{T}_0^3\hat{T}_1)' + \hat{T}_0'''$. Integrating from 0 to ∞ we get, using (30) and setting $\hat{T}_1^+ \stackrel{\text{def}}{=} \hat{T}_1(0^+)$,

$$\hat{T}_1'(0^+) = -4\mu^{-1}\hat{T}_0^{*3}\hat{T}_1^+ + 4\mu^{-3}\hat{T}_0^{*3}(\hat{T}_0^{*4} - T_+^4). \tag{32}$$

This equation expresses $\hat{T}_1'(0^+)$ in the unknown constant $\hat{T}_1^+$. The behavior of $\hat{q}_1$ near the origin is given by

$$\begin{aligned} \hat{q}_1(\hat{x}) &\sim \hat{T}_0'(0^+) - \mu\hat{T}_1^+ + (\hat{T}_0''(0^+) - \mu\hat{T}_1'(0^+))\,\hat{x} \\ &\sim -\mu^{-1}(\hat{T}_0^{*4} - T_+^4) - \mu\hat{T}_1^+ + 4\hat{T}_0^{*3}\hat{T}_1^+\,\hat{x} \qquad \text{as } \hat{x} \downarrow 0. \end{aligned}$$

The problem for $\hat{T}_2$ and $\hat{q}_2$ is

$$\begin{cases} 0 = \mu\hat{T}_2 + \hat{q}_2, \\ \hat{q}_2'' = \hat{q}_0 + (4\hat{T}_0^3\hat{T}_2)', \\ \hat{T}_2(+\infty) = 0,\ \hat{q}_2(+\infty) = 0, \\ \hat{T}_2'(-\infty) = \frac{\mu(\tilde{T}_0^* - T_-)}{4\tilde{T}_0^{*3}},\ \hat{q}_2'(-\infty) = -\frac{\mu^2(\tilde{T}_0^* - T_-)}{4\tilde{T}_0^{*3}}. \end{cases}$$

On the left, the equation reduces to $\mu\hat{T}_2'' = -4\tilde{T}_0^{*3}\hat{T}_2' + \mu(\tilde{T}_0^* - T_-)$, with solution, setting as usual $\hat{T}_2^- \stackrel{\text{def}}{=} \hat{T}_2(0^-)$,

$$\hat{T}_2(\hat{x}) = \hat{T}_2^- + \tfrac{\mu(\tilde{T}_0^* - T_-)}{4\tilde{T}_0^{*3}}\hat{x} \quad \text{and} \quad \hat{q}_2(\hat{x}) = -\mu\hat{T}_2^- - \tfrac{\mu^2(\tilde{T}_0^* - T_-)}{4\tilde{T}_0^{*3}}\hat{x} \qquad \text{for } \hat{x} < 0. \tag{33}$$

On the right, the equation becomes $\mu\hat{T}_2'' = -(4\hat{T}_0^3\hat{T}_2)' + \mu(\hat{T}_0 - T_+)$. Integrating from 0 to ∞ we get, setting $\hat{T}_2^+ \stackrel{\text{def}}{=} \hat{T}_2(0^+)$,

$$\hat{T}_2'(0^+) = -4\mu^{-1}\hat{T}_0^{*3}\hat{T}_2^+ - \int_0^\infty \big[\hat{T}_0(\hat{x}) - T_+\big]\,d\hat{x}. \tag{34a}$$

The last integral involves the function $\hat{T}_0(\hat{x})$, which we have not computed explicitly. The integral can be simplified using equation (29) for $\hat{T}_0$:

$$I_2 \stackrel{\text{def}}{=} \int_0^\infty \left[\hat{T}_0(\hat{x}) - T_+\right] d\hat{x} = -\int_0^{\hat{T}_0^* - T_+} \frac{\hat{T}_0 - T_+}{(\hat{T}_0 - T_+)'}\, d(\hat{T}_0 - T_+)$$

$$= \frac{\mu}{T_+^2} \int_0^{\hat{T}_0^*/T_+ - 1} \frac{t}{(t+1)^4 - 1}\, dt. \tag{34b}$$

We could compute the primitive, but that does not lead to more insight. Finally, the behavior of $\hat{q}_2$ near the origin is given by

$$\hat{q}_2(\hat{x}) \sim -\mu \hat{T}_2^+ + [4\hat{T}_0^{*3}\hat{T}_2^+ + \mu I_2]\,\hat{x} \qquad \text{as } \hat{x} \downarrow 0.$$

3.4. The inner problem. We are getting to the core of the problem. In the inner scale we want to solve

$$T' = \mu(T - T_\pm) + q,$$
$$q'' = \delta_1^2\delta_2 q + \delta_1 (T^4)'.$$

The boundary conditions are for $x \to -\infty$:

$$\begin{aligned} T(x) &= \tilde{T}_0^* + \delta_1 \hat{T}_1^- + \delta_2 \hat{T}_2^- + \delta_1^2 O(x^0) \\ &\quad + \delta_1\delta_2 \left[\tfrac{\mu(\tilde{T}_0^* - T_-)}{4\tilde{T}_0^{*3}} x + O(x^0)\right] + o(\delta_1^2, \delta_1\delta_2, \delta_2); \\ q(x) &= -\mu(\tilde{T}_0^* - T_-) - \delta_1 \mu \hat{T}_1^- - \delta_2 \mu \hat{T}_2^- + \delta_1^2 O(x^0) \\ &\quad - \delta_1\delta_2 \left[\tfrac{\mu^2(\tilde{T}_0^* - T_-)}{4\tilde{T}_0^{*3}} x + O(x^0)\right] + o(\delta_1^2, \delta_1\delta_2, \delta_2). \end{aligned} \tag{35}$$

For $x \to \infty$ the boundary conditions look complicated:

$$\begin{aligned} T(x) &= \hat{T}_0^* + \delta_1[-\mu^{-1}(\hat{T}_0^{*4} - T_+^4)x + \hat{T}_1^+] + \delta_2 \hat{T}_2^+ \\ &\quad + \delta_1^2 \hat{T}_0^{*3}\left[2\mu^{-2}(\hat{T}_0^{*4} - T_+^4)x^2 + 4\{-\mu^{-1}\hat{T}_1^+ + \mu^{-3}(\hat{T}_0^{*4} - T_+^4)\}x + O(x^0)\right] \\ &\quad + \delta_1\delta_2[(-4\mu^{-1}\hat{T}_0^{*3}\hat{T}_2^+ - I_2)x + O(x^0)] + o(\delta_1^2, \delta_1\delta_2, \delta_2); \\ q(x) &= -\mu(\hat{T}_0^* - T_+) + \delta_1[(\hat{T}_0^{*4} - T_+^4)x - \mu^{-1}(\hat{T}_0^{*4} - T_+^4) - \mu\hat{T}_1^+] - \delta_2\mu\hat{T}_2^+ \\ &\quad + \delta_1^2[-2\mu^{-1}\hat{T}_0^{*3}(\hat{T}_0^{*4} - T_+^4)x^2 + 4\hat{T}_0^{*3}\hat{T}_1^+ x + O(x^0)] \\ &\quad + \delta_1\delta_2[(4\hat{T}_0^{*3}\hat{T}_2^+ + \mu I_2)x + O(x^0)] + o(\delta_1^2, \delta_1\delta_2, \delta_2). \end{aligned} \tag{36}$$

These conditions follow from the analysis of the intermediate region; e.g., (35) follows from (28), (31), and (33), whereas (36) follows from (30), (32), and (34). Of course the boundary conditions for $T(x)$ and $q(x)$ as $x \to \pm\infty$ are related through the equation $q = T' - \mu(T - T_\pm)$. Furthermore, at the origin q, q' and T are continuous.

We expand T as

$$T \sim T_0 + \delta_1 T_1 + \delta_2 T_2 + \delta_1^2 T_3 + \delta_1\delta_2 T_4,$$

and analogously for q. As mentioned before, terms of order δ_2^2 do not need to be computed. We now solve subsequently the equations at zeroth, first, and second order in the small parameters δ_1 and δ_2.

3.4.1. Zeroth order. The equations for T_0 and q_0 are

$$\begin{cases} T_0' = \mu(T_0 - T_\pm) + q_0, \\ q_0'' = 0, \\ T_0(-\infty) = \tilde{T}_0^*,\ q_0(-\infty) = -\mu(\tilde{T}_0^* - T_-), \\ T_0(+\infty) = \hat{T}_0^*,\ q_0(+\infty) = -\mu(\hat{T}_0^* - T_+). \end{cases}$$

The functions T_0, q_0, and q_0' are continuous across the origin. This means that $q_0(x)$ is constant, and on the right $T_0(x)$ is constant as well. This implies

$$T_0^* \overset{\text{def}}{=} T_0(0) = T_0(+\infty) \qquad \text{and} \qquad q_0(-\infty) = q_0(+\infty),$$

and hence a comparison with the boundary conditions (and (17)) leads to

$$T_0^* = \hat{T}_0^* = \tilde{T}_0^* + T_+ - T_- = \tilde{T}_0^* + Y_-,$$

that is,

$$\hat{T}_0^* = T_0^*, \tag{37a}$$

$$\tilde{T}_0^* = T_0^* - Y_-. \tag{37b}$$

On the left, the solution $T(x)$ decays exponentially to $\tilde{T}_0^* = T_0^* - Y_-$, so

$$T_0(x) = \begin{cases} Y_-(e^{\mu x} - 1) + T_0^* & \text{for } x < 0, \\ T_0^* & \text{for } x \geq 0, \end{cases} \qquad \text{and} \quad q_0(x) = -\mu(T_0^* - T_+).$$

3.4.2. First order. The equations for T_1 and q_1 are (using (37a))

$$\begin{cases} T_1' = \mu T_1 + q_1, \\ q_1'' = (T_0^4)', \\ T_1(-\infty) = \hat{T}_1^-,\ q_1(-\infty) = -\mu\hat{T}_1^-, \\ T_1(x) \sim -\mu^{-1}(T_0^{*4} - T_+^4)x + \hat{T}_1^+ \ \text{ as } x \to \infty, \\ q_1(x) \sim (T_0^{*4} - T_+^4)x - \mu^{-1}(T_0^{*4} - T_+^4) - \mu\hat{T}_1^+ \ \text{ as } x \to \infty. \end{cases}$$

We start by integrating the second equation from $x = -\infty$:

$$q_1'(x) = T_0(x)^4 - T_0(-\infty)^4 = \begin{cases} [Y_-(e^{\mu x} - 1) + T_0^*]^4 - [T_0^* - Y_-]^4, & x < 0, \\ T_0^{*4} - [T_0^* - Y_-]^4, & x \geq 0. \end{cases} \tag{38}$$

We thus have, by comparing with the boundary conditions for q_1 as $x \to \infty$,

$$T_0^{*4} - T_+^4 = q_1'(+\infty) = T_0^{*4} - [T_0^* - Y_-]^4,$$

and hence

$$T_0^* = T_+ + Y_-. \tag{39}$$

Although we now have an expression for T_0^*, we keep using the notation T_0^* in the proceeding for notational convenience.

Another integration of (38) from $x = 0$ in both directions gives

$$\begin{aligned} &q_1(x) \to q_1(0) - I_1 && \text{as } x \to -\infty, \\ &q_1(x) = q_1(0) + (T_0^{*4} - T_+^4)\, x && \text{for } x \geq 0, \end{aligned}$$

where

$$I_1 \overset{\text{def}}{=} \int_{-\infty}^{0} (Y_- e^{\mu x} + T_+)^4 - T_+^4 \, dx = \mu^{-1}\big(4T_+^3 Y_- + 3T_+^2 Y_-^2 + \tfrac{4}{3} T_+ Y_-^3 + \tfrac{1}{4} Y_-^4\big). \tag{40}$$

To obtain $T_1(x)$ on the right we solve $T_1' - \mu T_1 = q_1(0) + (T_0^{*4} - T_+^4)\, x$, and we obtain

$$T_1(x) = T_1^* - \mu^{-1}(T_0^{*4} - T_+^4)\, x \qquad \text{for } x \geq 0, \tag{41}$$

where

$$T_1^* \overset{\text{def}}{=} T_1(0) = -\mu^{-1} q_1(0) - \mu^{-2}(T_0^{*4} - T_+^4). \tag{42}$$

On the left, the limit behavior of T_1 is (using (42))

$$\lim_{x \to -\infty} T_1(x) = -\mu^{-1} q_1(0) + \mu^{-1} I_1 = T_1^* + \mu^{-2}(T_0^{*4} - T_+^4) + \mu^{-1} I_1. \tag{43}$$

Comparing (41) and (43) with the boundary conditions gives the values for $\hat{T}_1^{\pm}$:

$$\begin{aligned} \hat{T}_1^- &= T_1^* + \mu^{-2}(T_0^{*4} - T_+^4) + \mu^{-1} I_1, \\ \hat{T}_1^+ &= T_1^*. \end{aligned}$$

Next, the equation at order δ_2 is

$$\begin{cases} T_2' = \mu T_2 + q_2, \\ q_2'' = 0, \\ T_2(-\infty) = \hat{T}_2^-,\ q_2(-\infty) = -\mu \hat{T}_2^-, \\ T_2(+\infty) = \hat{T}_2^+,\ q_2(+\infty) = -\mu \hat{T}_2^+. \end{cases}$$

Since the equations are satisfied on $\mathbb{R}$ the solution is constant, so $\hat{T}_2^- = \hat{T}_2^+ = T_2(0)$ and

$$T_2(x) = T_2^* \overset{\text{def}}{=} T_2(0) \qquad \text{for all } x \in \mathbb{R}.$$

3.4.3. Second order. The equation at order δ_1^2 reads

$$\begin{cases} T_3' = \mu T_3 + q_3, \\ q_3'' = 4(T_0^3 T_1)', \\ T_3'(-\infty) = 0,\ q_3'(-\infty) = 0, \\ T_3'(x) \sim 4\mu^{-2} T_0^{*3}(T_0^{*4} - T_+^4)\, x + 4\mu^{-3} T_0^{*3}(-\mu^2 T_1^* + T_0^{*4} - T_+^4) \ \text{ as } x \to \infty, \\ q_3'(x) \sim -4\mu^{-1} T_0^{*3}(T_0^{*4} - T_+^4)\, x + 4 T_0^{*3} T_1^* \ \text{ as } x \to \infty. \end{cases}$$

Integrating the second equation from $x = -\infty$ gives

$$q_3'(x) = 4T_0(x)^3 T_1(x) - 4T_+^3 T_1(-\infty),$$

and by using (41) and (43) we obtain for $x \geq 0$

$$q_3'(x) = -4\mu^{-1} T_0^{*3}(T_0^{*4} - T_+^4) x + 4T_0^{*3} T_1^* - 4T_+^3 [T_1^* + \mu^{-2}(T_0^{*4} - T_+^4) + \mu^{-1} I_1].$$

Comparing this with the boundary conditions for $q_3'(x)$ as $x \to \infty$ gives

$$T_1^* = -\mu^{-1} I_1 - \mu^{-2}\big((T_+ + Y_-)^4 - T_+^4\big), \tag{44}$$

with I_1 given in (40).

The equation at order $\delta_1\delta_2$ reads (using (37b))

$$\begin{cases} T_4' = \mu T_4 + q_4, \\ q_4'' = 4(T_0^3 T_2)', \\ T_4'(-\infty) = \frac{\mu Y_-}{4T_+^3},\ q_4'(-\infty) = -\frac{\mu^2 Y_-}{4T_+^3}, \\ T_4'(+\infty) = -4\mu^{-1} T_0^{*3} T_2^* - I_2,\ q_4'(+\infty) = 4T_0^{*3} T_2^* + \mu I_2. \end{cases}$$

Integrating the second equation from $x = -\infty$ to $x = \infty$ gives

$$q_4'(+\infty) - q_4'(-\infty) = 4[T_0^{*3} - T_+^3]T_2^*.$$

On the other hand, the boundary conditions say that

$$q_4'(+\infty) - q_4'(-\infty) = 4T_0^{*3} T_2^* + \mu I_2 + \frac{\mu^2 Y_-}{4T_+^3}.$$

Comparing these expressions for $q_4'(+\infty) - q_4'(-\infty)$ gives

$$T_2^* = -\frac{\mu}{4T_+^3} I_2 - \frac{\mu^2 Y_-}{16 T_+^6}\,, \tag{45}$$

with I_2 given in (34b).

4. The asymptotic law for the velocity. In this section we take a look at what the speed law tells us. We compare the asymptotic formula with numerical computations for small finite values of ε. In particular, we calculate bifurcation diagrams where the radiative parameters α and β are the continuation parameters. For this, the most delicate case where both terms in the right-hand side of (8) are present is the most interesting, i.e., (8a). In this limit the activation energy ε^{-1} is coupled with the radiative parameters via

$$\alpha = \alpha_0 \varepsilon^{3/2} \qquad \text{and} \qquad \beta = \beta_0 \varepsilon^{-1/2}.$$

The relation between the wave speed μ and α_0 and β_0 is thus

$$\ln(\mu\, Y_-) + \frac{\alpha_0 \beta_0 T_+^2}{\mu^2} E_1\Big(\tfrac{Y_-}{T_+}\Big) + \frac{\mu^2}{\beta_0^2 T_+^7} E_2\Big(\tfrac{Y_-}{T_+}\Big) = 0. \tag{46}$$

We note that due to our choice not to scale Y_- (see section 2.2) this expression should be invariant under the scaling

$$Y_- \to sY_-, \quad T_+ \to sT_+, \quad \mu \to s^{-1}\mu, \quad \alpha_0 \to s^{1/2}\alpha_0, \quad \beta_0 \to s^{-9/2}\beta_0,$$

and it is indeed. Furthermore, if one would replace the nonlinear term T^4 in (10c) by a linear approximation, then the solution can be (almost) explicitly calculated for any α and β. In the limit under consideration, an expression for the speed law analogous to (46) is found; see [2]. It is in that context that the stability investigation is being pursued.

The functions E_1 and E_2 depend only on the quotient Y_-/T_+, and since $T_+ = Y_- + T_-$, they thus depend on the ratio of the fuel mass fraction and the dimensionless temperature far ahead of the front. These two functions are plotted in Figure 5.

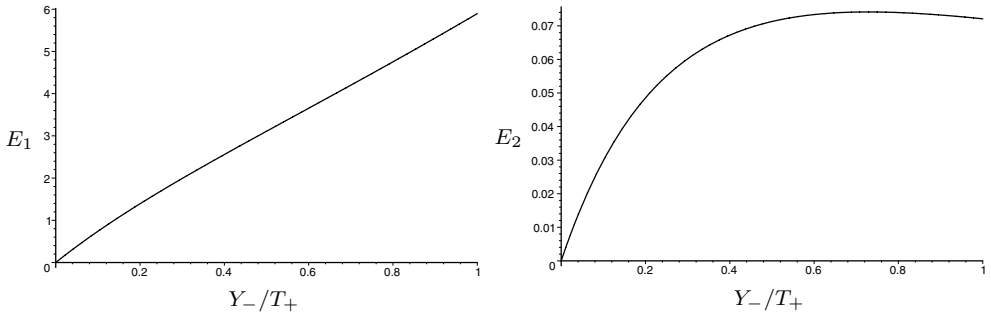


FIG. 5. *The functions* $E_1\left(\frac{Y_-}{T_+}\right)$ *and* $E_2\left(\frac{Y_-}{T_+}\right)$. *Notice that* $T_+ = T_- + Y_-$ *and thus* $0 < \frac{Y_-}{T_+} < 1$.

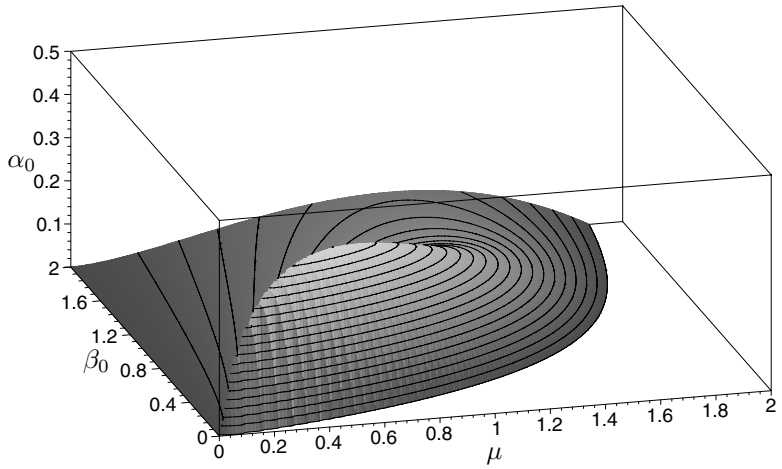


FIG. 6. *The surface in* (μ, α_0, β_0)*-space describing the speed law.*

For the subsequent numerical calculations we need to pick some values for the parameters, and we choose $Y_- = T_- = 0.5$ and hence $T_+ = 1$, throughout. The remaining variables in (46) are thus α_0, β_0, and μ. We can plot the surface most easily by writing α_0 as a function of μ and β_0, and the result is shown in Figure 6. When we fix β_0, then for small α_0 there are two solutions which merge in a saddle-node bifurcation as α_0 increases. On the other hand, when we fix α_0 and use β_0 as the bifurcation parameter we see that the set of solutions forms an *isola* in the (β_0, μ)-plane. This means that β_0 has to be carefully selected, not too large and not too small, for a flame with finite propagation speed to exist. If α_0 is too large, then there are no travelling waves. The maximum value of α_0 for which solutions exist can be calculated to be

$$\alpha_{\max} = \sqrt{\frac{T_+^3}{2e^3 Y_-^2 E_1^2 E_2}}.$$

To compare the asymptotic analysis with numerical computations, we implemented the travelling wave problem in the AUTO software package [7] for the continuation of solutions to ODEs. We treated the three different spatial regions with some care to reflect their respective scaling with ε (or with δ_1 and δ_2 to be more

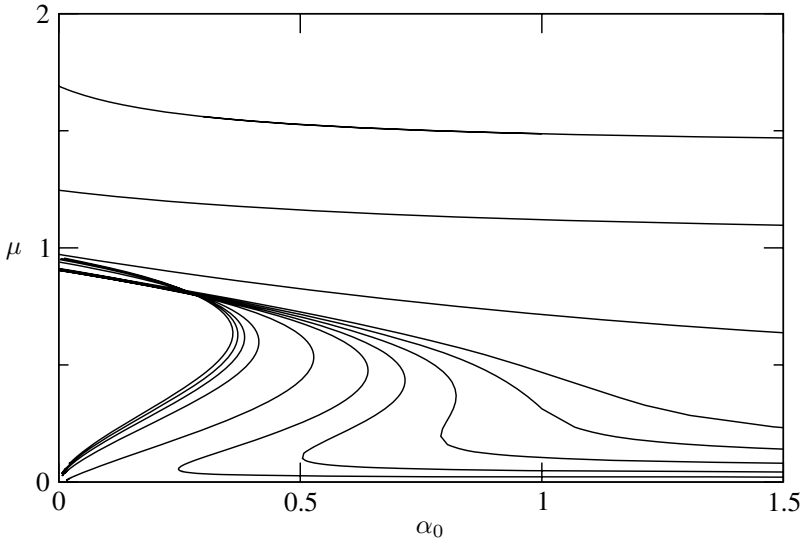


FIG. 7. *The solution curves in the* (α_0, μ)*-plane for fixed* $\beta_0 = 0.3$ *and* $\varepsilon = 1, 0.5, 0.2, 0.11, 0.1, 0.09, 0.08, 0.07, 0.05, 0.02, 0.01, 0.005, 0.001$. *As* ε *decreases, the solution curves shift inwards, i.e., the curve at the top corresponds to* $\varepsilon = 1$.

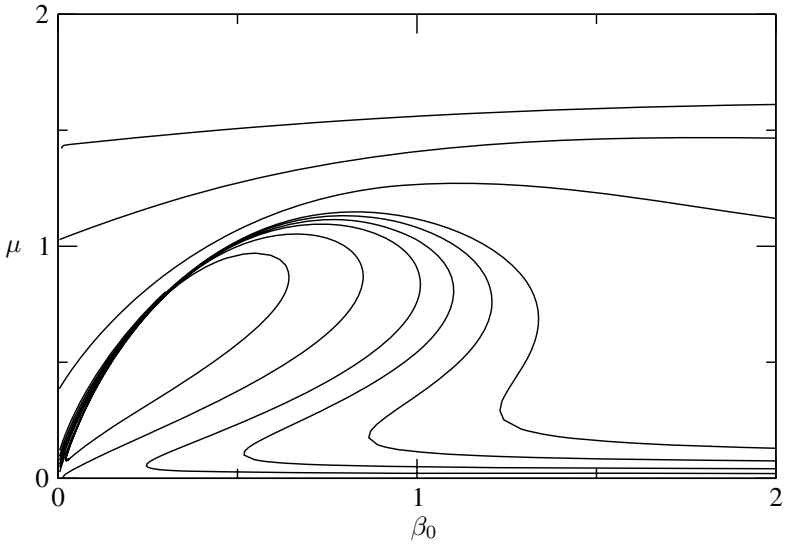


FIG. 8. *The solution curves in the* (β_0, μ)*-plane for fixed* $\alpha_0 = 0.3$ *and the same values of* ε *as in Figure* 7. *As* ε *decreases, the solution curves shift inward.*

precise). We calculated the bifurcation diagram using both α_0 and β_0 as parameters for a set of small values of ε. The resulting pictures are shown in Figures 7 and 8, and one can see how the asymptotic regime is approached. For the (α_0, μ)-diagram the solution curves become S-shaped as ε decreases and then approach a bell-shaped curve as $\varepsilon \to 0$. In the (β_0, μ)-diagram the solution branch curves back more and more and finally closes on itself as ε approaches 0.

To be able to compare with the analytic expression in the limit $\varepsilon \to 0$, we fixed

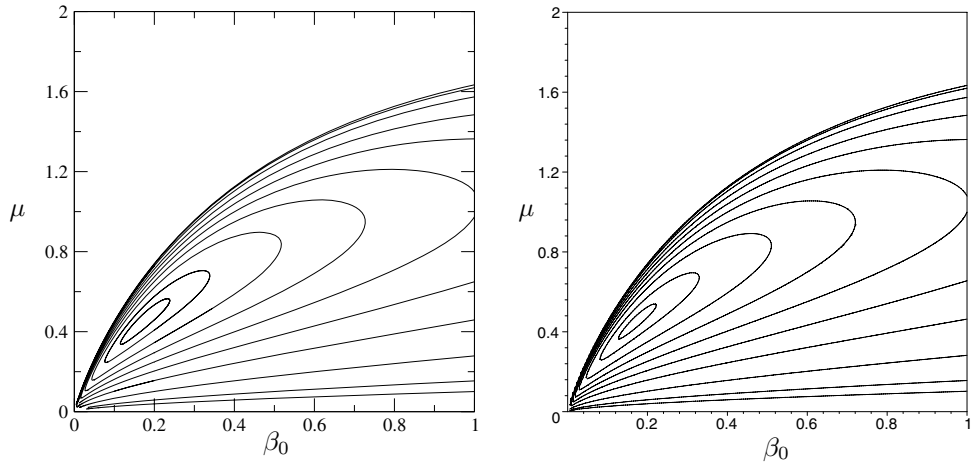


FIG. 9. *On the left is the* (β_0, μ) *bifurcation diagram for* $\varepsilon = 0.001$ *and* $\alpha_0 = 0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.375$. *As* α_0 *increases, the curves are moving inward. On the right are, for the same set of* α_0 *values, the contour lines of the surface (see Figure* 6*) describing the asymptotic speed law as* $\varepsilon \to 0$.

$\varepsilon = 0.001$ and computed the (β_0, μ)-diagram for various values of α_0. The resulting curves can thus be compared with the contour lines of the surface in Figure 6. The numerical computations and the contour lines of the analytic expression are depicted in Figure 9 side by side. The agreement is excellent.

5. Conclusion. In a thermodiffusive combustion model we have studied the influence of radiation effects on propagating flames. In particular, radiative heat transfer enhances the flame temperature and its propagation speed. This so-called Joulin effect is at its most prominent when the medium is fairly transparent while the radiative flux dominates convection. In this asymptotic regime the flame temperature approaches its upper bound. We have determined the law that describes the relation between the propagation speed of the flame and the control parameters.

We have arrived at distinguished limits in four cases. Case I exhibits a rich gamut of bifurcation diagrams, such as S-curves in the (α_0, μ)-plane and S-curves and isolas in the (β, μ)-plane. Case III is nothing but the classical bell-shaped curve in nonadiabatic flame models with heat loss. Cases II and IV correspond to straightforward travelling wave dynamics, and we have derived the corresponding laws for the sake of completeness.

Although the matched asymptotic analysis in this paper is fairly cumbersome, we think it would be fruitful to continue the work in the future in two directions. First, a less simplified Arrhenius reaction term will contain the mass fraction Y, leading to a more involved system of three equations instead of two. Second, a more detailed description of the radiative transfer than the Eddington equation can be taken into account. We hope this paper will serve as a guideline for these extensions.

Appendix A. The Eddington equation.

We consider radiative transfer in a medium of opacity κ at temperature T. A photon travelling at the light speed c covers a distance $L = 1/\kappa$ (the mean free path length of the photon) before being absorbed. Loosely speaking, $L = \infty$ ($\kappa = 0$) corresponds to a transparent medium (optically thin limit) and $L = 0$ ($\kappa = \infty$) to an

opaque medium (optically thick limit).

We start from the equation of radiative transfer for the radiative intensity $I = I(\mathbf{r}, \nu, \Omega, t)$,

$$\frac{1}{c}\frac{\partial I}{\partial t} + \Omega \cdot \nabla I = \kappa(\mathcal{B}(T,\nu) - I). \tag{47}$$

Here $\mathbf{r}$ is the position, t the time, ν the frequency, and Ω the unit vector in the direction of propagation. The Planck distribution $\mathcal{B}$ governs the emission of light by the medium and is given by

$$\mathcal{B}(T,\nu) = \frac{2h}{c^2}\frac{\nu^3}{e^{h\nu/(kT)} - 1},$$

where k and h are the Boltzmann and Planck constants.

Since we would like to consider the total amount of radiation, we denote by $\langle\phi\rangle$ the integral of a function ϕ over all frequencies and directions, rescaled with c:

$$\langle\phi\rangle = \frac{1}{c}\int_0^\infty \int_{S^2} \phi(\nu,\Omega)\, d\Omega\, d\nu.$$

Observing that

$$\langle\mathcal{B}(T)\rangle = aT^4 \quad \text{with} \quad a = \frac{8\pi^5 k^4}{15h^3c^3},$$

one obtains from (47) the system [14, 9, 10]

$$\frac{\partial E_R}{\partial t} + \nabla \cdot \mathbf{F_R} = c\kappa(aT^4 - E_R); \tag{48a}$$

$$\frac{1}{c}\frac{\partial \mathbf{F_R}}{\partial t} + c\nabla \mathbf{P_R} = -\kappa \mathbf{F_R}, \tag{48b}$$

for the radiative energy density E_R, the radiative flux $\mathbf{F_R}$, and the radiative pressure $\mathbf{P_R}$, defined by

$$\begin{aligned} E_R &= \langle I\rangle, \\ \mathbf{F_R} &= c\,\langle \Omega\, I\rangle, \\ \mathbf{P_R} &= \langle \Omega \otimes \Omega\, I\rangle. \end{aligned}$$

The factor c is, as usual, included in the definition of $\mathbf{F_R}$ so that it represents an energy flux. Notice that equations (48) do not form a closed system. They are the first members of a hierarchy, and the system still needs to be closed. If the emission and absorption would be isotropic, then we would have

$$\mathbf{F_R} = 0,$$

and also

$$\mathbf{P_R} = \frac{1}{3}E_R\mathbf{Id}. \tag{49}$$

In the so-called P_1-model, which leads to the Eddington equation, (49) is taken as a closure assumption, so that (48) is replaced by

$$\frac{\partial E_R}{\partial t} + \nabla \cdot \mathbf{F_R} = c\kappa(aT^4 - E_R); \tag{50a}$$

$$\frac{1}{c}\frac{\partial \mathbf{F_R}}{\partial t} + \frac{1}{3}c\nabla E_R = -\kappa \mathbf{F_R}. \tag{50b}$$

Since photons travel at light speed we may assume that the radiation is approximately at steady state at the typical time scale of a moving flame; i.e., the system (50) reduces to

$$\nabla \cdot \mathbf{F_R} = c\kappa(aT^4 - E_R); \qquad \frac{1}{3}c\nabla E_R = -\kappa \mathbf{F_R}.$$

It is not difficult to eliminate one of the unknowns, say E_R, by differentiating the first equation, whence

$$c\kappa\nabla E_R = c\kappa a\nabla(T^4) - \nabla(\nabla \cdot \mathbf{F_R}),$$

so that

$$\nabla(\nabla \cdot \mathbf{F_R}) = 4\kappa\sigma_{\rm sb}\nabla T^4 + 3\kappa^2 \mathbf{F_R}. \tag{51}$$

Here $\sigma_{\rm sb} = \frac{1}{4}ac$ is the Stefan–Boltzmann constant.

Equation (51) is the Eddington equation for the radiative flux $\mathbf{F_R}$, which is often written as

$$-L^2\nabla(\nabla \cdot \mathbf{F_R}) + 3\mathbf{F_R} + 4\sigma_{\rm sb}L\nabla T^4 = 0.$$

REFERENCES

[1] O. Baconneau, J.B. van den Berg, C.-M. Brauner, and J. Hulshof, *Multiplicity and stability of travelling wave solutions in a free boundary combustion-radiation problem*, European J. Appl. Math., 15 (2004), pp. 79–102.

[2] J.B. van den Berg, H. Elrofai, and J. Hulshof, *Stability analysis of travelling waves in a radiation-combustion free-boundary model for flame propagation*, in preparation.

[3] R. Blouquin, G. Joulin, and Y. Merhari, *Combustion regimes of particle-laden gaseous flames: Influences of radiation, molecular transports, kinetic-quenching, stoichiometry*, Combust. Theory Model., 1 (1997), pp. 217–242.

[4] C.-M. Brauner, J. Hulshof, and J.-F. Ripoll, *Existence of travelling wave solutions in a combustion-radiation model*, Discrete Contin. Dynam. Systems, 1 (2001), pp. 193–208.

[5] J.D. Buckmaster and T.L. Jackson, *The effects of radiation on the thermal-diffusive stability boundaries of premixed flames*, Combust. Sci. and Tech., 103 (1994), pp. 299–313.

[6] J.D. Buckmaster and G.S.S. Ludford, *Theory of Laminar Flames*, Cambridge University Press, Cambridge, UK, 1982.

[7] E.J. Doedel, A.R. Champneys, T.F. Fairgrieve, Y.A. Kuznetsov, B. Sandstede, and X. Wang, *AUTO97, Continuation and Bifurcation Software for Ordinary Differential Equations (with HomCont)*, 1997. Available via anonymous ftp from ftp.cs.concordia.ca from the directory pub/doedel/auto.

[8] J.W. Dold, R.W. Thatcher, and A.A. Shah, *High order effects in one step reaction sheet jump conditions for premixed flames*, Combust. Theory Model., 7 (2003), pp. 109–127.

[9] B. Dubroca and J.-L. Feugeas, *Etude théorique et numérique d'une hiérarchie de modèles aux moments pour le transfert radiatif*, C. R. Acad. Sci. Paris Sér. I. Math., 329 (1999), pp. 915–920.

[10] B. Dubroca and A. Klar, *Half moment closure for radiative transfer equations*, J. Comput. Phys., 180 (2002), pp. 584–596.

[11] A.S. Eddington, *The Internal Constitution of Stars*, Cambridge University Press, Cambridge, UK, 1926.

[12] G. Joulin and B. Deshaies, *On radiation-affected flame propagation in gaseous mixtures seeded with inert particles*, Combust. Sci. and Tech., 47 (1986), pp. 299–315.

[13] G. Joulin and M. Eudier, *The radiation-dominated propagation and extinction of slow, particle-laden gaseous flames*, in Proceedings of the 22nd International Symposium on Combustion, The Combustion Institute, Pittsburgh, PA, 1988, pp. 1579–1585.

[14] C. Levermore, *Moment closure hierarchies for kinetic theories*, J. Statist. Phys., 83 (1996), pp. 1021–1065.

[15] B.J. Matkowsky and G.I. Sivashinsky, *An asymptotic derivation of two models in flame theory associated with the constant density approximation*, SIAM J. Appl. Math., 37 (1979), pp. 686–699.
[16] M.F. Modest, *Radiative Heat Transfer*, Series in Mechanical Engineering, McGraw–Hill, New York, 1993.
[17] G.B. Rybicki and A.P. Lightman, *Radiative Processes in Astrophysics*, John Wiley and Sons, New York, 1979.
[18] F.A. Williams, *Combustion Theory*, Addison–Wesley, Reading, MA, 1994.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 433–446

POROUS MEDIA UPSCALING IN TERMS OF MATHEMATICAL EPISTEMIC COGNITION*

HWA-LUNG YU† AND GEORGE CHRISTAKOS†

Abstract. In this work we revisit the meaning of the term "solution" with respect to a mathematical model representing a physical media upscaling system. A central aim is to combine the science of mind with human mathematical ideas to solve real-world upscaling problems. We propose that in certain cases a (nonconventional) epistemic cognition solution (which assumes that the model describes incomplete knowledge about nature and focuses on conceptual mechanisms and thinking processes) can lead to a more realistic upscaling analysis than a (conventional) ontologic approach (which assumes that the model describes nature as is and focuses on form manipulations). In the present work we apply the epistemic cognition approach to the solution of the two-dimensional bounded porous media upscaling problem. A formal framework is presented, and implementation issues related to the epistemic cognition methodology are considered. Numerical experiments are presented involving effective conductivities in bounded two-dimensional spatial domains, and insight is gained by comparing the results to existing ontologic upscaling solutions. In addition to dealing with new and more general upscaling situations, the proposed approach can reproduce some well-known results, a fact that further demonstrates its power and nesting capabilities.

Key words. upscaling, porous media, epistemic, stochastic

AMS subject classifications. 15A15, 15A09, 15A23

DOI. 10.1137/040614438

1. Introduction.

In the porous media upscaling literature we find various ontologic techniques (analytical and numerical) for deriving useful values of effective hydraulic conductivities; see Cushman (1986), Dagan (1989), King (1989), Deutsch (1989), Rubin and Gomez-Hernandez (1990), Kitanidis (1990), Neuman and Orr (1993), Christakos, Miller, and Oliver (1993), Christakos, Hristopulos, and Miller (1995), Gelhar (1993), Wen and Gomez-Hernandez (1996), Hristopulos and Christakos (1997a and b, 1999), Tartakovsky et al. (2002), and Rubin (2003). Underlying these techniques is an ontologic view of reality, i.e., the organization inherent in the mathematical upscaling model describes nature in a realistic manner and, thus, one is focusing on form manipulations (e.g., extracting from the algebra effective conductivity values so that a system of equations is satisfied).

In an earlier work (Christakos, 2003), an alternative view to porous media upscaling was proposed based on an epistemic cognition approach (ECA), i.e., it involved the epistemically evaluated cognitive integration and processing of various forms of knowledge. The main ECA ideas were developed by Christakos (1992), and since then the approach has been applied with considerable success to a number of physical and life systems (e.g., Christakos, 1998, 2000; Kolovos et al., 2002; D'Or and Bogaert, 2003; Serre et al., 2003a and b; Douaik et al., 2004; Christakos et al., 2005; Quilfen et al., 2004; Parkin, Savelieva, and Serre, 2005). A recent review of the ECA conceptual framework and its various natural applications can be found in Christakos (2005) and references therein. A central aim of the ECA is to combine the science of

*Received by the editors September 3, 2004; accepted for publication (in revised form) June 10, 2005; published electronically December 2, 2005. The work has been supported by grants from NIEHS (P42 ES05948 and P30-ES10126).

http://www.siam.org/journals/siap/66-2/61443.html

†Department of Geography, San Diego State University, San Diego, CA 92182 (hlyu@geography.sdsu.edu, christak@geography.sdsu.edu).

mind with human mathematical ideas to solve real-world upscaling problems. One of its basic premises is that an examination of what constitutes a "solution" produces useful tools for physical modelling and suggests greater emphasis on sound scientific reasoning over technical criteria. Unlike the ontologic approach, the ECA seeks effective conductivity values that satisfy a set of epistemic principles (e.g., maximum information and adaptation) subject to all the core knowledge (including the system of equations mentioned above) as well as the various site-specific (and often uncertain) databases available. The term "epistemic" refers to the construction of models of the processes (perceptual and intellectual) by which knowledge and understanding about the physical system are achieved and communicated. In the ECA framework, the contribution of "cognition" is to identify basic knowledge-assimilation and problem-solving processes, which are then examined by means of the evaluative standards of epistemology. The organization structure inherent in the mathematical upscaling model expresses our incomplete knowledge about a real-world situation rather than nature itself, and its solution is neither a fixed nor an absolute matter in the conventional ontologic sense. Instead, an adequate solution is deeply rooted in the process that is concerned with conceptual mechanisms at work—not merely on form manipulations. A distinction is made, e.g., between the individual's way of thinking of a concept and its formal definition, thus distinguishing between mathematical upscaling solution as a mental activity vs. as a formal system.

In this work we apply the ECA to study upscaling in two-dimensional (2-D) porous media situations. A rigorous theoretical definition of the effective hydraulic conductivity (EHC) is given in terms of (2.1) below, which involves stochastic expectation operators. In practice, the EHC is often viewed as a representative conductivity value (spatial average or global property) associated with a porous media scale that is larger than the scale of the available data (Rubin, 2003). The implementation of the ECA starts by distinguishing between two major categories of knowledge, viz., the general knowledge base ($\mathcal{G}$-KB) and the specificatory knowledge base ($\mathcal{S}$-KB). The $\mathcal{G}$-KB may include human constructs like physical equations and stochastic laws, whereas the $\mathcal{S}$-KB consists of site-specific details of the specified porous medium, including exact ("hard") and uncertain ("soft") conductivity data and hydraulic head observations, as well as several secondary information sources. Subsequently, the upscaling reasoning we adopt consists of the following stages (a detailed discussion of the upscaling methodology and the relevant terminology can be found in Christakos, 2003):

1. Structural stage, which transforms the $\mathcal{G}$-KB available into a set of equations in terms of the EHC probability density function (pdf) model, $f_{\mathcal{G}}$. This transformation is achieved by means of a maximum expected information principle $\mathcal{T}$, which can be expressed, e.g., in terms of the Shannon information measures, i.e., $\mathcal{T}$: $\max_{f_{\mathcal{G}}} \overline{\log(f_{\mathcal{G}}^{-1})}$.

2. Specificatory stage, which represents the $\mathcal{S}$-KB in a form suitable for quantitative analysis and processing. Common forms include

(i) Exact numerical values across space (hard data).

(ii) Intervals, i.e., there is not a unique data value available at a spatial location but, instead, an interval of possible values (uncertain data).

(iii) Probability functions, i.e., the datum at the specified location is in the form of a probability distribution (uncertain data).

3. Integration stage, which blends the results of the previous stages by means of an adaptation principle $\mathcal{A}$, thus leading to the final solution in terms of an updated pdf model, $f_{\mathcal{K}}$, for the EHC (the subscript $\mathcal{K}$ denotes the total KB, i.e., the syn-

thesis of the $\mathcal{G}$- and the $\mathcal{S}$-KB). The $\mathcal{A}$ principle may involve Bayesian induction pdf, $f_{\mathcal{K}} = f_{\mathcal{G}}(\text{EHC}|\mathcal{S})$, or stochastic deduction pdf, $f_{\mathcal{K}} = f_{\mathcal{G}}(\mathcal{S} \rightarrow \text{EHC})$ and $f_{\mathcal{K}} = f_{\mathcal{G}}(\mathcal{S} \leftrightarrow \text{EHC})$; the symbols "|," "$\rightarrow$," and "$\leftrightarrow$" denote, respectively, Bayesian conditional (probability of EHC given $\mathcal{S}$), material conditional (probability that $\mathcal{S}$ implies EHC), and material biconditional (probability that EHC if and only if $\mathcal{S}$). In this work, the Bayesian pdf will be used to calculate EHC values.

Hence, there are several possibilities regarding the mathematical formulation of the ECA methodology 1–3 above. For details, see Christakos (2000, 2002, 2005). In the next section a mathematical formulation of the ECA is considered in the 2-D porous media upscaling context.

2. Formulation and solution of the 2-D upscaling problem. Consider the case of effective flow in a 2-D porous domain that is sufficiently characterized by the local mean value law (e.g., Rubin, 2003)

$$\overline{K(\mathbf{s})J_i(\mathbf{s})} = \sum_{j=1}^{2} K_{eff,ij}\overline{J_i(\mathbf{s})} \tag{2.1}$$

$(i = 1, 2)$, where $\mathbf{s} = (s_1, s_2)$ is the spatial location vector, the bar denotes stochastic expectation, $K(\mathbf{s})$ is the random conductivity field at point $\mathbf{s}$, $\overline{J_i(\mathbf{s})}$ is the mean hydraulic gradient in the i direction expressed in terms of boundary conditions (BC) and conductivity statistics, and the $K_{eff,ij}$ are the EHC components sought, i.e., $\mathbf{K}_{eff} = [K_{eff,11}, K_{eff,12} = K_{eff,21}, K_{eff,22}]^T$. The tensorial nature of EHC was established by Tartakovsky et al. (2002). Depending on the porous media flow situation, (2.1) can be associated with different BC. Note that (2.1) constitutes a local law. Nonlocal laws may be considered as well (e.g., for nonhomogeneous situations), but this is beyond the scope of the present work. In addition to the $\mathcal{G}$-KB or core knowledge that is expressed, in this case, by (2.1), in real-world situations a set of conductivity and/or hydraulic gradient measurements is also available at a number of points in space. These measurements constitute site-specific knowledge $\mathcal{S}$-KB that must be also taken into account in deriving meaningful EHC $\mathbf{K}_{eff}$ values.

When talking about "solving" (2.1) in the conventional ontologic sense, we mean extracting from the algebra $K_{eff,ij}$-values at specified points so that the system of (2.1) is satisfied (e.g., Zhang, 2002). In a different spirit, the proposed ECA considers that the intellectual content of a mathematical solution lies in its ideas—not in the symbols themselves—and it seeks $K_{eff,ij}$-values that satisfy a set of epistemic principles ($\mathcal{T}$ and $\mathcal{A}$) subject to all the $\mathcal{G}$-KB and $\mathcal{S}$-KB available. In view of the total KB considered above, $\mathcal{K}$ $(= \mathcal{G} \cup \mathcal{S})$, we can write the stochastic moment equations (2.1) as the following set of equations in a matrix form:

$$\begin{bmatrix} \Lambda_{s_1}[\kappa\zeta_1] \\ \Lambda_{s_1}[\kappa\zeta_2] \\ \vdots \\ \Lambda_{s_n}[\kappa\zeta_1] \\ \Lambda_{s_n}[\kappa\zeta_2] \end{bmatrix} - \begin{bmatrix} \Lambda_{s_1}[\zeta_1] & \Lambda_{s_1}[\zeta_2] & 0 \\ 0 & \Lambda_{s_1}[\zeta_1] & \Lambda_{s_1}[\zeta_2] \\ \vdots & \vdots & \vdots \\ \Lambda_{s_n}[\zeta_1] & \Lambda_{s_n}[\zeta_2] & 0 \\ 0 & \Lambda_{s_n}[\zeta_1] & \Lambda_{s_n}[\zeta_2] \end{bmatrix} \begin{bmatrix} K_{eff,11} \\ K_{eff,12} \\ K_{eff,22} \end{bmatrix} = 0, \tag{2.2}$$

where $\Lambda_{s_q}[\cdot] = \int d\kappa d\zeta_1 d\zeta_2 f_{\mathcal{K}}[\cdot]$ is an operator at each point $\mathbf{s}_q$ $(q = 1, \ldots, n)$ of the 2-D porous domain; the κ and ζ_i denote $K(\mathbf{s})$ and $J_i(\mathbf{s})$ realizations, respectively; and the $f_{\mathcal{K}} = f_{\mathcal{K}}(\mu, \mathbf{g})$ is the integrated pdf, in the sense that it integrates knowledge about the random fields $K(\mathbf{s})$ and $J_i(\mathbf{s})$ $(i = 1, 2)$.

TABLE 2.1
The g_β functions selected.

β	$g_\beta(\mathbf{s}_q)$, $q = 1, \ldots, n$
0	$g_0(\mathbf{s}_q) = 1$ (normalization)
1	$g_1(\mathbf{s}_q) = \kappa$
2	$g_2(\mathbf{s}_q) = \zeta_1$
3	$g_3(\mathbf{s}_q) = \zeta_2$
4	$g_4(\mathbf{s}_q) = [\kappa - \overline{K(\mathbf{s}_q)}]^2$
5	$g_5(\mathbf{s}_q) = [\zeta_1 - \overline{J_1(\mathbf{s}_q)}]^2$
6	$g_6(\mathbf{s}_q) = [\zeta_2 - \overline{J_2(\mathbf{s}_q)}]^2$
7	$g_7(\mathbf{s}_q) = [\kappa - \overline{K(\mathbf{s}_q)}][\zeta_1 - \overline{J_1(\mathbf{s}_q)}]$
8	$g_8(\mathbf{s}_q) = [\kappa - \overline{K(\mathbf{s}_q)}][\zeta_2 - \overline{J_2(\mathbf{s}_q)}]$
9	$g_9(\mathbf{s}_q) = [\zeta_1 - \overline{J_1(\mathbf{s}_q)}][\zeta_2 - \overline{J_2(\mathbf{s}_q)}]$

The shape of $f_{\mathcal{K}}$ as well as the form of the vectors μ and $\mathbf{g}$ depend on the principles $(\mathcal{T}, \mathcal{A})$ and the KB $(\mathcal{G}, \mathcal{S})$ of the ECA methodology above (see introduction); e.g., an important case is the exponential form

$$f_{\mathcal{K}} \propto \exp(\mu^T \mathbf{g}), \tag{2.3}$$

which is a general result obtained if we assume a maximum expected information principle (in the Shannon sense) in stage 1, an $\mathcal{S}$-KB consisting of hard (exact) $K(\mathbf{s})$ and $J_i(\mathbf{s})$ data in stage 2, and a Bayesian principle in stage 3. In the 2-D upscaling situation of interest in this work, the $\mathbf{g} = \{g_\beta;\ \beta = 0, 1, \ldots, N\}$ is a vector of functions of the $K(s)$ and $J_i(\mathbf{s})$ fields. For example, in Table 2.1 the g_β functions involve several one-point and two-point spatial statistics so that $N = 9$. In particular, the functions g_1–g_6 are associated with one-point statistics of $K(\mathbf{s})$ and $J_i(\mathbf{s})$ across space, and the g_7–g_9 with the two-point statistics; the g_0 is a normalization constraint (assuring that a mathematically proper pdf $f_{\mathcal{K}}$ is derived). Note that, in principle, any high-order or multiple-point spatial statistics can be incorporated into the upscaling analysis above.

The vector $\mu = \{\mu_\beta\}$ consists of space-dependent coefficients μ_β associated with $\mathbf{g}$ at every point $\mathbf{s}$. In this work these coefficients were calculated as follows: (i) A set of initial values $\mu = \mu^{(0)}$ is selected. (ii) The Bayesian updating of the values of the vector μ is expressed as

$$f(\mu|\mathcal{S}) \propto f(\mathcal{S}|\mu) f(\mu), \tag{2.4}$$

where $\mathcal{S}$ denotes the $K(\mathbf{s})$, $J_i(\mathbf{s})$ data available, as mentioned above, $f(\mathcal{S}|\mu)$ is the likelihood function, and the prior pdf $f_{\mathcal{G}}(\mu)$ is assumed to have a known form (e.g., Gaussian). (iii) The initial $\mu = \mu^{(0)}$ values are updated using (2.4) and the Monte Carlo Markov chain (MCMC), which assures fast convergence to the final $\mu = \mu^{(f)}$ values. These values are substituted into (2.2) and (2.3), which are subsequently solved for the EHC components, $K_{eff,ij}$. Numerical solution of (2.2) at $\mathbf{s}_q$ ($q = 1, \ldots, n$) can be derived by means of a regression technique using the least square criterion. In particular, the regression technique generates $\mathbf{K}_{eff}$ values so that the left-hand side of (2.2) is as close to zero as possible, in the least square sense, at every point $\mathbf{s}_q$ in the porous domain (see, also, Appendix A).

In the next section we discuss a few numerical experiments in an effort to gain insight regarding the performance of the proposed upscaling technique in the case of flow in 2-D bounded porous domains. The ECA is formulated in a way that can reproduce the results obtained by well-established techniques, which are its limiting cases (this property is sometimes called nesting).

3. Numerical experiences.

3.1. Experiment 1. We consider the simple case of uniform flow in a 2-D spatial domain A with an area of 12.8×12.8 units^2; Dirichlet boundary conditions (BC) are assumed on all sides of A; and the distribution of the hydraulic head $H(\mathbf{s})$ values at the boundaries are shown in Figure 1a. Hydraulic head is related to hydraulic gradient $J_i(\mathbf{s})$ by $J_i(\mathbf{s}) = -\partial H(\mathbf{s})/\partial s_i$ $(i = 1, 2)$. A hydraulic conductivity random field $K(\mathbf{s})$ is assumed to be lognormally distributed and exhibits a homogenous/isotropic spatial variation characterized by an exponential covariance model with mean $\overline{\log K} = 1$ unit, variance $\sigma^2_{\log K} = 0.1$ unit^2, and correlation range (length) $\epsilon_{\log K} = 0.8$ units. The above constitute part of the $\mathcal{G}$-KB of the porous media flow situation under consideration. Log-conductivity realizations were generated at $n = 50$ points $\mathbf{s}_q = (s_{q1}, s_{q2})$, $q = 1, \ldots, 50$, using the fast Fourier transform technique of Ruan and McLaughlin (1998). One such realization is plotted in Figure 1b, and then used to obtain hydraulic head values at the 50 points (Figure 1c) by means of a code based on the finite difference technique for solving saturated 2-D flow equations discussed in Wang and Anderson (1982). Using the K and H values at the 50 points as the $\mathcal{S}$-KB, the EHC values were derived by the proposed ECA upscaling technique as follows:

$$\mathbf{K}_{eff} = [K_{eff,11}, K_{eff,12} = K_{eff,21}, K_{eff,22}]^T = [2.83, 0.22, 2.53]^T \tag{3.1}$$

for the entire (bounded) domain A. Note that in a bounded domain the EHC $\mathbf{K}_{eff}$ values generally depend on the distance from the boundaries, although for uniform flow this dependency often dies out within a few correlation lengths from the boundaries. More specifically, the correlation length is defined as $\lambda = \int_0^\infty d\kappa \rho_{\log K}(\kappa)$, where $\rho_{\log K}$ is the correlation function of $\log K$. The λ assesses spatial dependence between hydraulic conductivity values. In this simulation example, the random field is exponentially correlated, in which case the correlation between two points dies out at a distance equal to 4λ (see Figure 2). Note that a similar conclusion concerning the effects of certain boundaries on EHC was presented in Paleologos, Neuman, and Tartakovsky (1996).

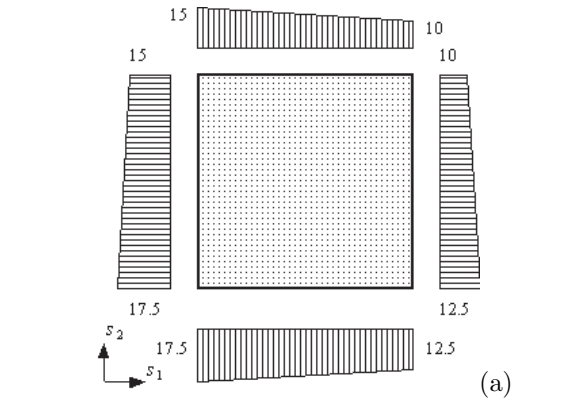
3.2. Experiment 2. Assuming a lognormal and statistically isotropic 2-D conductivity field in an unbounded (infinite) 2-D domain, a well-known upscaling result (e.g., Zhang, 2002) leads to the following EHC values:

$$\mathbf{K}_g = [K_g, 0, K_g]^T, \tag{3.2}$$

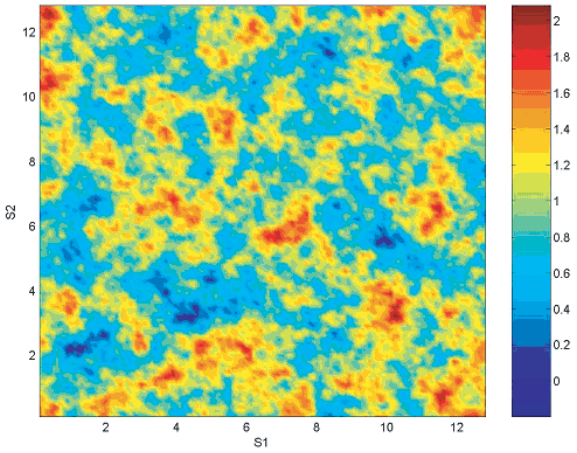
where K_g is the geometric mean of the $K(\mathbf{s})$ field. Despite certain differences in the underlying assumptions (e.g., bounded vs. unbounded domains), an attempt was made to compare the $\mathbf{K}_{eff}$ values obtained by the proposed upscaling technique with the $\mathbf{K}_g$ result of (3.2). Naturally, such an attempt should require an adequate simulation of the unboundedness condition of the $\mathbf{K}_g$ result, at least approximately. More specifically, in Figure 3 we plot the $K_{eff,11}$ and $K_{eff,22}$ values obtained by the proposed method assuming a series of 2-D conductivity $K(\mathbf{s})$ realizations (the $K_{eff,12}$ values are very close to zero in every realization and, thus, they are not plotted). Two cases were considered in this figure:

(i) Boundary effects were included, in the sense that the $K_{eff,11}$ and $K_{eff,22}$ values were the outcomes of the upscaling technique in Experiment 1 above (K and H values were used at points $\mathbf{s}_q$ located throughout the bounded domain A in Figure 1a).

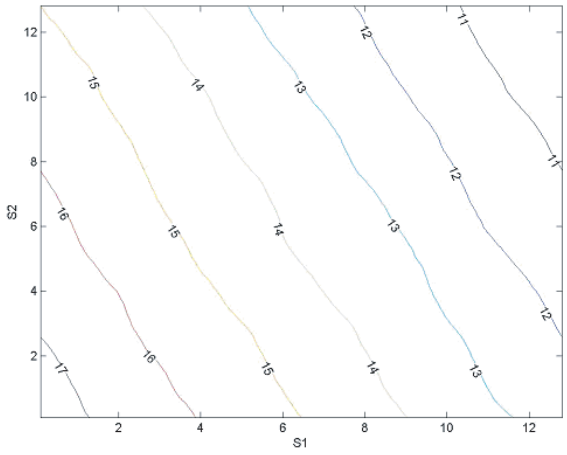
(ii) Approximately no boundary effects were assumed, in the sense that K and H



(a)



(b)



(c)

FIG. 1. (a) *Domain geometry and BC (hydraulic head distributions),* (b) *log-hydraulic conductivity realization, and* (c) *hydraulic head map.*

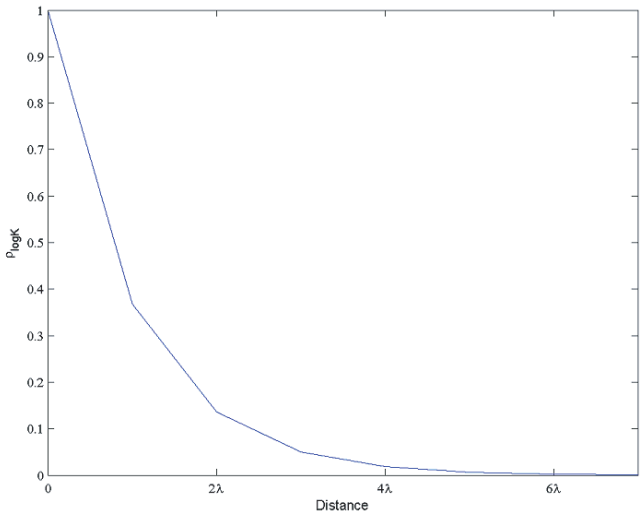


FIG. 2. *Conductivity correlation $\rho_{\log K}$ as a function of spatial distance (in λ units).*

values were used at points $\mathbf{s}_q$ located only within a small subdomain of A and away from its boundaries (in this way the boundary effects were somehow reduced and, hence, a better simulation of the unboundedness condition of the $\mathbf{K}_g$ result was achieved).

As should be expected, the $K_{eff,11}$ and $K_{eff,22}$ values differ for cases (i) and (ii) above. Note that the $K_{eff,11}$ and $K_{eff,22}$ values obtained in case (ii) are, in general, closer to the $\mathbf{K}_g$ values than the $K_{eff,11}$ and $K_{eff,22}$ values of case (i).

3.3. Experiment 3. Some more numerical tests concerning the (approximately) unbounded case (ii) of Experiment 2 above were considered here. More specifically, in Figure 4 we plot the $\mathbf{K}_{eff}$ values as a function of the $\log K$ variance $\sigma^2_{\log K}$. The calculated $K_{eff,11}$ and $K_{eff,22}$ values differ from K_g (since the assumptions of the unbounded K_g case do not exactly apply in the specific experimental setup considered here). Generally, the $K_{eff,11}$ and $K_{eff,22}$ values seem to follow the K_g variation, whereas the $K_{eff,12}$ values fluctuate around zero. Notice that as the $\sigma^2_{\log K}$ increases, the $K_{eff,11}$, $K_{eff,22}$ show larger fluctuations around K_g and the $K_{eff,12}$ larger fluctuations around zero. We subsequently plot the $\mathbf{K}_{eff}$ as a function of $\epsilon_{\log K}$ (Figure 5). For the specific 2-D situation, the calculated $\mathbf{K}_{eff}$ values are very close to the K_g values for $\epsilon_{\log K} < 1$; the $\mathbf{K}_{eff}$ values fluctuate considerably around the K_g values for larger $\epsilon_{\log K}$ values (which may be due to the small size of the domain considered). If the $\epsilon_{\log K}$ is too large, the EHC calculation will seriously depend on the locations from the boundaries. Furthermore, as the $\epsilon_{\log K}$ increases considerably, the EHC field is no longer isotropic ($K_{eff,11}$ exhibits a decreasing trend and $K_{eff,22}$ an increasing trend in Figure 5a). In Figure 6 we plot the $\mathbf{K}_{eff}$ components vs. the number n of points considered within the domain A. As might have been expected, after a certain number, $n = n^*$, further increasing the number n has no significant effect on the magnitude of the fluctuations of the EHC values around the values of the $\mathbf{K}_g$ result.

Several other cases can be handled by the proposed upscaling methodology. For

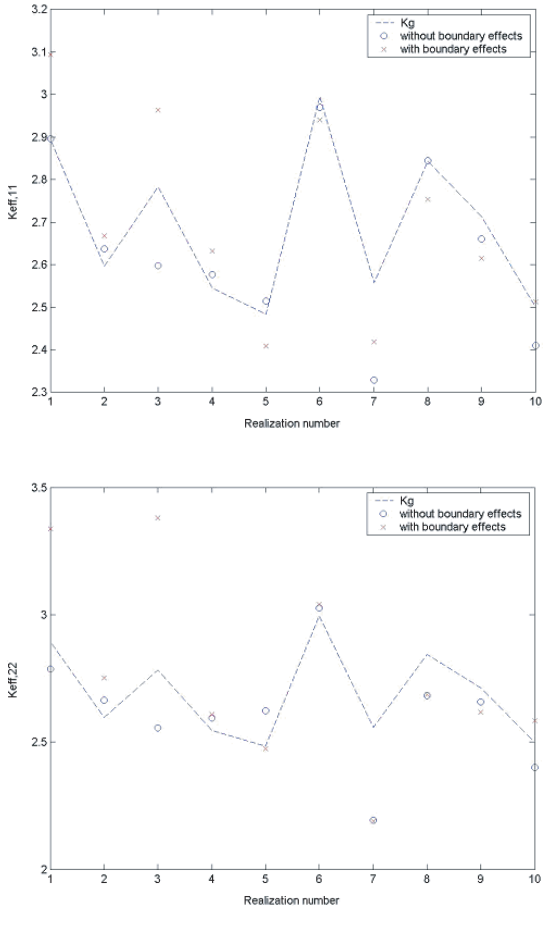


FIG. 3. *Realization number vs.* (a) *calculated* $K_{eff,11}$*, and* (b) *calculated* $K_{eff,22}$*; with boundary effects and without boundary effects. The corresponding* $\mathbf{K}_g$ *result is shown as a dashed line for comparison.*

illustration purposes, another numerical experiment with a different feature (pumping well) from the ones above is investigated next.

3.4. Experiment 4. In this experiment, the same arrangement as in Experiment 1 was considered with regards to the 2-D spatial domain A, the BC, and the hydraulic conductivity random field characteristics. However, in the present case a pumping well is assumed in the center of the domain A with a pumping rate of 100 units/time. Under these conditions, in Figure 7 we plot the log-conductivity field ($K_g = 2.73$) and the corresponding hydraulic head map. In this case, the calculated EHC components were found by the proposed upscaling technique, as follows: $\mathbf{K}_{eff} = [K_{eff,11}, K_{eff,12}, K_{eff,22}]^T = [2.87, 0.20, 2.83]^T$. We would like to conclude by noticing that a variety of different flow situations, BC, domain geometries, etc. can be studied by means of the upscaling technique. The discussion of some of these cases will be the topic of a future publication.

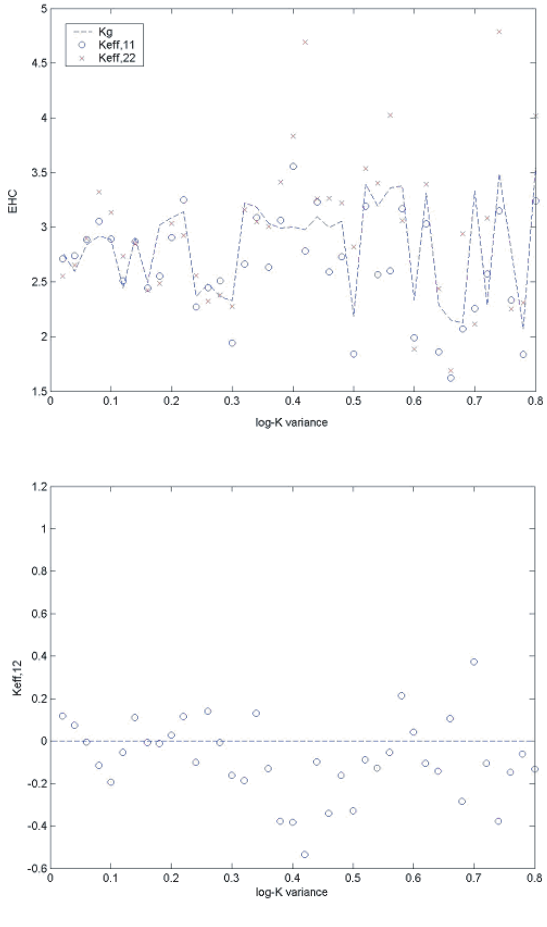
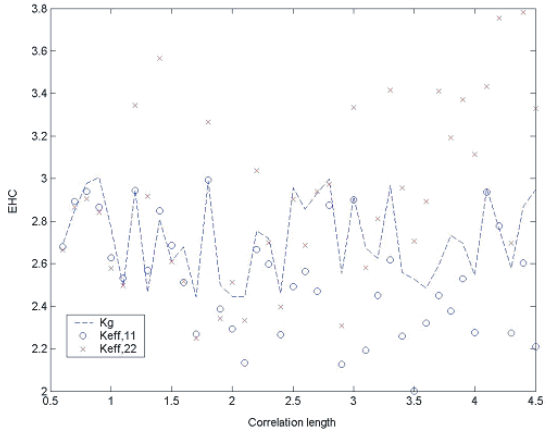
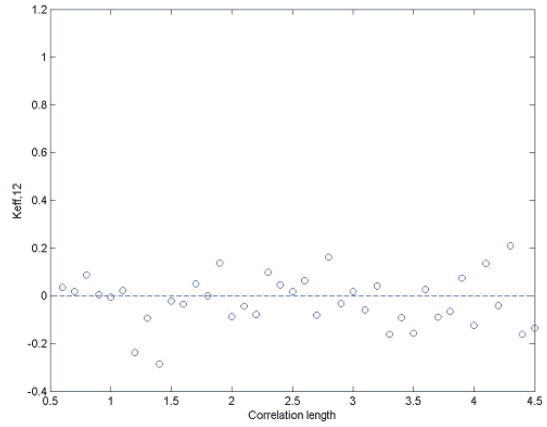


FIG. 4. Log K *variance* $\sigma^2_{\log K}$ *vs.* (a) *calculated* $K_{eff,ii}$ $(i = 1, 2)$*, and* (b) *calculated* $K_{eff,12}$. *The corresponding* $\mathbf{K}_g$ *result is shown as a dashed line for comparison.*

4. Summary. In this short paper, a new epistemic cognition approach was used to formulate and solve a bounded 2-D porous media upscaling problem. An ECA generates mind-based mathematical solutions of the upscaling problem, which do not rely solely on symbolic logic and form manipulations; the cognitive processes involved in the creation and conceptualization of the mathematical solutions are also considered. A systematic stochastic upscaling framework was presented and practical implementation issues were considered. Numerical experiments were discussed involving effective conductivities in 2-D domains, the effects of important flow parameters in the upscaling solution were examined, and comparisons were made with the results of previous ontologic upscaling solutions. Note that, in addition to dealing with new and more general porous media upscaling situations, the proposed approach can reproduce some well-known results, a fact that further demonstrates the



(a)



(b)

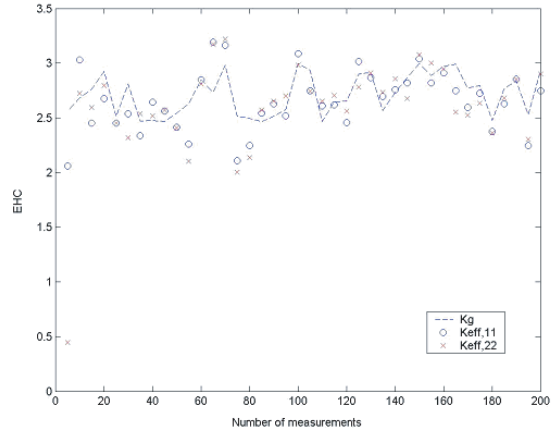
FIG. 5. *K correlation length $\epsilon_{\log K}$ vs.* (a) *calculated $K_{eff,ii}$ ($i = 1, 2$), and* (b) *calculated $K_{eff,12}$. The corresponding $\mathbf{K}_g$ result is shown as a dashed line for comparison.*

power of the ECA approach. By way of a summary, the ECA-based upscaling results obtained so far are promising, in which case further theoretical and application issues concerning the proposed upscaling technique will be the subject of future research.

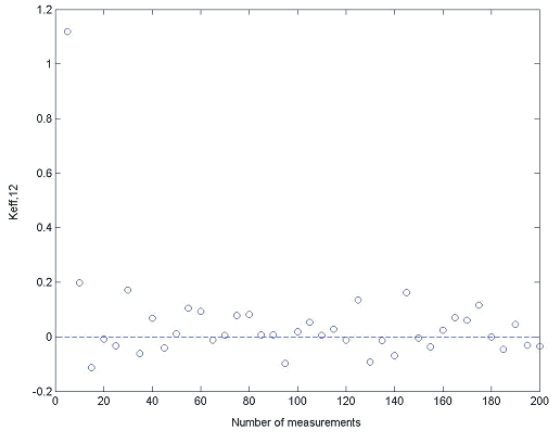
Appendix A.

A.1. The regression technique produces an EHC solution that satisfies (2.2) in the least square sense, at every point of the porous domain. The so-derived solution of (2.2) will have the form

$$\mathbf{K}_{eff} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{A.1}$$



(a)



(b)

FIG. 6. *Number n of points vs.* (a) *calculated* $K_{eff,ii}$ $(i = 1, 2)$*, and* (b) *calculated* $K_{eff,12}$*; the corresponding* $\mathbf{K}_g$ *result is shown as a dashed line for comparison.*

where

$$\mathbf{Y} = \begin{bmatrix} \Lambda_{s_1}[\kappa\zeta_1] \\ \Lambda_{s_1}[\kappa\zeta_2] \\ \vdots \\ \Lambda_{s_n}[\kappa\zeta_1] \\ \Lambda_{s_n}[\kappa\zeta_2] \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \Lambda_{s_1}[\zeta_1] & \Lambda_{s_1}[\zeta_2] & 0 \\ 0 & \Lambda_{s_1}[\zeta_1] & \Lambda_{s_1}[\zeta_2] \\ \vdots & \vdots & \vdots \\ \Lambda_{s_n}[\zeta_1] & \Lambda_{s_n}[\zeta_2] & 0 \\ 0 & \Lambda_{s_n}[\zeta_1] & \Lambda_{s_n}[\zeta_2] \end{bmatrix}, \text{ and } \mathbf{K}_{eff} = \begin{bmatrix} K_{eff,11} \\ K_{eff,12} \\ K_{eff,22} \end{bmatrix}.$$

Since the form (A.1) may involve numerical instabilities, it is recommended to reduce it by applying the QR decomposition on the matrix $\mathbf{X}$. The regression technique has been used in a variety of problems (see, e.g., White and Horne, 1987; Wen and Gomez-Hernandez, 1996).

A.2. Note that one set of EHC values is obtained from (2.2) or (A.1) that represents the EHC of the entire field. If, on the other hand, the EHC $\mathbf{K}_{eff}$ values at every point $\mathbf{s}_q$ $(q = 1, \ldots, n)$ within the porous domain are sought, the easiest way to

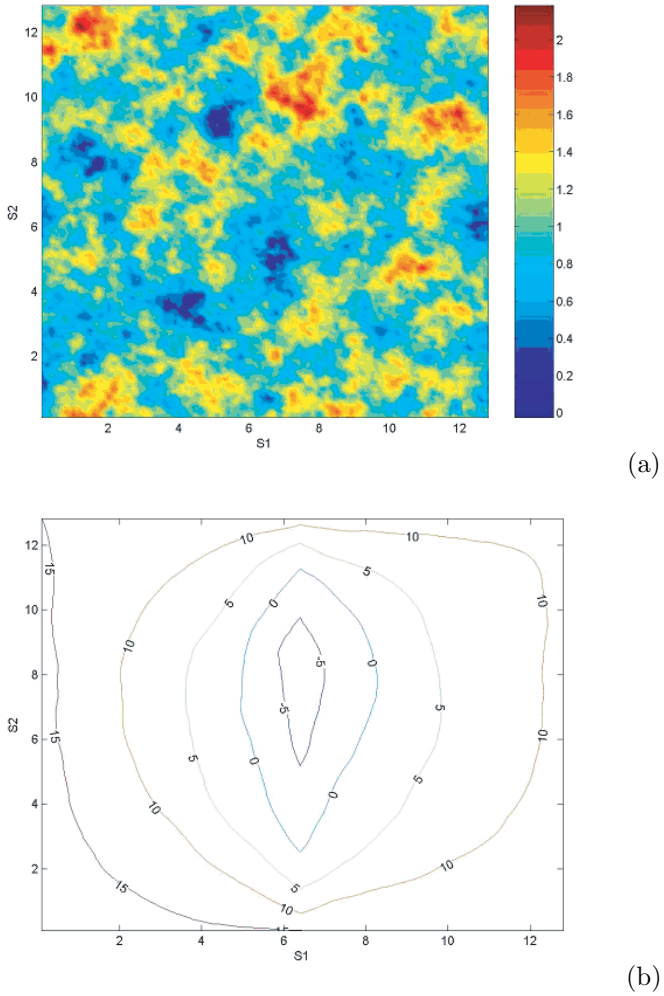


FIG. 7. (a) *Log-hydraulic conductivity realization; and* (b) *hydraulic head map.*

do this is to impose a constraint on (2.2); e.g., by solving (2.2) with the constraint

$$\text{(A.2)} \qquad \begin{cases} \Lambda_{\mathbf{s}_q}[K_{eff,11}\zeta_1 + K_{eff,12}\zeta_2 - \kappa\zeta_1] = 0, \\ \Lambda_{\mathbf{s}_q}[K_{eff,21}\zeta_1 + K_{eff,22}\zeta_2 - \kappa\zeta_2] = 0 \end{cases}$$

at every point $\mathbf{s}_q$, the resulting EHC values will be the local EHC values, $K_{eff,ij}(\mathbf{s}_q)$, at each point $\mathbf{s}_q$. In this case, if a single EHC value for the entire domain A is needed, one should calculate the arithmetic mean of the EHC over all n points, i.e., $\overline{K_{eff,ij}} = n^{-1}\sum_{q=1}^{n} K_{eff,ij}(\mathbf{s}_q)$. Assuming a homogeneous conductivity field $K(\mathbf{s})$, the arithmetic mean $\overline{K_{eff,ij}}$ and the previously obtained solution $K_{eff,ij}$ of (2.2) should be very close to each other.

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

REFERENCES

G. Christakos (1992), *Random Field Models in Earth Sciences*, Academic Press, San Diego (out of print). New edition, Dover, Mineola, NY, 2005.

G. Christakos (1998), *Spatiotemporal information systems in soil and environmental sciences*, Geoderma, 85, pp. 141–179.

G. Christakos (2000), *Modern Spatiotemporal Geostatistics*, Oxford University Press, New York.

G. Christakos (2002), *On a deductive logic-based spatiotemporal random field theory*, Teor. Ĭmovīr. Mat. Stat., 66, pp. 46–57; translation in Theory Probab. Math. Statist., 66 (2003), pp. 49–61.

G. Christakos (2003), *Another look at the conceptual fundamentals of porous media upscaling*, Stoch. Environ. Res. Risk Assess., 17, pp. 276–290.

G. Christakos, D. T. Hristopulos, and C. T. Miller (1995), *Stochastic diagrammatic analysis of groundwater flow in heterogeneous porous media*, Water Resour. Res., 31, pp. 1687–1703.

G. Christakos, C. T. Miller, and D. L. Oliver (1993), *Stochastic perturbation analysis of groundwater flow. Spatially variable soils, semi-infinite domains and large fluctuations*, Stoch. Hydrol. and Hydraul., 7, pp. 213–239.

G. Christakos (2005), *Recent conceptual developments in geophysical assimilation modelling*, Rev. Geophys., 43, pp. 1–10.

G. Christakos, R. A. Olea, M. L. Serre, H. L. Yu, and L. Wang (2005), *Interdisciplinary Public Health Reasoning and Epidemic Modelling: The Case of Black Death*, Springer-Verlag, New York.

J. H. Cushman (1986), *On measurement, scale, and scaling*, Water Resour. Res., 22, pp. 129–134.

G. Dagan (1989), *Flow and Transport in Porous Formations*, Springer-Verlag, Berlin.

C. V. Deutsch (1989), *Calculating effective absolute permeability in sandstone/shale sequences*, SPE Formation Evaluation, 4, pp. 343–348.

D. D'Or and P. Bogaert (2003), *Continuous-valued map reconstruction with the Bayesian maximum entropy*, Geoderma, 112, pp. 169–178.

A. Douaik, M. van Meirvenne, T. Toth, and M. L. Serre (2004), *Space-time mapping of soil salinity using probabilistic Bayesian maximum entropy*, Stoch. Environ. Res. Risk Assess., 18, pp. 219–227.

L. W. Gelhar (1993), *Stochastic Subsurface Hydrology*, Prentice–Hall, Englewood Cliffs, NJ.

D. T. Hristopulos and G. Christakos (1997a), *A variational calculation of the effective fluid conductivity of heterogeneous media*, Phys. Rev. E (3), 55, pp. 7288–7298.

D. T. Hristopulos and G. Christakos (1997b), *Diagrammatic theory of nonlocal effective hydraulic conductivity*, Stoch. Hydrol. and Hydraul., 11, pp. 369–395.

D. T. Hristopulos and G. Christakos (1999), *Renormalization group analysis of permeability upscaling*, Stoch. Environ. Res. Risk Assess., 13, pp. 131–160.

P. R. King (1989), *The use of renormalization for calculating effective permeability*, Transp. Porous Media, 4, pp. 37–58.

P. K. Kitanidis (1990), *Effective hydraulic conductivity for gradually varying flow*, Water Resour. Res., 26, pp. 1197–1208.

A. Kolovos, G. Christakos, M. L. Serre, and C. T. Miller (2002), *Computational BME solution of a stochastic advection-reaction equation in the light of site-specific information*, Water Resour. Res., 38, pp. 1318–1334.

S. P. Neuman and S. Orr (1993), *Prediction of steady-state flow in nonuniform geologic media by conditional moments: Exact nonlocal formalism, effective conductivities, and weak approximation*, Water Resour. Res., 29, pp. 341–364.

E. K. Paleologos, S. P. Neuman, and D. M. Tartakovsky (1996), *Effective hydraulic conductivity of bounded, strongly heterogeneous porous media*, Water Resour. Res., 32, pp. 1333–1341.

R. Parkin, E. Savelieva, and M. L. Serre (2005), *Oft geostatistical analysis of radioactive soil contamination*, in geoENV V-Geostatistics for Environmental Applications, P. Renard, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands.

Y. Quilfen, B. Chapron, F. Collard, and M. L. Serre (2004), *Calibration/validation of an altimeter wave period model and application to TOPEX/Poseidon and Jason-1 altimeters*, Marine Geodesy, 27, pp. 535–550.

F. Ruan and D. McLaughlin (1998), *An efficient multivariate random field generator using the fast Fourier transform*, Adv. in Water Resour., 21, pp. 385–399.

Y. Rubin (2003), *Applied Stochastic Hydrogeology*, Oxford University Press, New York.

Y. Rubin and J. J. Gomez-Hernandez (1990), *A stochastic approach to the problem of upscaling of conductivity in disordered media: Theory and unconditional numerical simulations*, Water Resour. Res., 26, pp. 691–701.

M. L. Serre, G. Christakos, H. Li, and C. T. Miller (2003a), *A BME solution of the inverse*

problem, Stoch. Environ. Res. Risk Assess., 17, pp. 354–369.

M. L. Serre, A. Kolovos, G. Christakos, and K. Modis (2003b), *An application of the holistochastic human exposure methodology to naturally occurring arsenic in Bangladesh drinking water*, Risk Analysis, 23, pp. 515–528.

D. M. Tartakovsky, A. Guadagnini, F. Ballio, and A. M. Tartakovsky (2002), *Localization of mean flow and equivalent transmissivity tensor for bounded randomly heterogeneous aquifers*, Transp. Porous Media, 49, pp. 41–58.

H. F. Wang and M. P. Anderson (1982), *Introduction to Groundwater Modelling: Finite Difference and Finite Element Methods*, W. H. Freeman, San Francisco.

X.-H. Wen and J. J. Gomez-Hernandez (1996), *Upscaling hydraulic conductivities in heterogeneous media: An overview*, J. Hydrology, 183, pp. ix–xxxii.

C. D. White and R. N. Horne (1987), *Computing absolute transmissivity in the presence of fine-scale heterogeneity*, Paper SPE, 16011, pp. 209–221.

D. Zhang (2002), *Stochastic Methods for Flow in Porous Media*, Academic Press, San Diego.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 447–467

A MIXTURE THEORY FOR THE GENESIS OF RESIDUAL STRESSES IN GROWING TISSUES II: SOLUTIONS TO THE BIPHASIC EQUATIONS FOR A MULTICELL SPHEROID*

ROBYN P. ARAUJO† AND D. L. SEAN MCELWAIN†

Abstract. This is the second paper in the series *A Mixture Theory for the Genesis of Residual Stresses in Growing Tissues*. While the first paper in the series elaborated a general formulation for such a theory, the present paper develops a simple biphasic model of residual stress evolution in a growing multicell spheroid comprising a linear-elastic cellular phase and an inviscid interstitial fluid. Both isotropic and anisotropic growth are considered in this study, highlighting the necessity to incorporate stress relaxation in order to predict an evolution of stresses over a period of growth.

The solutions to the biphasic equations are juxtaposed with the corresponding solutions to the single phase equations, illuminating the approximate nature of the single phase formulation for growing tissues. Moreover, the analysis demonstrates the significance of both interphase drag and the stress-relaxation characteristics of the solid phase in distinguishing between the single phase and multiphase paradigms.

Key words. multicell spheroid, poroelasticity, residual stresses, isotropic growth, anisotropic growth, diffusion, linear elasticity, biomechanics, constitutive equations, porous media

AMS subject classifications. 74A10, 74F10, 74L15, 76S05, 92B05

DOI. 10.1137/040607125

1. Introduction.

The genesis of growth-induced tissue stresses is an important consideration in the study of tumor growth. The experimental studies by Helmlinger et al. [15] on the growth of multicell tumor spheroids in agarose gels, for example, have demonstrated that "solid stress controls tumour growth at both the macroscopic and cellular levels, and thus influences tumour progression and delivery of therapeutic agents." Indeed, an increasing gel stiffness gave rise to smaller equilibrium-size spheroids, in addition to decreasing percentages of proliferating and apoptotic cells and increasing cellular densities. Later experiments by Koike et al. [20] illustrated that solid stress can facilitate the formation of spheroids in cell lines that do not form spheroids in free suspension, in addition to inhibiting their growth. Relieving this stress contributed to a loss of spheroid integrity. Furthermore, the experiments suggested that solid stress increases the synthesis of the extracellular matrix macromolecule, hyaluronan, by tumor cells.

While these two noteworthy experiments varied solid stress by changing the external hydrostatic pressure applied to the tumor spheroids at their outer boundaries, it is important to recognize that any *residual* stresses generated *within* the tumor either add to or relieve these externally applied stresses to increase or decrease the local state of stress. For this reason, an understanding of the evolution of residual stresses in growing tumors is of crucial importance.

Nevertheless, very few mathematical models of residual stress evolution in growing tissues have been proposed. Since residual stresses arise from incompatible growth, a mathematical model of such a phenomenon must consider the tissue's solid charac-

*Received by the editors January 5, 2004; accepted for publication (in revised form) July 5, 2005; published electronically December 2, 2005.
http://www.siam.org/journals/siap/66-2/60712.html
†School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia (s.mcelwain@qut.edu.au).

teristics via either an elastic or a viscoelastic constitutive law. Significantly, the vast majority of mechanical models of tumor growth employ fluid constitutive laws [5] which either incorporate [9, 11] or neglect [10, 24, 22] viscosity.

The first model to consider an evolution of residual stresses over a period of growth was that proposed by Jones et al. [16], who considered nutrient-regulated isotropic growth of a linear-elastic tissue—a model which highlighted the necessity to incorporate stress relaxation in the tissue's constitutive law in order to maintain boundedness of the solutions. The two models of tumor residual stress evolution proposed since that time have incorporated stress relaxation by various means—MacArthur and Please [21] by viscoelasticity and Araujo and McElwain [6] by anisotropic growth.

Nevertheless, all three mathematical models of residual stress evolution in growing tissues [6, 16, 21] are single phase models. The single phase approach consists of incorporating a source term in the balance of mass and a growth-expansion term in the constitutive equation, so that tissue mass is "created" in regions of cell proliferation and "destroyed" in regions of cell death. While this simple approach represents an excellent tool for studying the effects of the spatial nonuniformity of the growth process and is well suited to incorporating phenomenological relationships between growth and factors such as nutrient concentration and stress-modulated cell proliferation, it must be conceded that the behavior of any extracellular phases, such as the interstitial fluid, cannot be studied using this technique. In addition, a more serious shortcoming of this approach lies in the fact that the "Darcy-like" drag terms are neglected from the equilibrium equation, giving rise to an approximate form of the linear momentum equation which may not be valid in all situations.

The present paper is the second in a series of papers which studies the evolution of residual stresses using multiphase mechanics. The preceding paper in this series [1] (hereafter referred to as "Paper 1") presented a general method of deducing constitutive equations for multiphase materials which undergo a combination of continuous volumetric expansion and interphase mass exchange—a combination which reflects the process of biological growth. In this second paper, the simplest set of multiphase equations which allow residual stresses to develop—the biphasic equations for a linear elastic solid and an inviscid fluid—are solved in spherical polar coordinates in order to model the growth-induced stresses in a multicell spheroid.

Following the framework presented in Paper 1, a number of simplifying assumptions are made—constant volume fractions, intrinsically incompressible phases with equal densities, and a linear-elastic solid phase (regularized by anisotropic growth) with an inviscid fluid phase. While the assumption of constant volume fractions precludes a study of necrosis formation due to the comparatively high fluid content in such regions, the study of nonnecrotic tumors is not without basis. The experiments of Sutherland and Durand [27] demonstrated that multicell spheroids could reach a dormant size without central necrosis, suggesting an alternative cell loss mechanism in these tumors. It was the work of Kerr [18] and Kerr, Wyllie, and Currie [19] which demonstrated that apoptosis can always be detected in malignant neoplasms. The mathematical model by McElwain and Morris [23], which was proposed in response to these important experimental findings, was the first to study a tumor's arrival at a dormant state by incorporating apoptosis as a cell loss mechanism and is an antecedent to much of the subsequent mathematical literature relating to tumor development. It is the purpose of this paper to compare the solutions to these biphasic equations with their single phase counterparts, demonstrating the approximate nature of the single phase formulation for the stresses within the solid phase and illustrating the behavior of the interstitial fluid.

In section 2, the modelling equations underpinning the tissue's growth profile and the associated stress development are presented, with the solution procedure for this suite of equations being given in section 2.2. Numerical solutions are then presented in section 3 and are discussed in section 4. Section 5 presents some concluding remarks and outlines various avenues for future work in this important area of solid tumor growth research.

2. The mathematical model. A spherically symmetric multicell spheroid surrounded by a nutrient-rich medium and consisting exclusively of proliferating/apoptotic cells (the solid phase, with volume fraction ϕ_s) and interstitial fluid (the fluid phase, with volume fraction $\phi_f = 1 - \phi_s$) is considered. Both phases are assumed to be intrinsically incompressible with density ρ on account of the high water content of both phases. Following Jones et al. [16] and Araujo and McElwain [6], it is assumed that the growth of the spheroid is dependent upon the concentration, c, of a key nutrient which is supplied from the well-stirred surrounding medium and diffuses inwards. Assuming the nutrient diffusion to be very rapid, with constant diffusion coefficient, D_c, and the consumption of oxygen to be proportional to the local nutrient concentration, with proportionality constant, m, the nutrient profile maintains a pseudosteady-state distribution which satisfies

$$D_c \boldsymbol{\nabla}^2 c - mc = 0. \tag{2.1}$$

Further, the rate of mass supply to the solid (cellular) phase, Γ_s, due to the uptake of interstitial fluid during cell proliferation and growth is assumed to be proportional to both the nutrient concentration and the cell density ($\phi_s \rho$). Central to the development of the biphasic equations presented in Paper 1 was the assumption that the volume fractions of the two phases were constants, thereby precluding a consideration of necrotic regions which are characterized by a significantly higher porosity than the "live" tumor regions. For this reason, cell death is assumed to occur by apoptosis alone, with the associated rate of mass conversion to the fluid phase assumed proportional to the cell density. Hence,

$$\Gamma_s = \alpha \phi_s \rho c - k \phi_s \rho, \tag{2.2}$$

where α and k are the rate constants associated with cell proliferation and cell death, respectively. Now the balance of mass becomes

$$\boldsymbol{\nabla} \cdot \mathbf{v}_s = \frac{\Gamma_s}{\phi_s \rho} = \alpha c - k, \tag{2.3}$$

where $\mathbf{v}_s$ is the velocity of the solid phase. Recall from Paper 1 that the constitutive equation for the solid phase in the most general case of anisotropic growth is given by

$$\frac{D^s \mathbf{E}_s}{Dt} = \boldsymbol{\nabla} \cdot \mathbf{v}_s \Omega + \frac{1}{2\mu} \left(\frac{D^s \boldsymbol{\sigma}_s}{Dt} - \frac{1}{3} \frac{D^s}{Dt} (\mathrm{tr} \boldsymbol{\sigma}_s) \mathbf{I} + \phi_s \frac{D^s P}{Dt} (3\boldsymbol{\Omega} - \mathbf{I}) \right), \tag{2.4}$$

where $\mathbf{E}_s$ is the infinitesimal strain tensor, $\boldsymbol{\sigma}_s$ is the Cauchy stress tensor, P is the hydrostatic pressure arising due to the intrinsic incompressibility of the phases, $\mathbf{I}$ is the identity tensor, and $\frac{D^s \mathbf{E}_s}{Dt}$ is the material derivative of the strain tensor. Note that the corotational (or Jaumann [8, 17]) derivative of the constitutive equation has reduced to the material derivative in this case because of the assumed spherical symmetry. Note also that (2.4) is a rearrangement of (8.24) in Paper 1 to make the strain tensor

the subject of the equation, rather than the stress tensor. In addition, the anisotropy tensor has the form

$$\mathbf{\Omega} = \begin{bmatrix} \gamma_r & 0 & 0 \\ 0 & \gamma_\theta & 0 \\ 0 & 0 & \gamma_\phi \end{bmatrix}, \tag{2.5}$$

where spherical symmetry requires that $\gamma_\theta = \gamma_\phi$. Thus,

$$\gamma_r = 1 - 2\gamma_\theta, \tag{2.6}$$

since the trace of the anisotropy tensor must be unity (see [2]). Araujo and McElwain [6] have shown that, in order to incorporate stress relaxation via anisotropic growth, the anisotropic strain multipliers, γ_r and γ_θ, must be functions of the difference between the radial and circumferential stress components, $\beta_s = \sigma_{sr} - \sigma_{s\theta}$. Furthermore, a different set of anisotropic multipliers should apply to the two separate growth processes of cell proliferation (represented by αc in (2.3)) and apoptotic cell death (represented by k in (2.3)) since volumetric expansion should occur preferentially in the direction of least compressive stress, while volumetric contraction should occur preferentially in the direction of greatest compressive stress. Thus, the constitutive equation (2.4) becomes

$$\frac{D^s\mathbf{E}_s}{Dt} = \alpha c\mathbf{\Omega}_\eta - k\mathbf{\Omega}_\zeta + \frac{1}{2\mu}\left(\frac{D^s\boldsymbol{\sigma}_s}{Dt} - \frac{1}{3}\frac{D^s}{Dt}(\mathrm{tr}\boldsymbol{\sigma}_s)\mathbf{I} + \phi_s\frac{D^sP}{Dt}(3\mathbf{\Omega}_p - \mathbf{I})\right), \tag{2.7}$$

where

$$\mathbf{\Omega}_\eta = \begin{bmatrix} \eta_r & 0 & 0 \\ 0 & \eta_\theta & 0 \\ 0 & 0 & \eta_\phi \end{bmatrix} \tag{2.8}$$

represents the anisotropy tensor for the process of cell proliferation, and

$$\mathbf{\Omega}_\zeta = \begin{bmatrix} \zeta_r & 0 & 0 \\ 0 & \zeta_\theta & 0 \\ 0 & 0 & \zeta_\phi \end{bmatrix} \tag{2.9}$$

represents the anisotropy tensor for the process of cell death. Moreover, since P is a *hydrostatic pressure* acting on the mixture as a whole as a result of the intrinsic incompressibility of the phases, which is distributed to each phase by its volume fraction, the anisotropy tensor associated with its convected time derivative, $\mathbf{\Omega}_p$, should assume the isotropic value of

$$\mathbf{\Omega}_p = \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix}. \tag{2.10}$$

Following earlier work [6], the forms to be taken for these anisotropic strain multipliers in the present paper will be

$$\eta_r = \frac{1}{2} + \frac{1}{2}\left(\lambda\beta_s + \tanh^{-1}\left(-\frac{1}{3}\right)\right) \tag{2.11}$$

and

$$\zeta_r = \frac{1}{2} - \frac{1}{2}\left(\lambda\beta_s + \tanh^{-1}\left(\frac{1}{3}\right)\right). \tag{2.12}$$

These relations, depicted in Figure 1, allow the multipliers to adopt their isotropic values of $\frac{1}{3}$ when $\beta_s = \sigma_{sr} - \sigma_{s\theta} = 0$, while asymptotically approaching their maximum and minimum values of unity and zero, respectively, as β_s becomes large. Moreover, the stress-relaxation parameter, λ, reflects how readily the anisotropic nature of the growth process responds to the prevailing stresses. (Note that by the convention adopted here, compressive stresses are *negative*.)

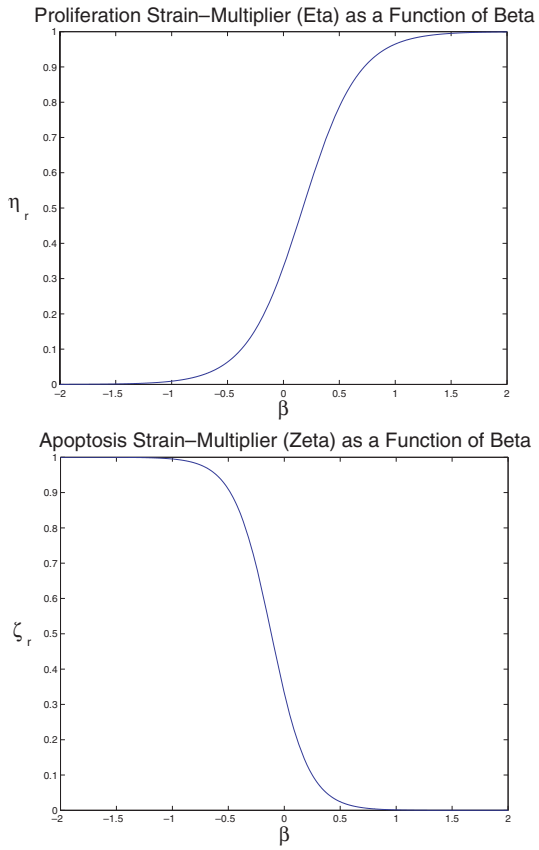


FIG. 1. *The dependence of the growth-strain multipliers on* β $(\lambda = 3)$.

Noting that, in spherical polar coordinates,

$$\frac{D^s\mathbf{E}_s}{Dt} = \begin{bmatrix} \frac{\partial v_s}{\partial r} & 0 & 0 \\ 0 & \frac{v_s}{r} & 0 \\ 0 & 0 & \frac{v_s}{r} \end{bmatrix}, \tag{2.13}$$

where v_s is the radial component of $\mathbf{v}_s$, the radial component of (2.7) is

$$\begin{aligned}\frac{\partial v_s}{\partial r} &= \alpha c\eta_r - k\zeta_r + \frac{1}{2\mu}\frac{D^s}{Dt}\left(\sigma_{sr} - \frac{1}{3}(\sigma_{sr} + 2\sigma_{s\theta})\right)\\ &= \alpha c\eta_r - k\zeta_r + \frac{1}{3\mu}\frac{D^s\beta_s}{Dt}.\end{aligned} \tag{2.14}$$

This equation may also be deduced from a consideration of the circumferential component of (2.7) by recalling that $\eta_\theta = \frac{1}{2}(1 - \eta_r)$. In the special case of isotropic growth, where $\eta_r = \zeta_r = \frac{1}{3}$, (2.14) reduces to

$$\frac{\partial v_s}{\partial r} = \frac{v_s}{r} + \frac{1}{2\mu}\frac{D^s\beta_s}{Dt}. \tag{2.15}$$

Combining (2.1) through (2.14) with the remaining biphasic equations presented in Paper 1 gives rise to the following suite of modelling equations:
Diffusion of nutrient:

$$D_c\boldsymbol{\nabla}^2 c - mc = 0. \tag{2.16}$$

Balance of mass:

$$\boldsymbol{\nabla}\cdot\mathbf{v}_s = \alpha c - k, \tag{2.17}$$

$$\rho\frac{D^s\phi_s}{Dt} = 0, \tag{2.18}$$

$$\boldsymbol{\nabla}\cdot(\phi_s\mathbf{v}_s + \phi_f\mathbf{v}_f) = 0. \tag{2.19}$$

Constitutive equations:

$$\frac{\partial v_s}{\partial r} = \alpha c\eta_r - k\zeta_r + \frac{1}{3\mu}\frac{D^s\beta_s}{Dt}, \tag{2.20}$$

where

$$\beta_s = \sigma_{sr} - \sigma_{s\theta}, \tag{2.21}$$

with

$$\eta_r = \frac{1}{2} + \frac{1}{2}\left(\lambda\beta_s + \tanh^{-1}\left(-\frac{1}{3}\right)\right)$$

and

$$\zeta_r = \frac{1}{2} - \frac{1}{2}\left(\lambda\beta_s + \tanh^{-1}\left(\frac{1}{3}\right)\right),$$

for anisotropic growth and $\eta_r = \zeta_r = \frac{1}{3}$ for isotropic growth,

$$\boldsymbol{\sigma}_f = -\phi_f P\mathbf{I}. \tag{2.22}$$

Momentum equations:

$$\boldsymbol{\nabla} \cdot \boldsymbol{\sigma}_s + \kappa(\mathbf{v}_f - \mathbf{v}_s) = 0, \tag{2.23}$$

$$\phi_f \boldsymbol{\nabla} P = -\kappa(\mathbf{v}_f - \mathbf{v}_s). \tag{2.24}$$

The momentum equations (2.23) and (2.24) may also be combined to give a single equation relating the stress tensor of the solid phase, $\boldsymbol{\sigma}_s$, to the hydrostatic pressure, P,

$$\boldsymbol{\nabla} \cdot \boldsymbol{\sigma}_s = \phi_f \boldsymbol{\nabla} P. \tag{2.25}$$

2.1. Nondimensionalization. The following dimensionless quantities are defined:

$$\hat{\mathbf{r}} = \frac{\mathbf{r}}{L}, \qquad \hat{t} = \frac{t}{T}, \qquad \hat{c} = \frac{c}{C_0}, \qquad \hat{\mathbf{v}} = \frac{\mathbf{v}T}{L}, \qquad \hat{\boldsymbol{\sigma}} = \frac{\boldsymbol{\sigma}}{2\mu}, \qquad \hat{P} = \frac{P}{2\mu},$$

where C_0 is the nutrient concentration at the tumor edge (assumed constant), with $L = \sqrt{\frac{D_c}{m}}$ and $T = \frac{1}{\alpha C_0}$. This transformation allows the modelling equations to be expressed in spherical polar coordinates in the following dimensionless forms:
Nutrient profile:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial c}{\partial r}\right) - c = 0. \tag{2.26}$$

Balance of mass:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 v_s\right) = c - \epsilon, \tag{2.27}$$

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \phi_s v_s + r^2 \phi_f v_f\right) = 0. \tag{2.28}$$

Constitutive equations:

$$\frac{\partial v_s}{\partial r} = c\eta_r - \epsilon\zeta_r + \frac{1}{3\mu}\frac{D^s \beta_s}{Dt}, \tag{2.29}$$

$$\sigma_{fr} = \sigma_{f\theta} = -\phi_f P. \tag{2.30}$$

Momentum equations:

$$\frac{\partial \sigma_{sr}}{\partial r} + \frac{2\beta_s}{r} + \chi\left(v_f - v_s\right) = 0, \tag{2.31}$$

$$\phi_f \frac{\partial P}{\partial r} = -\chi\left(v_f - v_s\right), \tag{2.32}$$

or

$$\frac{\partial}{\partial r}\left(\sigma_{sr} + \phi_f P\right) + \frac{2\beta_s}{r} = 0, \tag{2.33}$$

where the *hat* notation has been omitted for clarity, with $\epsilon = \frac{k}{\alpha C_0}$ and $\chi = \frac{\kappa L^2}{2\mu T}$. Thus, the parameter ϵ represents the rate of cell death as a proportion of the maximum possible growth rate. For a given diffusion length scale and tumor doubling time, the parameter χ gives a measure of the relative importance of interphase drag in comparison with the elasticity of the tissue.

The suite of equations (2.26) through (2.33) is now to be solved subject to the boundary conditions $v_s = 0$ and c finite at $r = 0$, and $\sigma_{fr} = \sigma_{f\theta} = P = 0$ and $c = 1$ at the tumor's outer boundary.

2.2. Solution procedure. Let $r = a(t)$ denote the position of the tumor boundary at time t, so that the tumor occupies the region $0 \leq r \leq a(t)$. Now, integrating (2.26) subject to the boundary conditions $c(a,t) = 1$ and finite c at $r = 0$ gives

$$c(r,t) = \frac{a \sinh r}{r \sinh a}. \tag{2.34}$$

Substituting (2.34) into (2.27) and integrating subject to the boundary condition $v_s = 0$ at $r = 0$ now gives

$$v_s(r,t) = \frac{a(r\cosh r - \sinh r)}{r^2 \sinh a} - \frac{\epsilon r}{3}. \tag{2.35}$$

Thus, the nutrient-regulated equilibrium size of the tumor is given by $a = a^*$, where

$$\coth a^* - \frac{1}{a^*} - \frac{\epsilon a^*}{3} = 0. \tag{2.36}$$

Now (2.28) may be integrated, noting that $v_f = v_s = 0$ at $r = 0$, to give

$$v_f = -\frac{\phi_s}{\phi_f} v_s = \frac{a\phi_s(\sinh r - r\cosh r)}{r^2 \phi_f \sinh a} + \frac{\phi_s \epsilon r}{3\phi_f}. \tag{2.37}$$

Substituting (2.35) and (2.37) into (2.32) gives

$$\frac{\partial P}{\partial r} = \chi \frac{v_s}{\phi_f^2} = \frac{\chi}{\phi_f^2}\left(\frac{a(r\cosh r - \sinh r)}{r^2 \sinh a} - \frac{\epsilon r}{3}\right), \tag{2.38}$$

which may be integrated subject to the boundary condition $P = 0$ at $r = a$ (i.e., scaling the fluid pressure in the medium surrounding the tumor to zero) to give

$$P = \frac{\chi}{\phi_f^2}\left(\frac{a \sinh r}{r \sinh a} - \frac{\epsilon}{6}\left(r^2 - a^2\right) - 1\right). \tag{2.39}$$

The hydrostatic pressure in the fluid, $\boldsymbol{\sigma}_f$, may now be determined from (2.22), which gives

$$\sigma_{fr} = \sigma_{f\theta} = \sigma_f = \frac{\chi}{\phi_f}\left(1 + \frac{\epsilon}{6}(r^2 - a^2) - \frac{a \sinh r}{r \sinh a}\right). \tag{2.40}$$

In order to proceed, (2.29) must now be solved numerically for $\beta_s = \sigma_{sr} - \sigma_{s\theta}$. In this paper, the integration is performed using the modified Lax–Wendroff scheme presented in [16] for isotropic growth and [6] for anisotropic growth. As noted by Araujo and McElwain [3], this modified scheme "is very useful for a growing domain since it consists of a fixed number of evenly-distributed moving gridpoints. Computing times are significantly reduced when compared with the method of characteristics, the latter method having the additional disadvantages of unequally-spaced gridpoints and the tendency to develop instabilities for solutions at large times." The interested reader is referred to [16] and [6] for the details of this numerical method.

Having determined β_s, (2.31) may then be integrated using a Runge–Kutta scheme (for example) to give the radial stress component in the solid phase, σ_{sr}, subject to the boundary condition $\sigma_{sr} = 0$ at $r = a$. The corresponding circumferential component, $\sigma_{s\theta}$, is then to be determined from the definition of β_s.

Note that residual stresses are those which exist in a solid body in the absence of, or in addition to, the stresses caused by external loads. The assumption of a zero hydrostatic pressure at the periphery of the tumor ($\sigma_r = P = 0$ at $r = a$ being the boundary conditions used to integrate (2.31) and (2.38)) for all time allows the model to emphasize the residual nature of any induced stresses, since they arise in the absence of any external loads.

3. Results. In the solutions to follow, the parameter ϵ will assume the value of 0.1, which implies a low cell death rate in comparison with the maximum possible cell proliferation rate. Jones et al. [16] explain that this assumption gives rise to a tumor structure in which the radius of the tumor boundary at equilibrium (when the processes of cell proliferation and cell death are in balance) is large in comparison with the diffusion length scale, $L = \sqrt{\frac{D_c}{m}}$. Indeed, the nutrient concentration is only significant in a region within a distance of L from the tumor surface, with the equilibrium tumor size being achieved by a large growth rate in the thin region near the tumor surface and a low death rate throughout the tumor. This tumor structure is most representative of experimentally observed multicell spheroids and avascular tumors [12, 13, 14, 18, 19]. Figure 2 depicts the solutions to (2.26), (2.27), and (2.28) based on this assumption.

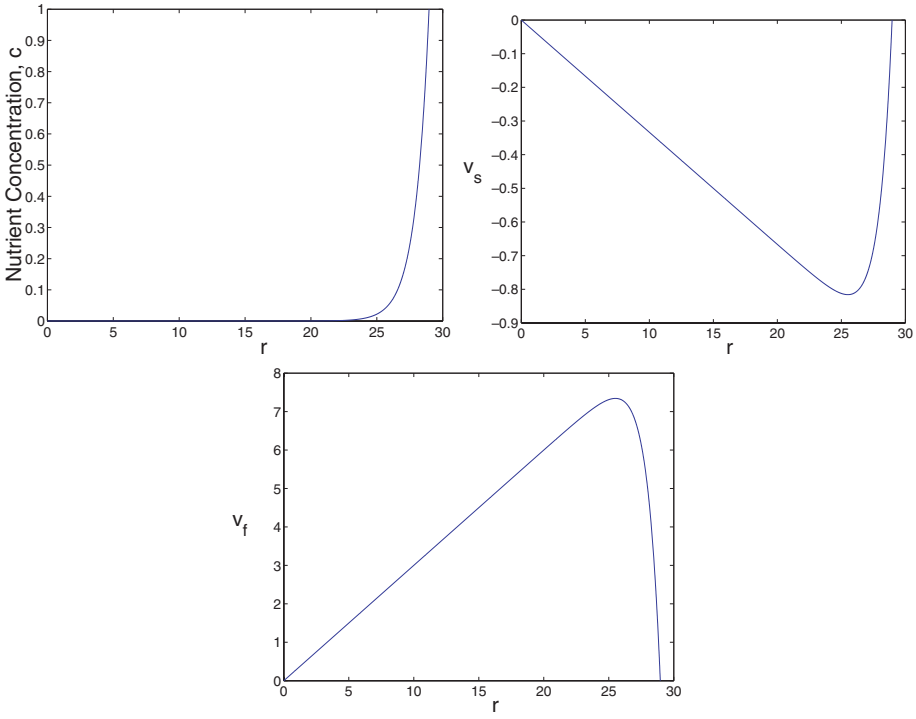


FIG. 2. *Distribution of nutrient, c, velocity profile for the cellular phase, v_s, and velocity profile for the interstitial fluid, v_f.* $\epsilon = 0.1$*;* $\phi_f = 0.1$.

Furthermore, to emphasize the development of growth-induced stresses for large times, most of the solutions for stresses depict the (admittedly artificial) situation in which the tumor begins stress-free at its nutrient-regulated equilibrium size, neglecting

the initial transient period required to attain this size. Figure 10 gives an example of the evolution of stresses in a tumor which begins stress-free at a size much smaller than the equilibrium size.

In addition to presenting the solutions of the biphasic equations developed in the present paper, Figures 3, 4, 7, 8, and 9 present the solutions of the single phase equations developed by Jones et al. [16] (in the case of isotropic growth) and Araujo and McElwain [6] (in the case of anisotropic growth) by way of comparison with the biphasic solutions. These single phase equations were summarized in Paper 1, in juxtaposition with the biphasic equations.

Note that, with the exception of Figure 10, the solutions for β_s are not shown since the difference between the radial and circumferential stress components in the solid phase is unaffected by the choice of single phase or multiphase frameworks.

3.1. Solutions to the stress equations with isotropic growth. Figures 3 and 4 show the solutions to the stress equations for isotropic growth, for which $\mathbf{\Omega}_\eta = \mathbf{\Omega}_\zeta = \frac{1}{3}\mathbf{I}$, or $\eta_r = \eta_\theta = \zeta_r = \zeta_\theta = \frac{1}{3}$, for ascending values of the interphase drag parameter, χ.

These solutions illustrate the problem with the combination of elasticity and isotropic growth—a combination which is unable to impart the crucial property of stress relaxation to the tissue's constitutive law and gives rise to singular behavior in the evolution of the stress profiles. Thus, although this combination is sufficient to give general information about the spatial variations in stresses, the elasticity must be regularized in order to predict the evolution of stresses over a period of growth. This may be accomplished by either incorporating a viscous term into the tissue's constitutive law, giving rise to a viscoelastic constitutive equation, or, as is the case here, by allowing the tissue to expand and contract anisotropically in response to the induced stress field. For anisotropic growth to regularize the elasticity of the tissue, expansion due to cell proliferation must occur in the direction of least compressive stress, while contraction due to cell death must occur in the direction of greatest compressive stress. In the context of spherical symmetry, these directional characteristics of the growth process are conferred by the anisotropy tensor which directs the expansion and contraction into the radial and circumferential principal directions. In this way, anisotropic growth imparts a *pseudoviscoelasticity* to growing tissues.

Nevertheless, these isotropic solutions enable comparisons to be made between the single phase and biphasic mathematical paradigms. As these results illustrate, a value as small as $\chi = 10^{-4}$ gives rise to stress profiles which differ only subtly from their single phase counterparts. By contrast, a larger value of $\chi = 10^{-1}$ affects the stress profiles quite dramatically.

Figures 5 and 6 present the distributions of the hydrostatic pressure, P, as well as the hydrostatic pressure in the fluid, σ_f. Note that these two pressures are unaffected by the choice of isotropic or anisotropic growth.

3.2. Solutions to the stress equations with anisotropic growth. Figures 7 through 9 show the solutions to the stress equations for anisotropic growth, for which $\mathbf{\Omega}_\eta$ and $\mathbf{\Omega}_\zeta$ are as defined by (2.8), (2.9), (2.11), and (2.12), for ascending values of the interphase drag parameter, χ.

These solutions illustrate the stress-relaxation properties of anisotropic growth, enabling steady-state stress profiles to prevail. With the exception of Figure 9, the "moderate" value of $\lambda = 10$ has been used for the stress relaxation parameter. A comparison of the results of the single phase and biphasic equations reveals that

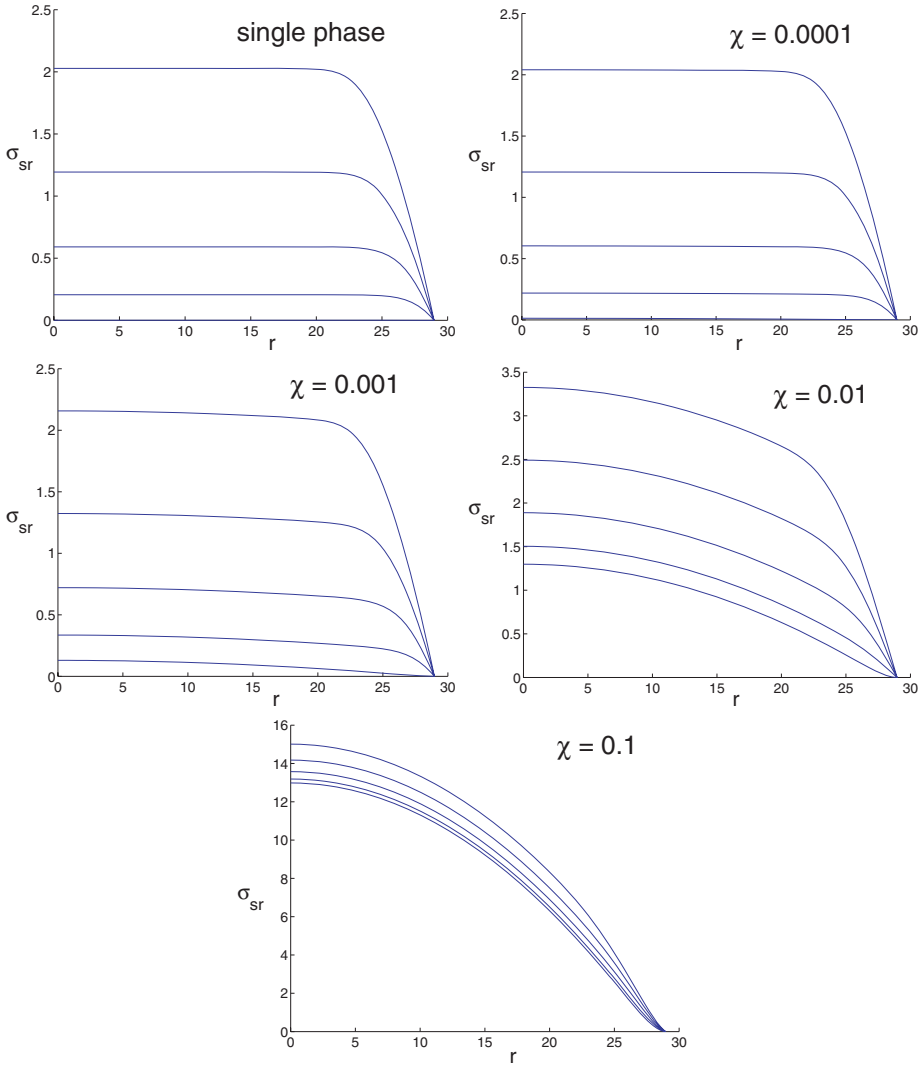


FIG. 3. *The distribution of radial stresses in the solid phase, σ_{sr}, for a single phase model, and for a biphasic model with $\chi = 0.0001, 0.001, 0.01, 0.1$ as labelled, for $t = 0$ to $t = 8$ in equal time increments. Growth is isotropic. The tumor is initially stress-free and at its nutrient-regulated equilibrium size ($a \backsim 29$), with $\epsilon = 0.1$. $\phi_f = 0.1$.*

even extremely small interphase drag, with $\chi = 10^{-4}$, gives rise to stress profiles which differ markedly from the single phase stress profiles. Figure 9 illustrates how this effect is accentuated by the larger stress-relaxation parameter of $\lambda = 100$. This intriguing result will be discussed in the next section. (It is also interesting to note that in Figures 7 and 8, when the comparatively large value of $\chi = 0.1$ has been used, the effect of introducing the second (fluid) phase is so pronounced that the evolution of solid stresses is almost imperceptible in comparison.)

4. Discussion. This paper illustrates the mechanical behavior of a spherically symmetric biphasic model of a growing multicell spheroid comprising a linear-elastic solid phase and an inviscid interstitial fluid phase. The growth process itself, which

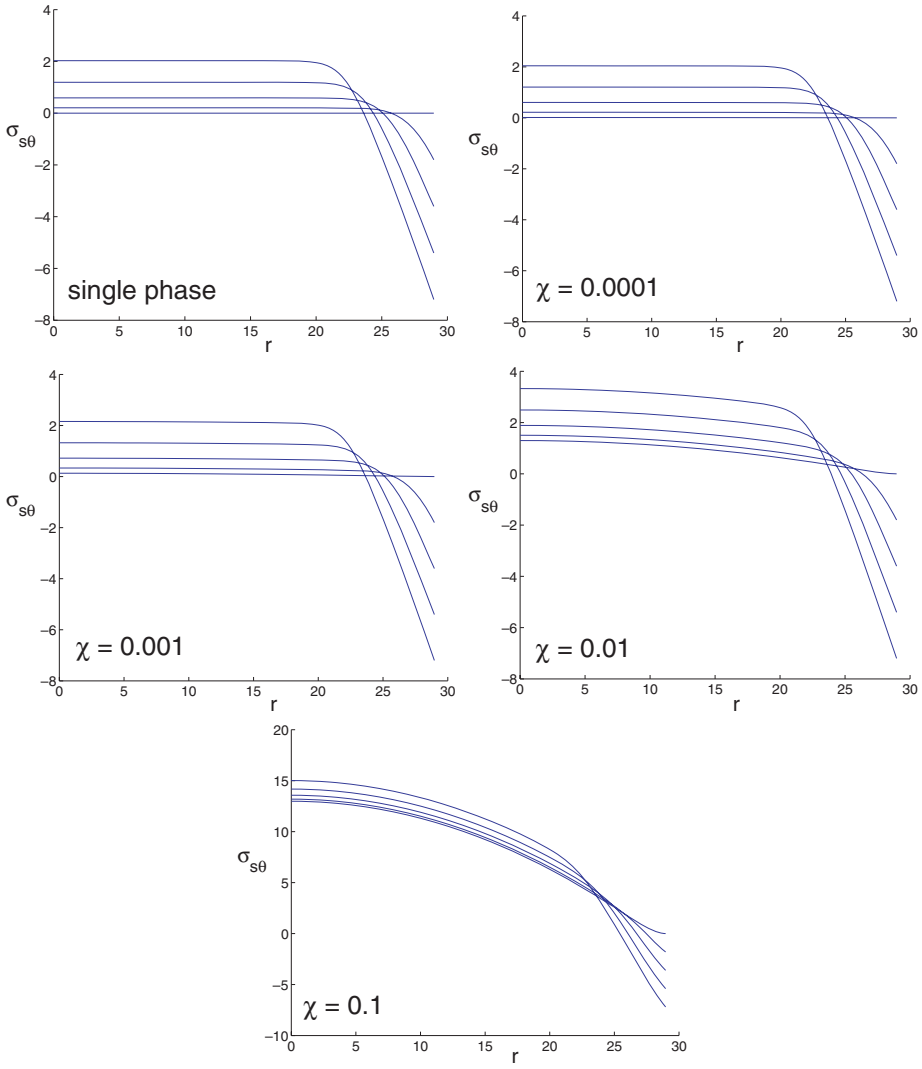


FIG. 4. *The distribution of circumferential stresses in the solid phase, $\sigma_{s\theta}$, for a single phase model, and for a biphasic model with $\chi = 0.001, 0.01, 0.1$ as labelled, for $t = 0$ to $t = 8$ in equal time increments. Growth is isotropic. The tumor is initially stress-free and at its nutrient-regulated equilibrium size ($a \backsim 29$), with $\epsilon = 0.1$. $\phi_f = 0.1$.*

may occur either isotropically or anisotropically, is regulated by an inwardly diffusing nutrient.

The solutions to the modelling equations give information about the distribution of nutrient concentration, c, within the tumor, the tumor's nutrient-regulated dormant radius, a^*, and the behavior of the individual phases—their movement (solid velocity, v_s, and fluid velocity, v_f) and their stress profiles (σ_f in the fluid phase, and σ_{sr} and $\sigma_{s\theta}$ for the radial and circumferential stress components, respectively, in the solid phase, as well as the hydrostatic pressure P acting on both phases). Of these quantities, several are independent of whether a single phase or multiphase modelling approach is taken—c, v_s, and a^*. Moreover, where a biphasic approach is adopted,

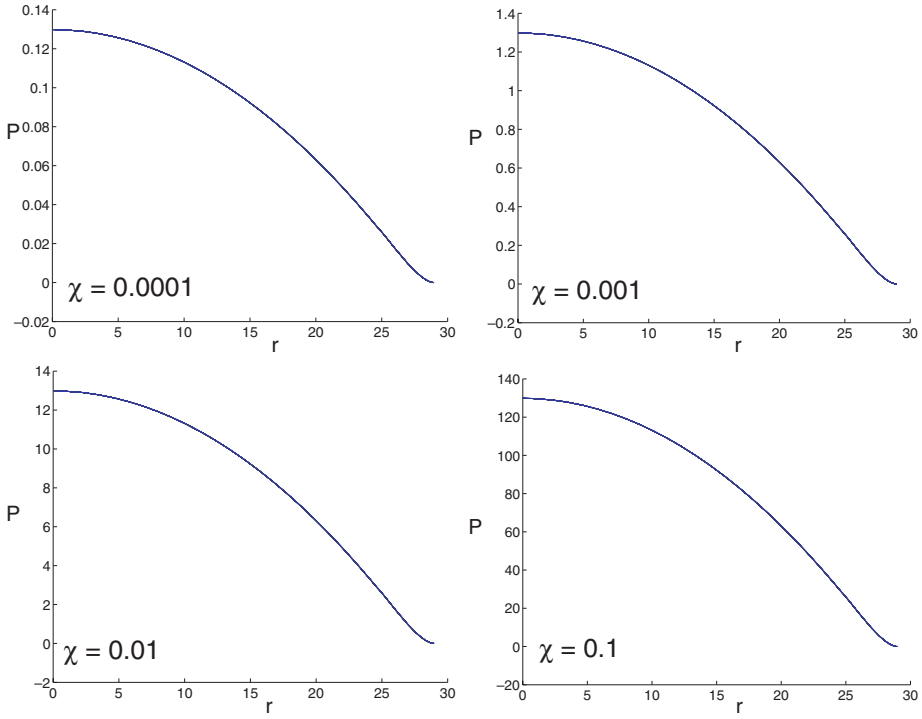


FIG. 5. *The distribution of the hydrostatic pressure acting on the mixture, P, for a biphasic model with $\chi = 0.001, 0.01, 0.1$ as labelled, for $t = 0$ to $t = 8$ in equal time increments. The tumor is initially stress-free and at its nutrient-regulated equilibrium size ($a \backsim 29$), with $\epsilon = 0.1$. $\phi_f = 0.1$.*

the variables associated with the fluid phase are unaffected by the choice of isotropic growth (which neglects stress relaxation) or anisotropic growth (which incorporates stress relaxation)—P, σ_f, and v_f.

The stresses in the solid phase, however, depend crucially on whether a single phase or a multiphase framework is employed. Comparing (2.25) and (2.33) with their single phase counterparts, $\boldsymbol{\nabla} \cdot \boldsymbol{\sigma}_s = \mathbf{0}$ and $\frac{\partial \sigma_r}{\partial r} + \frac{2\beta}{r} = 0$ (see Paper 1), gives some interesting insights into the distinction between the two mathematical paradigms. Whereas in the case of a spherically symmetric single phase solid the radial stress component (see (2.33)) is given by

$$\sigma_r = -\int_0^r \left(\frac{2\beta}{\hat{r}}\right) d\hat{r},$$

the multiphase momentum equation predicts a radial stress of

$$\sigma_{sr} = \phi_f P - \int_0^r \left(\frac{2\beta_s}{\hat{r}}\right) d\hat{r},$$

where β or β_s is determined from compatibility considerations via the constitutive equation (2.29). Thus, the two expressions for radial stress differ by the quantity $\phi_f P$, being the proportion of the hydrostatic pressure P distributed to the fluid phase. This pressure arises from the mechanical interactions between the two phases

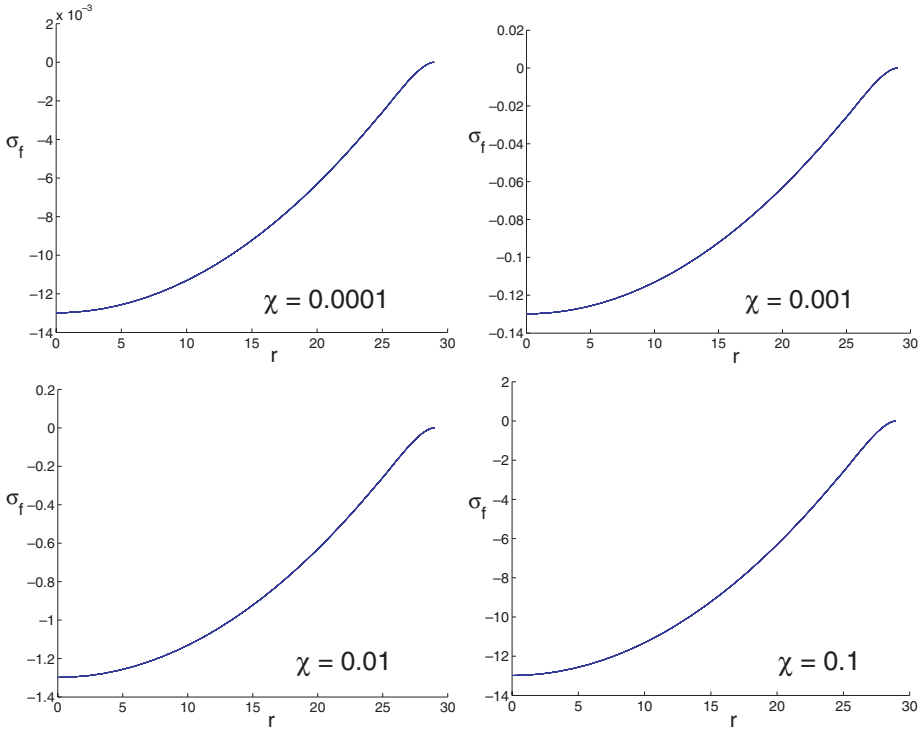


FIG. 6. *The distribution of the hydrostatic pressure in the fluid phase, σ_f, for a biphasic model with $\chi = 0.001, 0.01, 0.1$ as labelled, for $t = 0$ to $t = 8$ in equal time increments. The tumor is initially stress-free and at its nutrient-regulated equilibrium size ($a \backsim 29$), with $\epsilon = 0.1$. $\phi_f = 0.1$.*

and from (2.32) is given by

$$\phi_f P = -\chi \int_0^r (\mathbf{v}_f(\hat{r}) - \mathbf{v}_s(\hat{r}))\, d\hat{r}.$$

Therefore, since the circumferential stress component is given by $\sigma_{s\theta} = \sigma_{sr} + \beta_s$ from (2.21), *both* the circumferential and the radial stress components are augmented by the extra stress created by the relative motion between the phases. Note that this extra stress is directly proportional to the interphase drag parameter, χ.

Thus, in a multiphase model, solid stresses may be decomposed into two components—the temporally constant hydrostatic pressure,[1] $\phi_f P$, and a time-dependent component. Theoretically, if the processes of cell proliferation and cell death (apoptosis) were to cease, the hydrostatic component would vanish, leaving only the (previously) time-dependent component. The existence of the latter in the absence of any externally applied loads attests to its *residual* nature. Thus, the residual stresses in the solid phase are, of themselves, unaffected by the presence of additional phases; nevertheless, they do not occur in isolation but are augmented by a hydrostatic pressure through the presence of additional phases.

[1]Note that P is only constant with time once the tumor has reached its nutrient-regulated equilibrium size. When the tumor size is still evolving, the distribution of P is also evolving (see Figure 10).

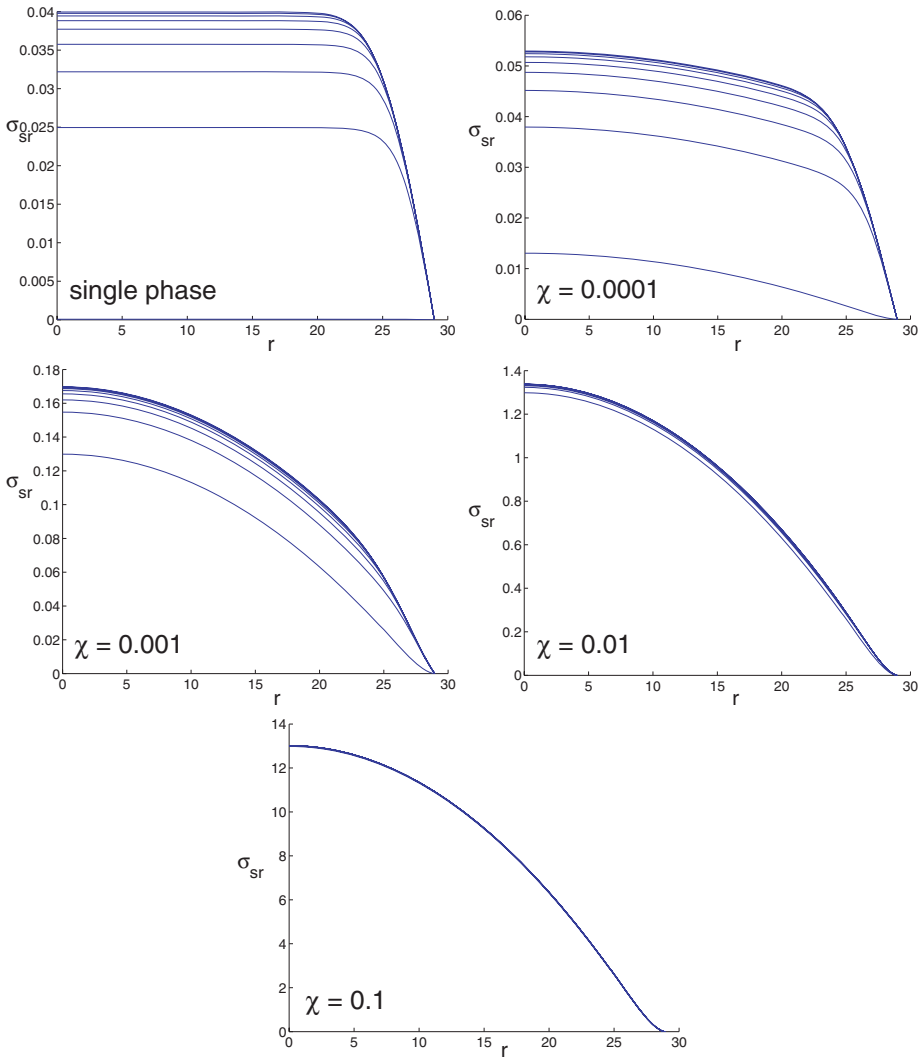


FIG. 7. *The distribution of radial stresses in the solid phase, σ_{sr}, for a single phase model, and for a biphasic model with $\chi = 0.0001, 0.001, 0.01, 0.1$ as labelled, for $t = 0$ to $t = 8$ in equal time increments. Growth is anisotropic with $\lambda = 10$. The tumor is initially stress-free and at its nutrient-regulated equilibrium size ($a \backsim 29$), with $\epsilon = 0.1$. $\phi_f = 0.1$.*

This study also demonstrates the important result that a consideration of multiple phases does nothing to alleviate the singularities associated with the combination of isotropic growth and an elastic constitutive law. Jones et al. [16] had supposed that this measure could resolve the problem, postulating that "in regions of the tumour where the number of live cells is low, the compensatory increase in the number of dead cells and extracellular water should prevent the stress tensor increasing indefinitely over time."

As one would intuitively expect, the magnitude of the interphase drag parameter, χ, strongly influences the extent to which the single phase and multiphase paradigms diverge, as reflected in Figures 3 through 8. In the case of isotropic growth, the

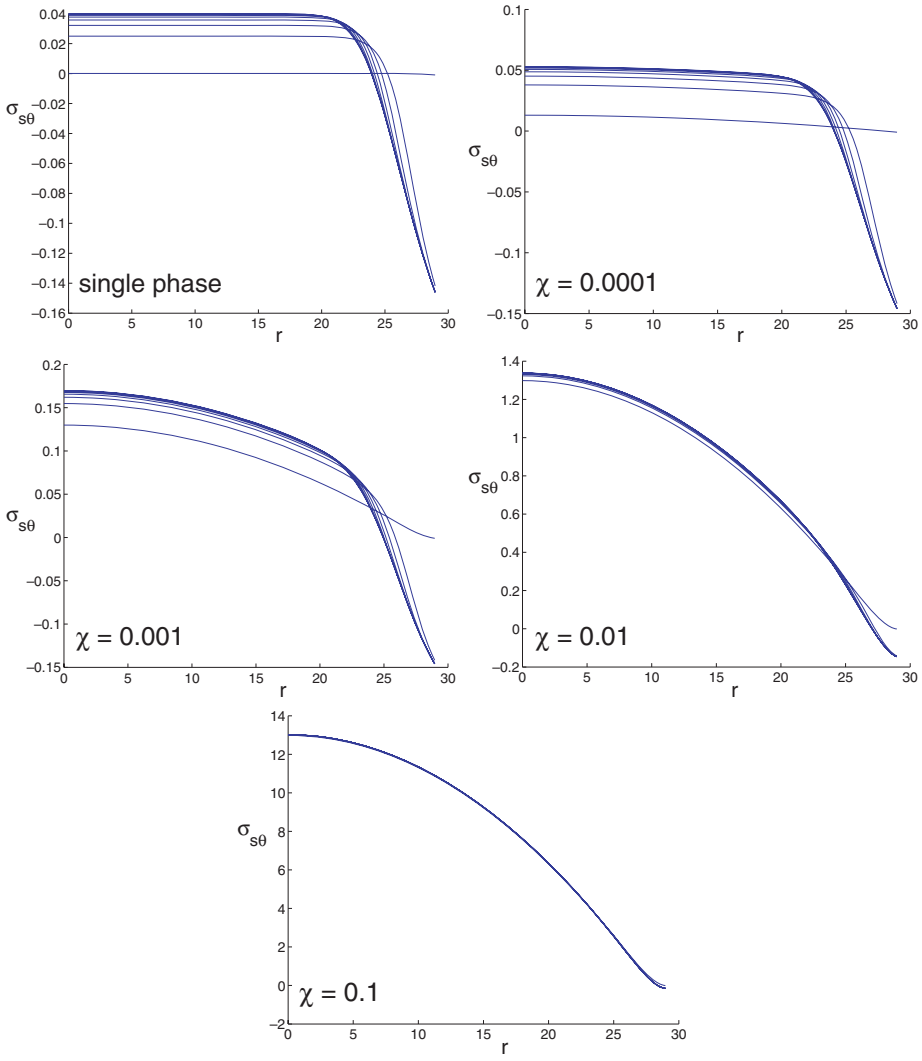


FIG. 8. *The distribution of circumferential stresses in the solid phase, $\sigma_{s\theta}$, for a single phase model, and for a biphasic model with $\chi = 0.0001, 0.001, 0.01, 0.1$ as labelled, for $t = 0$ to $t = 8$ in equal time increments. Growth is anisotropic with $\lambda = 10$. The tumor is initially stress-free and at its nutrient-regulated equilibrium size ($a \backsim 29$), with $\epsilon = 0.1$. $\phi_f = 0.1$.*

solutions are noticeably affected by the choice of modelling framework for $\chi > 0.01$, with $\chi > 0.1$ affecting the stress profiles quite dramatically. Interestingly, as noted in section 3.2, the stress profiles are much more sensitive to the choice of modelling framework when growth is anisotropic, where the biphasic results contrast with their single phase counterparts quite remarkably even with the minute value of $\chi = 0.0001$. This intriguing result may be understood by noting that, when growth is anisotropic, the evolution of residual stresses is gradually arrested in order to permit a steady-state stress profile at the tumor's nutrient-regulated dormant size. The greater the stress relaxation (corresponding to larger values of λ), the smaller are the steady-state stresses, and the shorter the time required to reach the steady-state profile. Therefore,

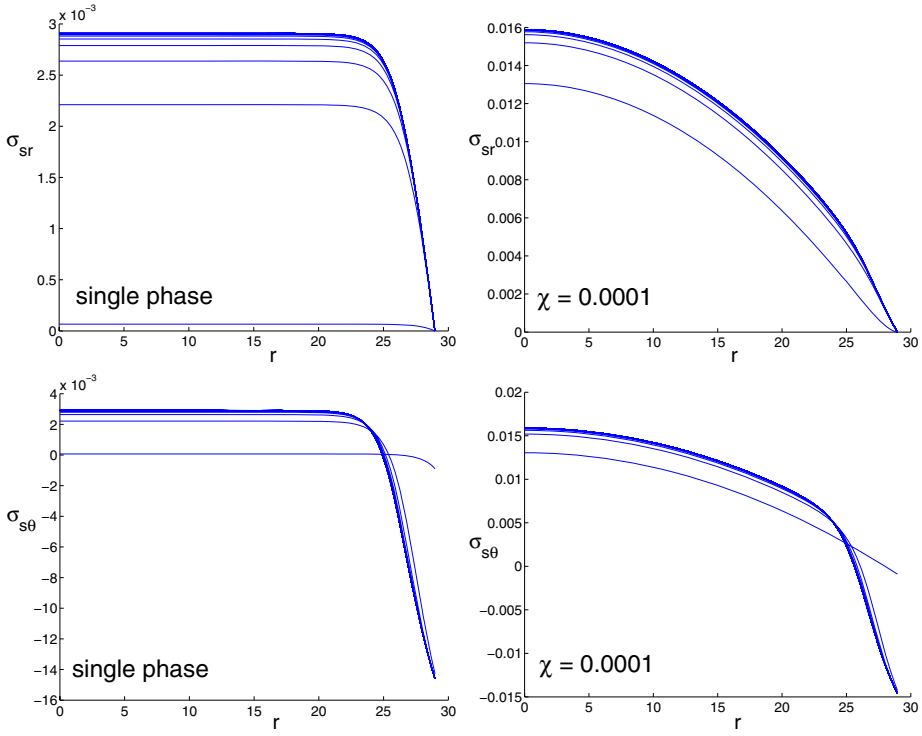


FIG. 9. *Radial stress distributions (top row) and circumferential stress distributions (bottom row) for a single phase model, and a biphasic model with* $\chi = 0.0001$*, as marked. Growth is anisotropic with* $\lambda = 100$*.* $t = 0$ *to* $t = 8$ *in equal time increments. The tumor is initially stress-free and at its nutrient-regulated equilibrium size (*$a \backsim 29$*), with* $\epsilon = 0.1$*.* $\phi_f = 0.1$*.*

the greater the stress relaxation, the more significant is the hydrostatic pressure, P, *in comparison with* the residual stresses. (This is to be compared with the single phase framework in which *only* residual stresses are induced.) As shown by (2.24) and illustrated in Figure 5, P is proportional to χ (all other things being equal), which implies that for a given value of χ, the difference between the stresses predicted by the single phase and multiphase models is more significant for a greater value of the stress-relaxation parameter.

Therefore, it is the combination of stress relaxation (a characteristic of the solid (cellular) phase, which may be associated with either anisotropic growth or viscoelasticity) and interphase drag (associated with the existence of multiple phases) which separates the single phase and multiphase approaches. *Both* properties must be considered in order to determine if single phase techniques give a reasonable approximation to the stresses in the solid phase. This is a most significant outcome of this study, given that stress-relaxation *must* occur in growing biological tissues. As demonstrated by Jones et al. [16] and by Figures 3 and 4 in the present paper, failure to incorporate this crucial property causes stresses to evolve indefinitely, eventually becoming infinite.

Thus, this study deracinates the assertion by Skalak [26] that volumetric growth of biological tissues is entirely analogous to thermal expansion—an analogy which lies at the very heart of the single phase paradigm.

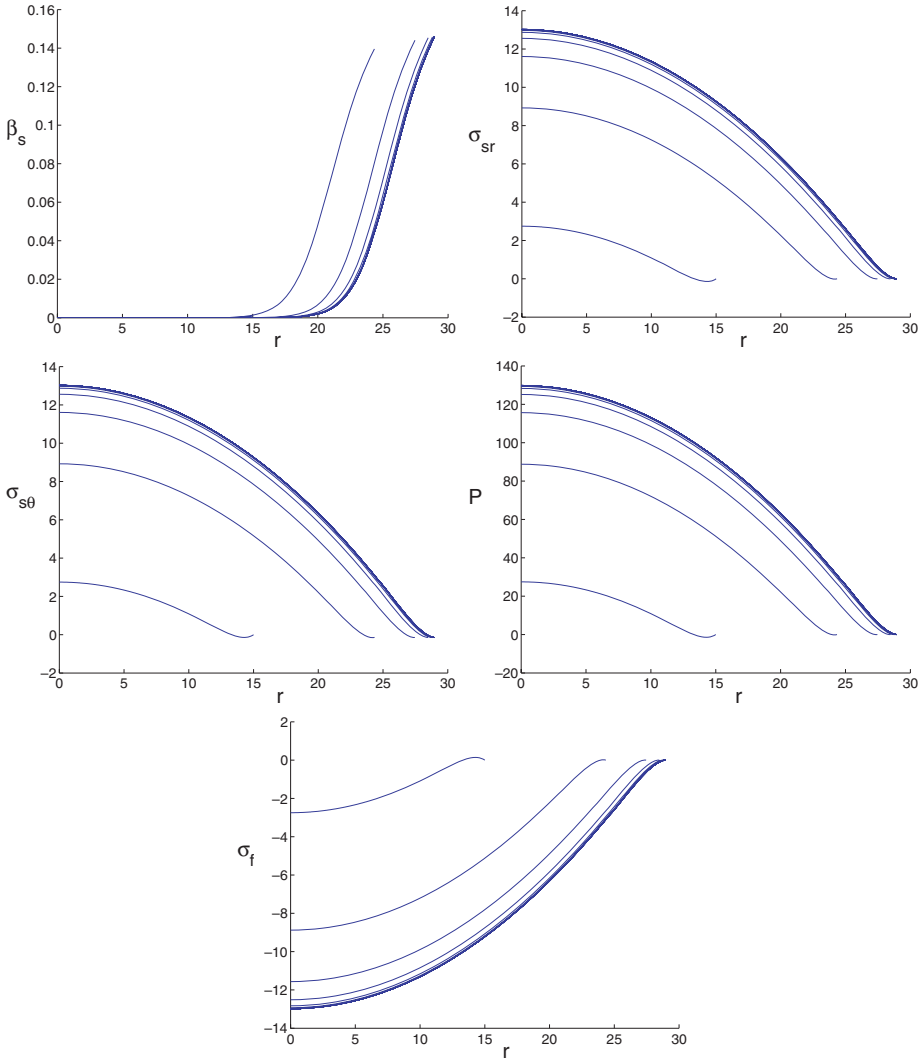


FIG. 10. *Stress distributions for a tumor which begins stress-free and much smaller than its nutrient-regulated equilibrium size,* $a = 15$*, for a biphasic model with* $\chi = 0.1$*. Here,* $t = 0$ *to* $t = 700$ *in time increments of* 35*. The tumor appears to have reached a steady state in both its size and its stress distributions by about* $t = 280$*.* $\phi_f = 0.1$*;* $\epsilon = 0.1$*.*

Caveat. As noted earlier, two distinctly different types of stresses contribute to the total stress in the solid phase—residual stresses, and a hydrostatic mixture stress, P, associated with the pressure in the fluid phase. However, it is essential to bear in mind that, unlike residual stresses, fluid stresses cannot sustain themselves. Therefore, if the growth process were sufficiently slow, the hydrostatic pressure, P, may dissipate. This would have two important implications for mathematical models of stresses in tumors:

1. Multiphase models of residual stresses with only one solid phase (such as the present model) could be replaced by single phase models, since the existence of fluid phases would not influence stresses in the solid phases.

2. Any mathematical model of tumor stresses in which the tissue is composed entirely of fluids would be invalid, since any stresses predicted by the model would dissipate over the timescale of growth. Note that, with the exception of the models by Jones et al. [16], Araujo and McElwain [1, 2, 3, 4, 6, 7], and MacArthur and Please [21], *all* mathematical models of tumor stresses currently in the literature consider the tissue to be composed *entirely of fluids*.

5. Concluding remarks. This paper has studied the evolution of residual stresses in a biphasic model of a multicell spheroid comprising a linear-elastic cellular phase and an inviscid fluid phase. In addition, the solutions to the biphasic equations have been compared with their single phase counterparts, elucidating the approximate nature of the single phase formulation for a growing tissue.

A number of important results have been uncovered by this study. In particular, in view of the minute values of the interphase drag parameter, χ, required to reconcile the single phase and multiphase modelling frameworks (generally $\chi \ll 0.001$), it is likely that single phase models give a poor approximation to the evolving stress profiles in many situations. Most significantly, for a given value of χ, the difference between the two frameworks is more significant for greater values of the stress-relaxation parameter. In the present model, stress relaxation has been imparted via anisotropic growth, with stress relaxation parameter, λ.

While stress relaxation has been addressed qualitatively in both this model and previous models [6, 21], it remains to determine realistic values of stress-relaxation parameters by experimental investigation. It is acknowledged that realistic values of the interphase drag parameter, χ, for real tissues may be somewhat more difficult to determine. Nevertheless, a sustained input by experimental investigators, in complementing these novel theoretical studies, will be the key to continued progress in this important area of tumor growth research.

Furthermore, this study has demonstrated that the consideration of multiple phases alone is unable to alleviate the singularities associated with the isotropic growth of an elastic tissue. Thus, in the context of growth-induced stresses, viscoelasticity and anisotropic growth remain the only two methods of regularizing elasticity.

The model presented here may be extended to consider additional phases such as extracellular matrix and blood vessels, permitting a much wider class of problems to be pursued using these novel theoretical tools. In addition, as discussed in the concluding section of Paper 1, more complicated relationships between interphase mass exchange and solid phase expansion may be proposed, enabling these models to consider the formation of necrotic regions in tumors. Necrosis formation remains a poorly understood aspect of tumor development, and recent investigators [21, 24, 25] have argued that mechanical factors are paramount. Since the formation of necrotic regions appears to correlate with tumor aggressiveness, further insights into this important phenomenon may yield fresh information on tumor invasion and metastasis.

Thus, these new multiphase techniques have the potential to make a powerful contribution to a variety of future projects in cancer research, embryogenesis, and tissue engineering. Indeed, in considering interactions amongst multiple phases, both solid and fluid, the models furnish valuable insights into both the stress profiles within the tissue, as well as the flow of interstitial fluid, permitting a unified study of many diverse aspects of tissue growth. With these techniques, a study of drug delivery may be conducted within an investigation of tumor vascular collapse, for example, where growth-induced stresses compress the weak-walled tumor blood vessels. In addition, the complex relationships between solid stresses and a variety of cellular and molecular

responses in tumors, as illustrated by the experimental work of Helmlinger et al. [15] and Koike et al. [20], may be considered in such theoretical investigations.

Acknowledgment. The authors would like to thank the anonymous referees for their helpful clarifying suggestions that have improved the final version of this paper.

REFERENCES

[1] R. P. Araujo and D. L. S. McElwain, *A mixture theory for the genesis of residual stresses in growing tissues* I: *A general formulation*, SIAM J. Appl. Math., 65 (2005), pp. 1261–1284.
[2] R. P. Araujo and D. L. S. McElwain, *The nature of the stresses induced during tissue growth*, Appl. Math. Lett., 18 (2005), pp. 1081–1085.
[3] R. P. Araujo and D. L. S. McElwain, *An anisotropic model of vascular tumor growth: Implications for vascular collapse*, in Proceedings of the Second M.I.T. Conference on Computational Fluid and Solid Mechanics, K. J. Bathe, ed., Elsevier Ltd., Oxford, UK, 2003, pp. 1613–1616.
[4] R. P. Araujo and D. L. S. McElwain, *The genesis of residual stresses and vascular collapse in solid tumours*, in Proceedings of the Sixth Engineering Mathematics and Applications Conference, R. L. May and W. F. Bluth, eds., Engineering Mathematics Group, ANZIAM, Sydney, Australia, 2003, pp. 1–6.
[5] R. P. Araujo and D. L. S. McElwain, *A history of the study of solid tumour growth: The contribution of mathematical modelling*, Bull. Math. Biol., 66 (2004), pp. 1039–1091.
[6] R. P. Araujo and D. L. S. McElwain, *A linear-elastic model of anisotropic tumour growth*, European J. Appl. Math., 15 (2004), pp. 365–384.
[7] R. P. Araujo and D. L. S. McElwain, *New insights into vascular collapse and growth dynamics in solid tumours*, J. Theoret. Biol., 228 (2004), pp. 335–346.
[8] R. B. Bird, R. C. Armstrong, and O. Hassager, *Dynamics of Polymeric Liquids*, John Wiley & Sons, New York, 1977.
[9] C. J. W. Breward, H. M. Byrne, and C. E. Lewis, *The role of cell-cell interactions in a two-phase model for avascular tumour growth*, J. Math. Biol., 45 (2002), pp. 125–152.
[10] H. M. Byrne, J. R. King, D. L. S. McElwain, and L. Preziosi, *A two-phase model of solid tumour growth*, Appl. Math. Lett., 16 (2003), pp. 567–673.
[11] H. M. Byrne and L. Preziosi, *Modelling solid tumor growth using the theory of mixtures*, Math. Med. Biol., 20 (2003), pp. 341–366.
[12] J. Carlsson, *A proliferation gradient in three-dimensional colonies of cultured human glioma cells*, Int. J. Cancer, 20 (1977), pp. 129–136.
[13] J. Folkman and M. Hochberg, *Self-regulation of growth in three dimensions*, J. Exp. Med., 138 (1973), pp. 745–753.
[14] K. Groebe and W. Mueller-Klieser, *Distributions of oxygen, nutrient and metabolic waste concentrations in multicellular spheroids and their dependence on spheroid parameters*, Eur. Biophys. J., 19 (1991), pp. 169–181.
[15] G. Helmlinger, P. A. Netti, H. D. Lichtenbeld, R. J. Melder, and R. K. Jain, *Solid stress inhibits the growth of multicellular tumour spheroids*, Nature Biotechnology, 15 (1997), pp. 778–783.
[16] A. F. Jones, H. M. Byrne, J. S. Gibson, and J. W. Dold, *A mathematical model of the stress induced during avascular tumour growth*, J. Math. Biol., 40 (2000), pp. 473–499.
[17] D. D. Joseph, *Fluid Dynamics of Viscoelastic Liquids*, Springer-Verlag, New York, 1990.
[18] J. F. R. Kerr, *Shrinkage necrosis: A distinct mode of cellular death*, J. Path., 105 (1971), pp. 13–20.
[19] J. F. R. Kerr, A. H. Wyllie, and A. R. Currie, *Apoptosis: A basic biological phenomenon with wide-ranging implications in tissue kinetics*, Br. J. Cancer, 26 (1972), pp. 239–257.
[20] C. Koike, T. D. McKee, A. Pluen, S. Ramanujan, K. Burton, L. L. Munn, Y. Boucher, and R. K. Jain, *Solid stress facilitates spheroid formation: Potential involvement of hyaluronan*, Br. J. Cancer, 86 (2002), pp. 947–953.
[21] B. D. MacArthur and C. P. Please, *Residual stress generation and necrosis formation in multi-cell tumour spheroids*, J. Math. Biol., 49 (2004), pp. 537–552.
[22] D. L. S. McElwain, R. Callcott, and L. E. Morris, *A model of vascular compression in solid tumours*, J. Theoret. Biol., 78 (1979), pp. 405–415.
[23] D. L. S. McElwain and L. E. Morris, *Apoptosis as a volume loss mechanism in mathematical models of solid tumor growth*, Math. Biosci., 39 (1978), pp. 147–157.

[24] C. P. Please, G. J. Pettet, and D. L. S. McElwain, *A new approach to modelling the formation of necrotic regions in tumours*, Appl. Math. Lett., 11 (1998), pp. 89–94.
[25] C. P. Please, G. J. Pettet, and D. L. S. McElwain, *Avascular tumour dynamics and necrosis*, Math. Models Methods Appl. Sci., 9 (1999), pp. 569–579.
[26] R. Skalak, *Growth as a finite displacement field*, in Proceedings of the IUTAM Symposium on Finite Elasticity, D. E. Carlson and R. T. Shield, eds., Martinus Nijhoff, The Hague, The Netherlands, 1981, pp. 347–355.
[27] R. M. Sutherland and R. E. Durand, *Hypoxic cells in an in vitro tumour model*, Int. J. Radiat. Biol., 23 (1973), pp. 235–246.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 468–488

CLASS-D AUDIO AMPLIFIERS WITH NEGATIVE FEEDBACK*

STEPHEN M. COX† AND BRUCE H. CANDY‡

Abstract. There are many different designs for audio amplifiers. Class-D, or switching, amplifiers generate their output signal in the form of a high-frequency square wave of variable duty cycle (ratio of on time to off time). The square-wave nature of the output allows a particularly efficient output stage, with minimal losses. The output is ultimately filtered to remove components of the spectrum above the audio range. Mathematical models are derived here for a variety of related class-D amplifier designs that use negative feedback. These models use an asymptotic expansion in powers of a small parameter related to the ratio of typical audio frequencies to the switching frequency to develop a power series for the output component in the audio spectrum. These models confirm that there is a form of distortion intrinsic to such amplifier designs. The models also explain why two approaches used commercially succeed in largely eliminating this distortion; a new means of overcoming the intrinsic distortion is revealed by the analysis.

Key words. class-D amplifier, total harmonic distortion, mathematical model

AMS subject classifications. 34E13, 37N20

DOI. 10.1137/040617467

1. Introduction.

Class-D audio amplifiers are becoming increasingly popular, particularly at the high end of the hi-fi audio amplification market. The key feature of their design is that they switch their output between two voltage levels at a very high frequency (typically 500kHz), well above the audio range. The audio signal is essentially encoded in the relative durations of the pulses at the two output voltage levels. The discrete nature of the switching then allows the output stage to be highly efficient; the audio signal is recovered by low-pass filtering of the output. Although the concept of class-D amplifiers using this pulse-width modulation technique has been known for at least fifty years [1], it is only much more recently that electronic components have become available that make its practical implementation feasible. Several commercial amplifiers at the high end of the audio market use class-D amplifier technology.

In its simplest manifestation, the class-D amplifier is known to be capable of producing no distortion to audio signals [1, 4, 5], at least when the mathematical model assumes, as we shall do, that electronic components perform in an ideal fashion, and that the circuit is free from noise. (Significant effort has also been applied to devising remedies for the effects of imperfections in the circuit components [2], for example, nonlinearities in a carrier waveform that is generally modeled mathematically as a piecewise-linear (triangular or sawtooth) wave [6].)

Unfortunately, the simplest design is prone to noise (including thermal and output-stage power-supply noise), due to a lack of negative feedback, and so more sophisticated versions of the class-D design have been developed, incorporating such feedback, in an attempt to counter the poor noise performance. While these negative-feedback

*Received by the editors October 21, 2004; accepted for publication (in revised form) August 16, 2005; published electronically December 2, 2005. This work was supported by Extraordinary Technology Pty Ltd, Australia. This work appeared in preliminary form in the Proceedings of the 117th Audio Engineering Society Convention, San Francisco, 2004.
http://www.siam.org/journals/siap/66-2/61746.html
†School of Mathematical Sciences, University of Adelaide, Adelaide 5005, Australia (stephen.cox@adelaide.edu.au).
‡Halcro, 118 Hayward Avenue, Torrensville 5031, Australia (brucec@adam.com.au).

designs do indeed have better noise performance, they also significantly distort the output, even with perfect components, and there have been various attempts to develop further negative-feedback designs to counter this *intrinsic* distortion.

Despite the great practical value of the application, and the variety of "engineering" solutions available, there appears to be a dearth of mathematical models for class-D amplifier designs with negative feedback. By contrast, the no-feedback case was analyzed over fifty years ago by Black in his treatise [1] and was shown to allow distortion-free output of sinusoidal input signals. More recently the same problem was reconsidered in greater depth [4, 5], and it was shown that there is no distortion to *any* audio signal, sinusoidal or otherwise. The latter result is significant because the amplifier design is nonlinear, and thus the distortion characteristics of an arbitrary signal cannot be inferred from those of its Fourier components.

We develop mathematical models for class-D amplifiers with negative feedback. The models proceed from the governing differential equations that relate the voltage signals at the various parts of the device, assuming perfect components. The resulting system of equations may be formally integrated to yield what is essentially a set of nonlinear difference equations for the various internal signals at multiples of the switching period. The solution to these equations is then developed in an asymptotic series based on the separation of scales between the (relatively high-frequency) switching stage and the (relatively low-frequency) audio signal. The analysis is continued as far as the first term in the series that reveals the inherent distortion of the system. We then show how two successful commercial approaches to significantly reducing this component of the distortion can be modeled, and confirm what is already known empirically, that they do indeed work. The analysis reveals a third means of reducing the intrinsic distortion. We conclude by briefly considering the effects of nonlinear distortion to the carrier wave upon the audio output.

2. Mathematical model: General considerations. The "classical" class-D amplifier design, without negative feedback, is illustrated in Figure 2.1. The audio input signal is denoted by $s(t)$; generally this signal comprises a Fourier spectrum in the audible range up to 20kHz. This audio signal is added to a triangular carrier wave $v(t)$, with period T, that satisfies

$$v(t) = \begin{cases} 1 - \dfrac{4t}{T} & \text{for } 0 \le t < \dfrac{T}{2}, \\ -3 + \dfrac{4t}{T} & \text{for } \dfrac{T}{2} \le t < T, \end{cases} \tag{2.1}$$

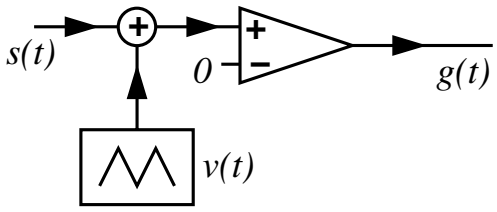


FIG. 2.1. *Classical class-D amplifier (without negative feedback). The audio input signal is* $s(t)$*; this is summed with a high-frequency triangular carrier wave* $v(t)$ *and input to the noninverting input (+) of a comparator, whose inverting input (−) is grounded. The output of the comparator is* $g(t)$*, given by* (2.2).

and $v(t+T) = v(t)$ for all t. Thus $v(nT) = 1$ and $v((n+\frac{1}{2})T) = -1$, for any integer n, and $v(t)$ is piecewise linear between these two values. It will be significant for the analysis that follows that if ω is a typical audio frequency, then $\omega T \ll 1$. The main circuit element is a comparator, which compares the voltage at its noninverting input (denoted by a "+" in the figure) with the voltage at its inverting input (denoted by a "−") and gives an output $g(t)$ that satisfies

$$g(t) = \begin{cases} +1 & \text{if } s(t) + v(t) > 0, \\ -1 & \text{if } s(t) + v(t) < 0. \end{cases} \tag{2.2}$$

Note that the output voltages have been normalized to ± 1; furthermore, we assume throughout this paper that $-1 < s(t) < 1$ for all t. The switching times of $g(t)$ are thus governed by $s(t) + v(t) = 0$; we denote the switching times from $+1$ to -1 by $t = nT + \alpha_n$, with the reverse switchings at times $t = nT + \beta_n$. For the classical design in Figure 2.1 these switching times are governed by

$$0 < \alpha_n = \frac{T}{4}\left(1 + s(nT + \alpha_n)\right) < \frac{T}{2} < \beta_n = \frac{T}{4}\left(3 - s(nT + \beta_n)\right) < T. \tag{2.3}$$

Note that the equations in (2.3) give α_n and β_n only implicitly. We shall consider their solution later.

We now examine how the switching times are used in computing the component of $g(t)$ in the audio spectrum, i.e., the amplifier output.

2.1. Comparator output $g(t)$. In all class-D designs, regardless of the details, the output $g(t)$ takes the form

$$g(t) = \begin{cases} +1 & \text{if } nT < t < nT + \alpha_n \quad \text{or} \quad nT + \beta_n < t < (n+1)T, \\ -1 & \text{if } nT + \alpha_n < t < nT + \beta_n, \end{cases} \tag{2.4}$$

for some switching times $t = nT + \alpha_n$ and $nT + \beta_n$. Thus we may write

$$g(t) = 1 - 2\sum_{n=-\infty}^{\infty} \mathcal{H}_n(t), \tag{2.5}$$

where

$$\mathcal{H}_n(t) = H(t - (nT + \alpha_n)) - H(t - (nT + \beta_n)) \tag{2.6}$$

and where $H(t)$ is the Heaviside step function ($H(t) = 0$ for $t < 0$ and $H(t) = 1$ for $t > 0$). Note that each $\mathcal{H}_n(t)$ has finite support:

$$\mathcal{H}_n(t) = \begin{cases} 1, & nT + \alpha_n < t < nT + \beta_n, \\ 0 & \text{otherwise}, \end{cases} \tag{2.7}$$

so the sum in (2.5) is well defined almost everywhere (i.e., except at the exact switching times). The corresponding Fourier transform

$$\hat{g}(\omega) \equiv \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} g(t) \mathrm{e}^{-\mathrm{i}\omega t}\, \mathrm{d}t \tag{2.8}$$

is then

$$\hat{g}(\omega) = (2\pi)^{1/2}\delta(\omega) - \frac{2\mathrm{i}}{(2\pi)^{1/2}\omega} \sum_{n=-\infty}^{\infty} \left\{ \mathrm{e}^{-\mathrm{i}\omega(nT+\beta_n)} - \mathrm{e}^{-\mathrm{i}\omega(nT+\alpha_n)} \right\}. \tag{2.9}$$

A formal analysis of this expression allows us to write the output $g(t)$ more usefully in terms of the switching times. We begin by denoting by $\mathcal{S}$ the sum in (2.9). Then

$$\begin{aligned}\mathcal{S} &= \sum_{n=-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega nT}\left(\mathrm{e}^{-\mathrm{i}\omega\beta_n} - \mathrm{e}^{-\mathrm{i}\omega\alpha_n}\right)\\ &= \sum_{n=-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega nT} \sum_{m=1}^{\infty} \frac{1}{m!}(-\mathrm{i}\omega)^m(\beta_n^m - \alpha_n^m)\\ &= \sum_{m=1}^{\infty} \frac{1}{m!}(-\mathrm{i}\omega)^m \sum_{n=-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega nT}(\beta_n^m - \alpha_n^m). \qquad (2.10)\end{aligned}$$

Now we note that

$$\mathrm{e}^{-\mathrm{i}\omega nT}(\beta_n^m - \alpha_n^m) = \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega t}\left[B^m(t) - A^m(t)\right]\delta(t - nT)\,\mathrm{d}t, \qquad (2.11)$$

where $A(t)$ and $B(t)$ are any smooth functions that satisfy

$$A(nT) = \alpha_n, \qquad B(nT) = \beta_n. \qquad (2.12)$$

We shall refer to $A(t)$ and $B(t)$ as generalized switching-time functions. Substituting (2.11) into (2.10) and using the discrete Fourier transform identity

$$\sum_{n=-\infty}^{\infty} \delta(t - nT) = \frac{1}{T}\sum_{n=-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\omega_c nt}, \qquad (2.13)$$

where $\omega_c = 2\pi/T$ is the carrier-wave frequency, we find

$$\begin{aligned}\mathcal{S} &= \frac{1}{T}\sum_{m=1}^{\infty} \frac{1}{m!}(-\mathrm{i}\omega)^m \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}(\omega - n\omega_c)t}\left[B^m(t) - A^m(t)\right]\,\mathrm{d}t\\ &= \frac{(2\pi)^{1/2}}{T}\sum_{m=1}^{\infty} \frac{1}{m!}(-\mathrm{i}\omega)^m \sum_{n=-\infty}^{\infty} \left[\hat{B}_m(\omega - n\omega_c) - \hat{A}_m(\omega - n\omega_c)\right]. \qquad (2.14)\end{aligned}$$

Here $\hat{A}_m$ and $\hat{B}_m$ are, respectively, the Fourier transforms of A^m and B^m.

Now we use the separation of time scales between the audio signal and the carrier wave to generate a simpler approximation to this expression for $\mathcal{S}$. We begin by noting that since the signal $s(t)$ varies only slowly over a single period of the carrier wave, the switching times α_n and β_n vary only slowly with n. Thus we assume that $A(t)$ and $B(t)$ are chosen to vary only on the relatively slow time scale of the signal and contain no components that vary on the shorter time scale of the carrier wave. (Of course, in practice, this "bandwidth limiting" is only approximate.) The upshot of this assumption is that we take into account only the term $n = 0$ in the sum in (2.14); then, with this approximation,

$$\mathcal{S} = \frac{(2\pi)^{1/2}}{T}\sum_{m=1}^{\infty} \frac{1}{m!}(-1)^m(\mathrm{i}\omega)^m\left[\hat{B}_m(\omega) - \hat{A}_m(\omega)\right]. \qquad (2.15)$$

Correspondingly, from (2.9) it follows that

$$\hat{g}(\omega) = (2\pi)^{1/2}\delta(\omega) + \frac{2}{T}\sum_{m=1}^{\infty} \frac{1}{m!}(-1)^m(\mathrm{i}\omega)^{m-1}\left[\hat{B}_m(\omega) - \hat{A}_m(\omega)\right], \qquad (2.16)$$

and hence, upon inverting the Fourier transform, we obtain for the component of $g(t)$ in the audio spectrum (cf. [4, 5])

$$g_a(t) = 1 + \frac{2}{T}\sum_{m=1}^{\infty} \frac{1}{m!}(-1)^m \frac{\mathrm{d}^{m-1}}{\mathrm{d}t^{m-1}}\left[B^m(t) - A^m(t)\right]. \tag{2.17}$$

This expression applies regardless of the details of the class-D amplifier design: the differences between the various designs lie in the specific relationships between the generalized switching-time functions and the signal $s(t)$; these correspondingly result in different audio outputs $g_a(t)$.

2.1.1. Discussion. The infinite sum in (2.17) should, of course, be viewed with some caution, given the approximations underlying it. Even if $A(t)$ and $B(t)$ are signals whose frequency spectra lie entirely within the audio range, the powers A^m and B^m include successively higher frequencies in their spectra. For example, if A and B are pure sinusoidal signals, each with frequency ω, then A^m and B^m involve frequencies up to $m\omega$. For sufficiently large m, when $|\omega_c \mp m\omega|$ lies in the audio range, terms with $n = \pm 1$ must be included in the sum (2.14), rendering inappropriate our assumption that only terms with $n = 0$ contribute to the output audio spectrum (for larger values of m, additional values of n also become relevant). However, in the analysis that follows we shall consider only the first few terms in (2.17), because these are sufficient to determine the principal distortion characteristics of the amplifier designs; hence for our purposes the difficulty with large values of m in (2.17) is immaterial.

2.2. Alternative expression for $g_a(t)$. This section may be omitted by readers interested only in the class-D amplifier designs with negative feedback, since it is primarily of importance for the classical design without feedback, in which case we shall see below that the equations governing $A(t)$ and $B(t)$ take the form

$$A(t) = \bar{A}(t + A(t)), \qquad B(t) = \bar{B}(t + B(t)), \tag{2.18}$$

for some functions $\bar{A}$ and $\bar{B}$. Clearly some conditions must be imposed upon $A(t)$ and $B(t)$ in order that (2.18) define $\bar{A}$ and $\bar{B}$ uniquely; it is sufficient that $A(t)$ and $B(t)$ should vary sufficiently slowly, i.e., $|A'(t)| < 1$ and $|B'(t)| < 1$ (cf. [5]). It then proves useful to introduce "warped times"

$$t_A = t + A(t), \qquad t_B = t + B(t), \tag{2.19}$$

so that

$$t_A = t + \bar{A}(t_A), \qquad t_B = t + \bar{B}(t_B). \tag{2.20}$$

Now to obtain a simpler expression for $g_a(t)$, we note from the definition of the Fourier transform (2.8) and (2.15) that $\mathcal{S}$ may be written as

$$\mathcal{S} = \frac{1}{T}\sum_{m=1}^{\infty} \frac{(-\mathrm{i}\omega)^m}{m!} \int_{-\infty}^{\infty} [B^m(t) - A^m(t)]\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}t, \tag{2.21}$$

and hence, provided the order of the summation and integration may be interchanged,

$$\begin{aligned} \mathcal{S} &= \frac{1}{T}\int_{-\infty}^{\infty} \sum_{m=1}^{\infty} \frac{(-\mathrm{i}\omega)^m}{m!}[B^m(t) - A^m(t)]\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}t \\ &= \frac{1}{T}\int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega(t+B(t))} - \mathrm{e}^{-\mathrm{i}\omega(t+A(t))}\,\mathrm{d}t, \end{aligned} \tag{2.22}$$

from which it follows that

$$\begin{aligned} \mathrm{i}\omega\hat{g}(\omega) &= \frac{2}{T}\frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega(t+B(t))} - \mathrm{e}^{-\mathrm{i}\omega(t+A(t))}\,\mathrm{d}t \\ &= \frac{2}{T}\frac{1}{(2\pi)^{1/2}}\left\{\int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega t_B}\frac{\mathrm{d}t}{\mathrm{d}t_B}\,\mathrm{d}t_B - \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega t_A}\frac{\mathrm{d}t}{\mathrm{d}t_A}\,\mathrm{d}t_A\right\} \\ &= \frac{2}{T}\frac{\mathrm{i}\omega}{(2\pi)^{1/2}}\left\{\int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega t_B}t(t_B)\,\mathrm{d}t_B - \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega t_A}t(t_A)\,\mathrm{d}t_A\right\}. \end{aligned} \tag{2.23}$$

Thus, by inverting the Fourier transform, we find that

$$g_a(t) = C_a + \frac{2}{T}\left[g_A(t) + g_B(t)\right], \tag{2.24}$$

where C_a is a constant of integration, $g_A(t_A) = -t(t_A)$, and $g_B(t_B) = t(t_B)$. From (2.20), $t(t_A) = t_A - \bar{A}(t_A)$ and $t(t_B) = t_B - \bar{B}(t_B)$. Thus

$$g_A(t_A) = -t_A + \bar{A}(t_A), \qquad g_B(t_B) = t_B - \bar{B}(t_B), \tag{2.25}$$

or, equivalently,

$$g_A(t) = -t + \bar{A}(t), \qquad g_B(t) = t - \bar{B}(t), \tag{2.26}$$

so that, finally, we obtain from (2.24) and (2.26)

$$g_a(t) = 1 + \frac{2}{T}\left[\bar{A}(t) - \bar{B}(t)\right]. \tag{2.27}$$

The constant of integration $C_a = 1$ has been fixed by noting that for zero input signal ($s(t) \equiv 0$) it follows from (2.2) that $g_a(t) \equiv 0$, while $\bar{A}(t) \equiv T/4$ and $\bar{B}(t) \equiv 3T/4$. Thus when the problem for the generalized switching times is of the form (2.18), the audio output takes a particularly simple form, which does not appear to have been noted previously.

2.3. Classical class-D amplifier. For the classical class-D amplifier illustrated in Figure 2.1, we find from (2.3) and (2.12) that

$$A(nT) = \frac{T}{4}\left[1 + s(nT + A(nT))\right], \qquad B(nT) = \frac{T}{4}\left[3 - s(nT + B(nT))\right]. \tag{2.28}$$

It is then straightforward to extend this definition of $A(t)$ and $B(t)$ appropriately to other times by globally mapping $nT \mapsto t$ in (2.28). Then, in view of (2.18) and (2.27), it follows that

$$g_a(t) = 1 + \frac{2}{T}\left\{\frac{T}{4}\left[1 + s(t)\right] - \frac{T}{4}\left[3 - s(t)\right]\right\} = s(t). \tag{2.29}$$

Thus (as is well known [1, 4, 5]) this amplifier gives no distortion to the signal (given the modeling assumptions).

Note in particular that no assumption has been made regarding the shape of the audio waveform $s(t)$ other than that its spectrum lies entirely in the audio band, well below the carrier-wave frequency; most significantly, the input signal need not be a pure sinusoid. Our derivation of this result differs somewhat from that given recently [4, 5], since we use (2.27) rather than applying a theorem in complex analysis due to Lagrange [4, 5].

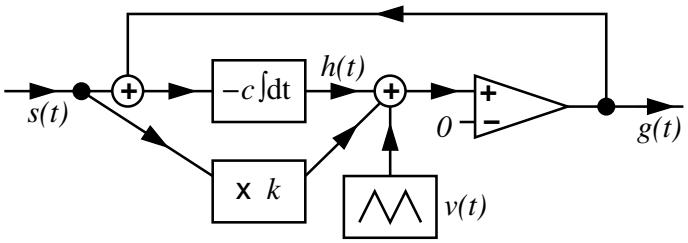


FIG. 3.1. *Class-D amplifier with negative feedback. The signal $s(t)$ is fed into a device that multiplies it by a constant k, and also into an integrator, whose output we denote by $h(t)$. The outputs of the integrator and the multiplier are summed, together with a high-frequency triangular carrier wave $v(t)$ and input to the noninverting input of a comparator whose inverting input is grounded. The output of the comparator is $g(t)$.*

3. Class-D amplifier with negative feedback. With negative feedback, the basic class-D amplifier design is as illustrated in Figure 3.1.

In analyzing the amplifier design, we find it convenient to introduce $f(t)$, the integral of the input signal, defined by $f'(t) = s(t)$; the constant of integration in determining $f(t)$ uniquely is not important in what follows. The triangular carrier wave $v(t)$ again satisfies (2.1) and the periodicity condition $v(t+T) = v(t)$ for all t. The output $g(t)$ of the comparator now satisfies

$$g(t) = \begin{cases} +1 & \text{if } h(t) + ks(t) + v(t) > 0, \\ -1 & \text{if } h(t) + ks(t) + v(t) < 0. \end{cases} \tag{3.1}$$

Finally, the integrator output is given by

$$h(t) = -c \int^t g(\tau) + s(\tau)\,\mathrm{d}\tau. \tag{3.2}$$

The time constant c is such that $cT = O(1)$. Since $-1 < s(t) < 1$, $h(t)$ alternately increases and decreases, when $g(t)$ is, respectively, negative and positive. The relationships between $v(t)$, $g(t)$, and $h(t)$ are illustrated in Figure 3.2. We note that for illustrative purposes the figure shows $h(t)$ as a piecewise linear function of time, which is appropriate only for a constant input signal; otherwise $h(t)$ has a slight nonlinearity.

3.1. Analysis of the model. We analyze the model by first constructing a system of nonlinear implicit difference equations for the switching times α_n and β_n. To do so, we consider a time interval $nT < t < (n+1)T$. Referring to the waveform in Figure 3.2, we see that at the start and end of this interval, $h(t)$ is decreasing; in-between, $h(t)$ is increasing. We define three subintervals:

$$\begin{array}{rll} \text{I:} & nT < t < nT + \alpha_n, & h'(t) < 0 \ [g(t) = 1], \\ \text{II:} & nT + \alpha_n < t < nT + \beta_n, & h'(t) > 0 \ [g(t) = -1], \\ \text{III:} & nT + \beta_n < t < (n+1)T, & h'(t) < 0 \ [g(t) = 1], \end{array} \tag{3.3}$$

and consider each in turn.

Subinterval I. By integrating (3.2) we find

$$h(t) = h(nT) - c[f(t) - f(nT)] - c(t - nT). \tag{3.4}$$

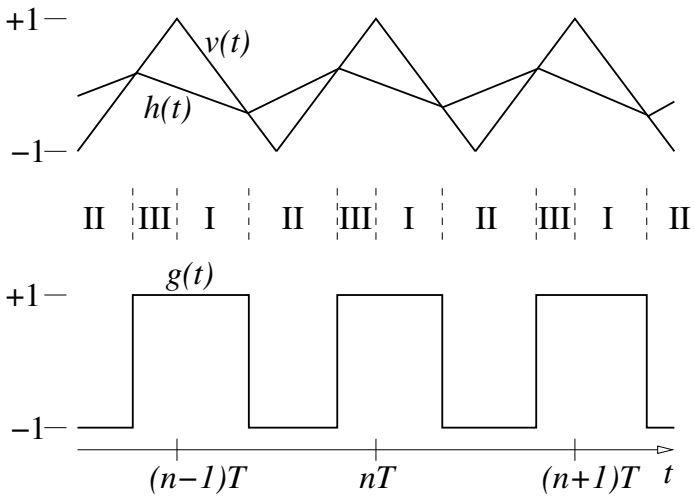


FIG. 3.2. *Diagram showing relationships between $v(t)$, $g(t)$, and $h(t)$ for the class-D amplifier with negative feedback. Note that, although $h(t)$ is drawn here as piecewise linear, it is in fact nonlinear for any nontrivial input signal $s(t)$. The subintervals* I, II, *and* III *are indicated, as defined in* (3.3).

According to (3.1), the value of α_n is defined by

$$h(nT+\alpha_n) + ks(nT+\alpha_n) + v(nT+\alpha_n) = 0; \tag{3.5}$$

that is,

$$h(nT) - c[f(nT+\alpha_n) - f(nT)] - c\alpha_n + ks(nT+\alpha_n) + 1 - \frac{4\alpha_n}{T} = 0. \tag{3.6}$$

Subinterval II. By integrating (3.2) and enforcing continuity of $h(t)$ at time $t = nT + \alpha_n$, we find

$$h(t) = h(nT) - c[f(t) - f(nT)] - c\alpha_n + c(t - nT - \alpha_n). \tag{3.7}$$

From (3.1), the value of β_n is defined by

$$h(nT+\beta_n) + ks(nT+\beta_n) + v(nT+\beta_n) = 0; \tag{3.8}$$

that is,

$$h(nT) - c[f(nT+\beta_n) - f(nT)] + c(\beta_n - 2\alpha_n) + ks(nT+\beta_n) - 3 + \frac{4\beta_n}{T} = 0. \tag{3.9}$$

Subinterval III. By integrating (3.2) and enforcing continuity of $h(t)$ at time $t = nT + \beta_n$, we find

$$h(t) = h(nT) - c[f(t) - f(nT)] + c(\beta_n - 2\alpha_n) - c(t - nT - \beta_n). \tag{3.10}$$

For the remaining analysis, we note that at the end of this subinterval

$$h((n+1)T) = h(nT) - c[f((n+1)T) - f(nT)] + c(2\beta_n - 2\alpha_n - T). \tag{3.11}$$

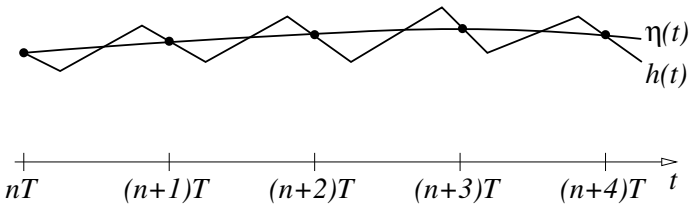


FIG. 3.3. *Diagram showing the relationship between $h(t)$ and $\eta(t)$. The two functions agree at times nT (for integers n), but $h(t)$ varies significantly at intermediate times, whereas η is only slowly varying.*

3.2. Solution of the governing equations. Our goal is to use (2.17) to determine the audio output of the amplifier. To do so we first need to determine the switching times of $g(t)$. Given a signal $s(t)$, we determine these switching times, together with the values of $h(nT)$, from the coupled equations (3.6), (3.9), and (3.11), which are, so far, exact.

We first use (2.12) to substitute the generalized switching-time functions for α_n and β_n; the result is a system of three equations involving $A(nT)$, $B(nT)$, and $h(nT)$. These equations are readily extended to intermediate times by mapping $nT \mapsto t$:

$$\left(\frac{4}{T}+c\right)A(t) = 1 + ks(t+A(t)) + \eta(t) - c[f(t+A(t)) - f(t)], \tag{3.12}$$

$$\left(\frac{4}{T}+c\right)B(t) = 3 - ks(t+B(t)) - \eta(t) + c[f(t+B(t)) - f(t)] + 2cA(t), \tag{3.13}$$

$$\eta(t+T) = \eta(t) - c[f(t+T) - f(t)] + c[2B(t) - 2A(t) - T]. \tag{3.14}$$

Since the functions $A(t)$ and $B(t)$ vary only on the time scale of the audio signal and not on that of the carrier wave, the function $h(t)$ is replaced in these expressions by a slowly varying function $\eta(t)$ such that

$$\eta(nT) = h(nT) \tag{3.15}$$

(see Figure 3.3).

Given the mild restrictions on the form of the input signal $s(t)$ and its first integral $f(t)$, it seems unlikely that a general solution can be found to the coupled nonlinear equations (3.12)–(3.14). Furthermore, it seems unlikely that any solution will be unique, although we are unable to demonstrate nonuniqueness for the system of three equations as posed. (A suggestion of nonuniqueness comes from the following thought experiment. Suppose $A(t)$ and $B(t)$ are known and independent of $\eta(t)$. Then the solution $\eta(t)$ to (3.14) is unique only up to the addition of a function of period T [3]. Note that here the nonuniqueness involves high-frequency components only.)

We find that we are able to construct a solution to (3.12)–(3.14) with frequency spectrum in the audio range, as follows. The derivation is admittedly rather informal. We introduce the small parameter $\epsilon = \omega_{\text{typ}}/\omega_c \ll 1$, where ω_{typ} is a typical audio frequency component of the input. Then we note that (3.12)–(3.14) relate to variations in s, A, B, and η on a time scale $t = O(T)$, and that such variations satisfy

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n} = O(\epsilon^n).$$

We then expand $A(t)$, $B(t)$, and $\eta(t)$ as series

$$A(t) = \sum_{m=0}^{\infty} A_m(t), \qquad B(t) = \sum_{m=0}^{\infty} B_m(t), \qquad \eta(t) = \sum_{m=0}^{\infty} \eta_m(t), \tag{3.16}$$

where the terms in (3.16) satisfy

$$A_m, B_m, \eta_m = O(\epsilon^m).$$

For all functions in (3.12)–(3.14) not evaluated at time t, we make use of Taylor expansions such as

$$\eta(t+T) = \sum_{n=0}^{\infty} T^n \frac{\eta^{(n)}(t)}{n!} \tag{3.17}$$

to write them in terms of functions and derivatives evaluated at time t.

As an aside, we note that an alternative mathematical solution may possibly be developed using the calculus of finite differences [3], where problems such as (3.14), of the form $u(t+T) - u(t) = U(t)$, are examples of the so-called "summation problem." However, the added complication of (3.12) and (3.13) makes a complete explicit solution unlikely. We note, however, that (3.14) has the exact formal solution

$$\eta(t) = -cf(t) + c\sum_{n=1}^{\infty} \left[2B(t-nT) - 2A(t-nT) - T\right].$$

Some of the key terms in the various expansions are

$$\begin{aligned}
f(t+T) - f(t) &= Ts(t) + \tfrac{1}{2}T^2 s'(t) + \tfrac{1}{6}T^3 s''(t) + O(\epsilon^3),\\
\eta(t+T) - \eta(t) &= T\eta_0'(t) + \left[T\eta_1'(t) + \tfrac{1}{2}T^2\eta_0''(t)\right] + O(\epsilon^3),\\
s(t+A(t)) &= s(t) + A_0(t)s'(t) + \left[\tfrac{1}{2}A_0^2(t)s''(t) + A_1(t)s'(t)\right] + O(\epsilon^3),\\
s(t+B(t)) &= s(t) + B_0(t)s'(t) + \left[\tfrac{1}{2}B_0^2(t)s''(t) + B_1(t)s'(t)\right] + O(\epsilon^3),\\
f(t+A(t)) - f(t) &= A_0(t)s(t) + \left[\tfrac{1}{2}A_0^2(t)s'(t) + A_1(t)s(t)\right]\\
&\quad + \left[\tfrac{1}{6}A_0^3(t)s''(t) + A_0(t)A_1(t)s'(t) + A_2(t)s(t)\right] + O(\epsilon^3),\\
f(t+B(t)) - f(t) &= B_0(t)s(t) + \left[\tfrac{1}{2}B_0^2(t)s'(t) + B_1(t)s(t)\right]\\
&\quad + \left[\tfrac{1}{6}B_0^3(t)s''(t) + B_0(t)B_1(t)s'(t) + B_2(t)s(t)\right] + O(\epsilon^3).
\end{aligned}$$

In view of (2.17), we also have the expansion

$$\begin{aligned}
g_a(t) = 1 &- \frac{2}{T}(B_0 - A_0)\\
&+ \frac{1}{T}\left[(B_0^2 - A_0^2)' - 2(B_1 - A_1)\right]\\
&+ \frac{1}{T}\left[-\tfrac{1}{3}(B_0^3 - A_0^3)'' + 2(B_0B_1 - A_0A_1)' - 2(B_2 - A_2)\right] + O(\epsilon^3)
\end{aligned} \tag{3.18}$$

for the audio output. With obvious notation, we write this as

$$g_a(t) = g_0(t) + g_1(t) + g_2(t) + O(\epsilon^3). \tag{3.19}$$

Let us now consider the problem (3.12)–(3.14) at successive powers of ϵ, starting with terms of $O(\epsilon^0)$. In doing so, it proves useful to note that upon adding (3.12) to (3.13) we eliminate the unknown function $\eta(t)$ and arrive at an equation that at $O(\epsilon^n)$ takes the form

$$\{4 - cT\,[1 - s(t)]\}\,A_n(t) + \{4 + cT\,[1 - s(t)]\}\,B_n(t) = P_n(t), \tag{3.20}$$

where $P_n(t)$ is known in terms of quantities calculated at previous stages in the calculation. Furthermore, (3.14) can be written in the form

$$B_n(t) - A_n(t) = Q_n(t), \tag{3.21}$$

where again $Q_n(t)$ comprises known quantities. This system is readily solved to give

$$A_n(t) = \tfrac{1}{8}\left\{P_n(t) - [4 + cT(1 - s(t))]\,Q_n(t)\right\}, \tag{3.22}$$

$$B_n(t) = \tfrac{1}{8}\left\{P_n(t) + [4 - cT(1 - s(t))]\,Q_n(t)\right\}. \tag{3.23}$$

At $O(\epsilon^0)$, we find $P_0 = 4T$ and $Q_0 = \frac{1}{2}T(1 + s(t))$. Correspondingly,

$$A_0 = \tfrac{1}{16}T(1 - s(t))\,[4 - cT(1 + s(t))]\,, \tag{3.24}$$

$$B_0 = \tfrac{1}{2}T + \tfrac{1}{16}T(1 + s(t))\,[4 - cT(1 - s(t))]\,, \tag{3.25}$$

and thus the switching times approximately satisfy

$$\alpha_n = \tfrac{1}{16}T(1 - s(nT))\,[4 - cT(1 + s(nT))]\,, \tag{3.26}$$

$$\beta_n = \tfrac{1}{2}T + \tfrac{1}{16}T(1 + s(nT))\,[4 - cT(1 - s(nT))]\,. \tag{3.27}$$

Of most significance is the result, which now follows from (3.24), (3.25), and (3.18), that

$$g_0(t) = -s(t). \tag{3.28}$$

Thus to this order there is no distortion from signal to output, apart from a sign change, which is unimportant for audio applications. However, in contrast to the classical design, the next orders in the expansion of the audio output reveal distortion inherent in the nonlinear-feedback design. (The minus sign in (3.28) is an artifact of applying the triangular wave input to the noninverting input of the comparator; if it is instead applied to the inverting input, then there is no sign change to the output.)

3.3. Amplifier output. The next steps in the calculation are algebraically cumbersome and shed little further light on the problem, so the full details are not presented here. Our primary interest lies in the audio output, and this turns out to be

$$\begin{aligned} g_a(t) = {} & -s(t) + \frac{1+k}{c}s'(t) \\ & - \frac{1}{48c^2}\left\{[48(1+k) - c^2T^2]s(t) - c^2T^2s^3(t)\right\}'' + O(\epsilon^3). \end{aligned} \tag{3.29}$$

Note that there arises a nonlinear (cubic) distortion term; this term is to leading order independent of k, so it cannot be removed by any choice of this parameter. This nonlinear term represents the "intrinsic" distortion of class-D amplifiers with negative feedback to which we have alluded above.

The linear terms in (3.29) also represent a form of distortion to the signal since they affect different frequency components to different extents. This distortion can be removed by making an appropriate choice of k such that the linear terms in (3.29) form the beginnings of a Taylor series for a slightly delayed signal $-s(t-(1+k)/c)$. It is readily determined that the appropriate value of k satisfies

$$k^2 = 1 - \frac{1}{24}c^2T^2; \tag{3.30}$$

correspondingly, the audio output is then given by

$$g_a(t) = -s\left(t - \frac{(1+k)}{c}\right) + \frac{1}{48}T^2\left(s^3(t)\right)'' + O(\epsilon^3). \tag{3.31}$$

Note that the delay to the output indicated here is independent of signal amplitude and frequency, and thus is entirely benign. Furthermore, $(1+k)/c$ is a time of the order of the carrier-wave period, so the delay to the audio signal is slight. However, the nonlinear distortion term remains.

For the specific case of a sinusoidal input signal $s(t) = s_0 \sin \omega t$, (3.29) becomes

$$\begin{aligned} g_a(t) &= -s_0 \sin \omega t + (1+k)\mu s_0 \cos \omega t \\ &\quad + \frac{\mu^2}{192}\left\{\left[192(1+k) - (4+3s_0^2)c^2T^2\right] s_0 \sin \omega t + 9c^2T^2 s_0^3 \sin 3\omega t\right\} + O(\mu^3), \end{aligned} \tag{3.32}$$

where $\mu = O(\epsilon)$: specifically,

$$\mu = \frac{\omega T}{cT} \ll 1. \tag{3.33}$$

We note from (3.32) that the intrinsic nonlinear distortion manifests itself through both a nonlinear influence on the amplitude of the fundamental and the presence of a third-harmonic term.

3.4. Alternative expression for $g_a(t)$. An alternative expression for the audio output may be derived as follows. First we note that the switching times for $g(t)$ satisfy

$$\begin{aligned} \alpha_n &= \frac{T}{4}\left[1 + h(nT+\alpha_n) + ks(nT+\alpha_n)\right], \\ \beta_n &= \frac{T}{4}\left[3 - h(nT+\beta_n) - ks(nT+\beta_n)\right]. \end{aligned}$$

If we introduce two new slowly varying functions $\eta(t;\alpha)$ and $\eta(t;\beta)$ defined so that

$$\eta(nT+\alpha_n;\alpha) = h(nT+\alpha_n), \qquad \eta(nT+\beta_n;\beta) = h(nT+\beta_n), \tag{3.34}$$

then it follows from (2.29) that

$$g_a(t) = \frac{1}{2}\left[\eta(t;\alpha) + \eta(t;\beta)\right] + ks(t). \tag{3.35}$$

Although this expression does not yield a useful explicit exact formula for $g_a(t)$, it does provide an alternative means of calculating $g_a(t)$. This in turn gives us an independent check on our results, which we have used to verify expressions such as (3.29).

Having highlighted the third-harmonic distortion generated by the simplest class-D amplifier design with negative feedback, we proceed to describe some remedies.

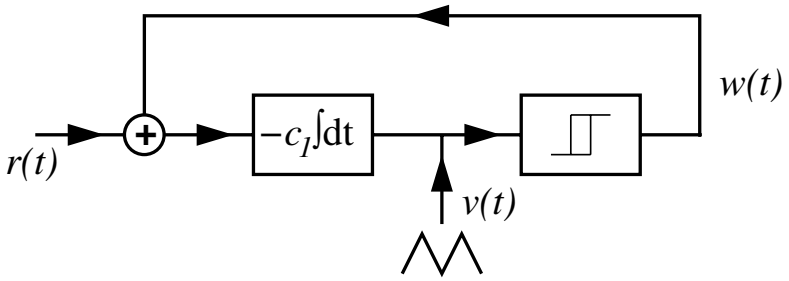


FIG. 4.1. *Circuit diagram for modulation of the carrier wave. The modulation function $r(t)$ is input to an integrator, whose output $v(t)$ is input to a hysteresis loop. The output $w(t)$ of the last device takes the value $+1$ once its input has reached the value $+1$; thereafter $w(t)$ remains at $+1$ until the input $v(t)$ falls to -1, from which point onwards $w(t) = -1$ until the input $v(t)$ reaches $+1$ again, and so on.*

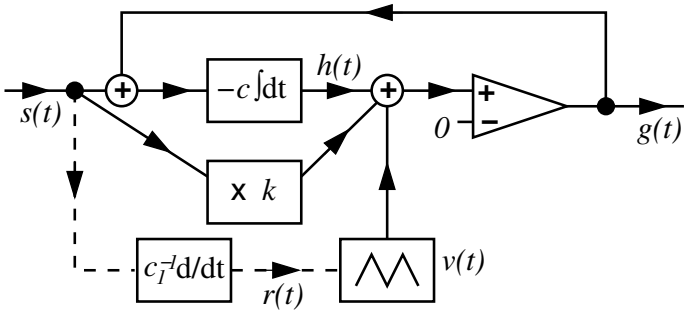


FIG. 4.2. *Class-D amplifier with negative feedback and modulation of the carrier-wave symmetry. Note that the entire circuitry of Figure* 4.1 *is represented by the single box marked "$v(t)$." The appropriate modulation signal $r(t) = c_1^{-1}s'(t)$ is indicated on the diagram, as determined in section* 4.1.2.

4. Modulation of the carrier wave. We now examine one means of eliminating the "intrinsic distortion" term in (3.29). The key to the technique is to modulate the carrier wave in such a way that the switching times of $g(t)$ are slightly altered in a fashion appropriate to countering the distortion. This technique is used successfully for distortion reduction in amplifiers manufactured by Halcro (www.halcro.com).

We suppose that the carrier wave is modulated by a slowly varying input signal $r(t)$, where $r(t) = e'(t)$, for some function $e(t)$; since $r(t)$ involves a first derivative, it is taken to be $O(\epsilon)$ and the modulation of the carrier wave correspondingly slight. The modulation circuit is illustrated in Figure 4.1; the full amplifier circuit is caricatured in Figure 4.2. Now the carrier wave $v(t)$ is governed by

$$v(t) = -c_1 \int^t w(\tau) + r(\tau)\,\mathrm{d}\tau, \tag{4.1}$$

where c_1 is a time constant associated with the integrator in the modulation circuit (see Figure 4.1), and is no longer quite piecewise linear. The signal $w(t)$ is a square wave of variable duty cycle, taking the values $w(t) = \pm 1$, depending on the carrier wave $v(t)$ as follows. First, suppose that $w = -1$; then v is increasing. When v reaches $+1$, w changes to $+1$ and v starts to decrease. Once v has decreased to -1, w changes to -1; then v starts to increase once more, until it reaches $+1$ again, at

which point the cycle starts over.

4.1. Analysis of the model. Suppose that $w(t)$ changes to the value $w = 1$ at times $t = T_n$, and changes to the value $w = -1$ at times $t = U_n$. Then for $T_n < t < U_n$, $w(t) = 1$ and hence

$$v(t) = 1 - c_1[e(t) - e(T_n)] - c_1(t - T_n). \tag{4.2}$$

For $U_n < t < T_{n+1}$, $w(t) = -1$ and

$$v(t) = -1 - c_1[e(t) - e(U_n)] + c_1(t - U_n). \tag{4.3}$$

The time constant c_1 (determined below) is such that $c_1 T = O(1)$. Constants of integration have been chosen so that each of these expressions gives the correct value for $v(t)$ at the start of each time interval. Imposing the appropriate value for $v(t)$ at the end of each time interval then gives the two conditions

$$c_1[e(U_n) - e(T_n)] + c_1(U_n - T_n) = 2, \tag{4.4}$$

$$-c_1[e(T_{n+1}) - e(U_n)] + c_1(T_{n+1} - U_n) = 2. \tag{4.5}$$

For the special case in which $r(t) \equiv r_0$ is constant, the carrier-wave $v(t)$ is time-periodic, with period of oscillation

$$T = \frac{4}{(1 - r_0^2)c_1} \sim \frac{4}{c_1}(1 + r_0^2).$$

Furthermore, in this case

$$U_n - T_n = \frac{2}{(1 + r_0)c_1} \sim \frac{2}{c_1}(1 - r_0).$$

Note that T is in general *increased* by the presence of a nonzero modulation signal (i.e., the frequency of the carrier wave is reduced). In what follows, we shall require corrections to the carrier-wave due to modulation only up to $O(\epsilon)$, and thus, since $r^2 = O(\epsilon^2)$, we may take T as fixed. Then the time constant c_1 must be chosen so that

$$c_1 = \frac{4}{T}. \tag{4.6}$$

With this approximation, it turns out that we may consistently calculate terms in $g_a(t)$ up to $O(\epsilon^2)$, which is sufficient to determine the effects of carrier-wave modulation on the amplifier's distortion characteristics.

If we write the times at which the slope of the triangular wave changes as

$$T_n = nT + a_n, \qquad U_n = nT + b_n, \tag{4.7}$$

where $0 < a_n < b_n < T$, then these times are now governed by

$$\begin{aligned} 0 = {} & h(nT) - c[f(nT + \alpha_n) - f(nT)] - c\alpha_n + ks(nT + \alpha_n) \\ & + 1 - c_1[e(nT + \alpha_n) - e(nT + a_n)] - c_1(\alpha_n - a_n), \end{aligned} \tag{4.8}$$

$$\begin{aligned} 0 = {} & h(nT) - c[f(nT + \beta_n) - f(nT)] + c(\beta_n - 2\alpha_n) + ks(nT + \beta_n) \\ & - 1 - c_1[e(nT + \beta_n) - e(nT + b_n)] + c_1(\beta_n - b_n), \end{aligned} \tag{4.9}$$

rather than by (3.6) and (3.9). As in the absence of modulation, we have (3.14).

The solution technique is just as with no modulation. Again we seek slowly varying generalized switching times $A(t)$ and $B(t)$ of the output $g(t)$, but now we need two further slowly varying functions, $C(t)$ and $D(t)$, such that

$$C(nT) = a_n, \qquad D(nT) = b_n. \tag{4.10}$$

The equations to be solved arise from (3.14), (4.4), (4.5), (4.8), and (4.9), and are

$$c_1[e(t + D(t)) - e(t + C(t))] + c_1(D(t) - C(t)) = 2, \tag{4.11}$$

$$-c_1[e(t + T + C(t + T)) - e(t + D(t))] + c_1(T + C(t + T) - D(t)) = 2, \tag{4.12}$$

$$\begin{aligned} &\eta(t) - c[f(t + A(t)) - f(t)] - cA(t) + ks(t + A(t)) \\ &+ 1 - c_1[e(t + A(t)) - e(t + C(t))] - c_1(A(t) - C(t)) = 0, \end{aligned} \tag{4.13}$$

$$\begin{aligned} &\eta(t) - c[f(t + B(t)) - f(t)] + c(B(t) - 2A(t)) + ks(t + B(t)) \\ &- 1 - c_1[e(t + B(t)) - e(t + D(t))] + c_1(B(t) - D(t)) = 0, \end{aligned} \tag{4.14}$$

$$\eta(t + T) - \eta(t) + c[f(t + T) - f(t)] - c[2B(t) - 2A(t) - T] = 0. \tag{4.15}$$

As with the simpler case of an unmodulated carrier-wave, we expand all unknown functions (here A, B, C, D, and η) as series, as in (3.16), and solve in succession for the terms in these series at the first few orders.

4.1.1. Discussion. When expanded, (4.11) and (4.12) each yield at $O(1)$ and at $O(\epsilon)$ identical equations of the forms

$$D_0(t) - C_0(t) = \frac{1}{2}T, \qquad D_1(t) - C_1(t) = -\frac{1}{2}Tr(t), \tag{4.16}$$

respectively. The fact that only the difference between the times C and D may be determined partly reflects an arbitrariness in the time origin for the circuit that generates the carrier wave. However, if we continue to the next order we find that the two equations for $D_2(t) - C_2(t)$ are in fact inconsistent, reflecting the more serious limitation imposed upon the analysis by our assumption that the mean carrier-wave period is unaltered by the modulation. Fortunately, a consistent calculation of C and D up to terms C_1 and D_1 proves sufficient to determine the audio output of the amplifier up to g_2, which allows us to calculate the elimination of the distortion.

4.1.2. Elimination of the distortion. We find the audio output to be

$$\begin{aligned} g_a(t) = &-s(t) + \frac{1+k}{c}s'(t) \\ &- \frac{1}{3c_1^2c^2}\left\{3c_1c^2(rs^2)' + \left[3(1+k)c_1^2 - c^2\right]s'' - 3c_1c^2r' - c^2(s^3)''\right\} + O(\epsilon^3), \end{aligned} \tag{4.17}$$

where c_1 is given by (4.6). There are two nonlinear distortion terms in this expression, proportional to $(rs^2)'$ and $(s^3)''$. If we set $r = \nu s'(t)$, we may eliminate both of them by choosing $\nu = 1/c_1 = T/4$. Then

$$g_a(t) = -s(t) + \frac{1+k}{c}s'(t) - \frac{1}{12c^2}\left[12(1+k) - c^2T^2\right]s''(t) + O(\epsilon^3). \tag{4.18}$$

Thus all nonlinear distortion is removed, at least to the order calculated. A key result of the present analysis is that the appropriate modulation of the carrier wave is through a derivative of the input signal; this is, in fact, the method used in practice.

Now if we choose $k^2 = 1 - \frac{1}{6}c^2T^2$, then the audio output is of the form

$$g_a(t) = -s\left(t - \frac{(1+k)}{c}\right) + O(\epsilon^3), \tag{4.19}$$

and, to the order calculated, there is no distortion beyond a slight delay to the output.

5. "Sample-and-hold" class-D amplifier with negative feedback. Modulation of the carrier wave is not the only means by which the leading-order nonlinear distortion can be removed from class-D amplifiers with negative feedback. We now describe an alternative means of eliminating this distortion, without carrier-wave modulation. This alternative amplifier design has also been constructed, in prototype, by Halcro and is illustrated in Figure 5.1. There is no modulation of the carrier-wave symmetry.

Now the output of the integrator inputs to a sample-and-hold device, which samples $h(t)$ at times $t = nT$ and $(n + \frac{1}{2})T$; its output $p(t)$ is then a piecewise-constant function. For $nT \le t < (n + \frac{1}{2})T$, $p(t)$ takes the value $h(nT)$, while for $(n + \frac{1}{2})T \le t < (n + 1)T$, $p(t)$ takes the value $h((n + \frac{1}{2})T)$. Aside from this new feature, most details of the model remain essentially unchanged. The triangular wave $v(t)$ again satisfies (2.1) and $v(t + T) = v(t)$ for all t. The output $g(t)$ of the comparator is now $+1$ if $p(t) + ks(t) + v(t) > 0$, and -1 if the inequality is reversed; the switching times thus satisfy

$$p(nT + \alpha_n) + ks(nT + \alpha_n) + v(nT + \alpha_n) = 0,$$
$$p(nT + \beta_n) + ks(nT + \beta_n) + v(nT + \beta_n) = 0.$$

The integrator output $h(t)$ is again given by (3.2). In any interval $nT < t < (n+1)T$, there are three subintervals, as in (3.3); we describe these below. The analysis is somewhat simplified by the sampling.

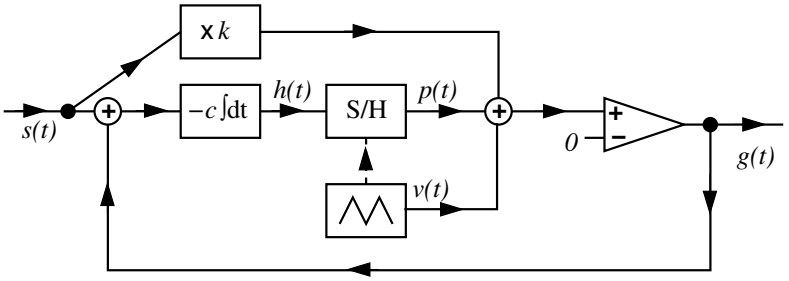


FIG. 5.1. *"Sample-and-hold" class-D amplifier with negative feedback. The sample-and-hold (S/H) device is synchronized with the carrier-wave generator and gives an output $p(t)$ equal to its input $h(t)$ sampled at times $t = nT$ and $t = (n + \frac{1}{2})T$. Thus for $nT \le t < (n + \frac{1}{2})T$, $p(t) = h(nT)$; correspondingly, for $(n + \frac{1}{2})T \le t < (n + 1)T$, $p(t) = h((n + \frac{1}{2})T)$.*

Subinterval I. By integrating (3.2) we find that $h(t)$ is again given by (3.4), but now the value of α_n is defined by

$$p(nT + \alpha_n) + ks(nT + \alpha_n) + v(nT + \alpha_n) = 0; \tag{5.1}$$

that is,

$$h(nT) + ks(nT + \alpha_n) + 1 - \frac{4\alpha_n}{T} = 0. \tag{5.2}$$

Subinterval II. By integrating (3.2) we again find (3.7) for $h(t)$; the value of β_n is defined by

$$p(nT+\beta_n)+ks(nT+\beta_n)+v(nT+\beta_n)=0; \tag{5.3}$$

that is,

$$h((n+\tfrac{1}{2})T)+ks(nT+\beta_n)-3+\frac{4\beta_n}{T}=0. \tag{5.4}$$

Subinterval III. By integrating (3.2) we find (3.10) for $h(t)$; it follows that $h((n+1)T)$ is again given by (3.11).

5.1. Solution of the governing equations. The three governing equations are now

$$\frac{4}{T}\alpha_n = 1+ks(nT+\alpha_n)+h(nT), \tag{5.5}$$

$$\frac{4}{T}\beta_n = 3-ks(nT+\beta_n)-h\left(\left(n+\frac{1}{2}\right)T\right), \tag{5.6}$$

$$h((n+1)T)=h(nT)-c[f((n+1)T)-f(nT)]+c(2\beta_n-2\alpha_n-T). \tag{5.7}$$

Furthermore, by considering subinterval II, we find that

$$h\left(\left(n+\frac{1}{2}\right)T\right)=h(nT)-c\left[f\left(\left(n+\frac{1}{2}\right)T\right)-f(nT)\right]+c\left(\frac{T}{2}-2\alpha_n\right). \tag{5.8}$$

As above, we introduce the slowly varying functions $A(t)$, $B(t)$, and $\eta(t)$, which now satisfy

$$\begin{aligned}
\frac{4}{T}A(t) &= 1+ks(t+A(t))+\eta(t),\\
\frac{4}{T}B(t) &= 3-ks(t+B(t))-\eta(t)+c\left[f\left(t+\frac{T}{2}\right)-f(t)\right]-c\left[\frac{T}{2}-2A(t)\right],\\
\eta(t+T) &= \eta(t)-c[f(t+T)-f(t)]+c[2B(t)-2A(t)-T],
\end{aligned}$$

and solve these equations at successive orders in ϵ. The audio output is eventually found to be

$$\begin{aligned}
g_a(t) = -s(t)+\frac{1+k}{c}s'(t)&\\
+\frac{1}{96c^2(4-cT)}\Big\{&\left[c^3T^3-(28+24k)c^2T^2+192(1+k)cT-384(1+k)\right]s(t)\\
&+\left[cT-4(1+2k)\right]c^2T^2s^3(t)\Big\}''+O(\epsilon^3).
\end{aligned} \tag{5.9}$$

Now the nonlinear distortion term proportional to $(s^3)''$ may be removed from this expression by choosing

$$k=-\frac{1}{2}+\frac{1}{8}cT, \tag{5.10}$$

in which case

$$g_a(t)=-s(t)+\frac{4+cT}{8c}s'(t)-\frac{24-c^2T^2}{48c^2}s''(t)+O(\epsilon^3). \tag{5.11}$$

Once the nonlinear distortion term has been so removed, we may similarly remove linear distortion, so that the audio output suffers only a slight delay, i.e.,

$$g_a(t) = -s\left(t - \frac{\frac{1}{8}(4+cT)}{c}\right) + O(\epsilon^3),$$

by choosing $cT = 12(2\sqrt{3}-1)/11 \approx 2.6881$.

5.2. Alternative means of sampling. Suppose now that the sampling is carried out only at times $t = nT$ (i.e., not also at times $t = (n+\frac{1}{2})T$). Then the equations to be solved for $A(t)$, $B(t)$, and $\eta(t)$ are somewhat simplified:

$$\frac{4}{T}A(t) = 1 + ks(t + A(t)) + \eta(t), \tag{5.12}$$

$$\frac{4}{T}B(t) = 3 - ks(t + B(t)) - \eta(t), \tag{5.13}$$

$$\eta(t+T) - \eta(t) = -c[f(t+T) - f(t)] + c[2B(t) - 2A(t) - T]. \tag{5.14}$$

The audio output is, correspondingly, found to be

$$\begin{aligned} g_a(t) = -s(t) + \frac{1+k}{c}s'(t) + \frac{1}{96c^2}\big\{&-(2k+1)c^2T^2(3s^2(t)+s^3(t))'' \\ &+ \left[-c^2T^2 - 6(1+k)c^2T^2 + 48(1+k)cT - 96(1+k)\right]s''(t)\big\}. \end{aligned} \tag{5.15}$$

In view of the asymmetrical sampling, there are now second and third harmonics at $O(\epsilon^2)$, but these can *simultaneously* be removed by choosing

$$k = -\frac{1}{2}.$$

If this choice is made, then $g_a(t)$ becomes

$$g_a(t) = -s(t) + \frac{1}{2c}s'(t) - \frac{1}{24c^2}\left[12 - 6cT + c^2T^2\right]s''(t) + O(\epsilon^3). \tag{5.16}$$

This is a delayed version of the original signal ($g_a(t) = -s(t - \frac{1}{2}c) + O(\epsilon^3)$) if the choice $cT = 3$ is made.

6. A new class-D amplifier, with reduced distortion. We now describe a third modification to the standard negative-feedback class-D amplifier, which eliminates the intrinsic distortion at $O(\epsilon^2)$. While the two designs described above in sections 4 and 5 were developed first on physical principles and subsequently modeled here mathematically, this new design arose as a consequence of the mathematical models described herein. Prototypes do indeed enjoy significant distortion reduction.

To see how this new design is derived, we consider adding to the noninverting input of the comparator a function $F(t)$ such that $F(t)$ is constant over each interval $nT \le t < (n+1)T$. At present the values taken by this function over each interval are arbitrary; we shall compute the effects of $F(t)$ on the audio output spectrum, then choose it so as to cancel out the intrinsic distortion.

With this additional design feature, the audio output of the amplifier is found to be

$$\begin{aligned} g_a(t) = -s(t) &+ \frac{1}{c}\left[(1+k)s(t) + \theta(t)\right]' \\ &- \frac{1}{48c^2}\left\{\left(48 + 24cT - 3c^2T^2\right)\theta'(t) + \left(48(1+k) - c^2T^2\right)s'(t)\right. \\ &\qquad \left. + 3c^2T^2\left[\theta(t) - s(t)\right]'s^2(t)\right\}' + O(\epsilon^3), \end{aligned} \tag{6.1}$$

where $\theta(t)$ is a slowly varying function that agrees with $F(t)$ at times $t = nT$. The last line of (6.1) represents the nonlinear distortion, and it is clear that this component can be eliminated by choosing

$$F(nT) = s(nT), \tag{6.2}$$

and hence $\theta(t) = s(t)$. Fortunately, no further distortion is introduced by this choice for $F(t)$, and we have

$$g(t) = -s(t) + \frac{2+k}{c}s'(t) - \frac{12(2+k) + 6cT - c^2T^2}{12c^2}s''(t) + O(\epsilon^3), \tag{6.3}$$

which is free from any nonlinear distortion. As in the other models above, it is possible to choose k so that the output is, to the order calculated, a delayed version of the input signal and suffers no further distortion beyond the slight delay.

7. Nonlinearity in the carrier wave. We now consider one way in which imperfect electronic components can introduce distortion into the output. Specifically, we note that it is difficult in practice to generate a high-frequency triangular carrier wave whose slopes are precisely linear. In general the wave comprises sections of exponential functions, which approximate very closely the desired piecewise-linear profile [6]. For instance, let us suppose that instead of (2.1) we have for the carrier wave

$$v(t) = \begin{cases} 1 - \dfrac{2(\mathrm{e}^{t/t_0} - 1)}{\mathrm{e}^{T/2t_0} - 1} \equiv v_1(t) & \text{for } 0 \le t < \dfrac{T}{2}, \\[2ex] -1 + \dfrac{2(\mathrm{e}^{(t-T/2)/t_0} - 1)}{\mathrm{e}^{T/2t_0} - 1} \equiv v_2(t) & \text{for } \dfrac{T}{2} \le t < T, \end{cases} \tag{7.1}$$

and $v(t+T) = v(t)$ for all t. Note that the piecewise-linear profile of (2.1) is recovered as $t_0/T \to \infty$.

7.1. "Classical" class-D amplifier design. Let us first consider the effects of carrier-wave nonlinearity on the classical class-D amplifier design, without negative feedback. Here switching of $g(t)$ takes place whenever $v(t) + s(t) = 0$, i.e., when

$$v_1(nT + \alpha_n) + s(nT + \alpha_n) = 0 \quad \text{or} \quad v_2(nT + \beta_n) + s(nT + \beta_n) = 0. \tag{7.2}$$

These expressions are readily rearranged to give implicit equations for the switching times:

$$\alpha_n = t_0 \log\left\{1 + \tfrac{1}{2}[1 + s(nT + \alpha_n)](\mathrm{e}^{T/2t_0} - 1)\right\}, \tag{7.3}$$

$$\beta_n = \tfrac{1}{2}T + t_0 \log\left\{1 + \tfrac{1}{2}[1 - s(nT + \beta_n)](\mathrm{e}^{T/2t_0} - 1)\right\}. \tag{7.4}$$

It now follows readily from (2.27) that the audio output is

$$g_a(t) = \frac{2t_0}{T} \log \frac{1 + \frac{1}{2}[1 + s(t)](\mathrm{e}^{T/2t_0} - 1)}{1 + \frac{1}{2}[1 - s(t)](\mathrm{e}^{T/2t_0} - 1)}. \tag{7.5}$$

(It is straightforward from this expression to check that $g(t) \sim s(t)$ as $t_0/T \to \infty$, in accordance with (2.29).) Since it follows by Taylor expansion that

$$g_a(t) \sim \frac{4t_0}{T} \sum_{n=0}^{\infty} \frac{1}{2n+1} \left[\frac{(\mathrm{e}^{T/2t_0} - 1)s(t)}{\mathrm{e}^{T/2t_0} + 1}\right]^{2n+1}, \tag{7.6}$$

we may in principle compute the audio spectrum, for instance, due to a sinusoidal input $s(t)$.

We contrast the result here, where $g_a \not\equiv s$ and there is nonlinear distortion of $O(T/t_0)$, with that for a perfectly piecewise-linear carrier wave, where there is no distortion to the output.

7.2. Class-D amplifier with negative feedback. We now consider the effects of carrier-wave nonlinearity on the class-D amplifier with negative feedback (as illustrated in Figure 3.1). It turns out that, provided $T/t_0 \ll 1$, then to leading order in the nonlinearity, the output $g_a(t)$ as given by (3.29) is augmented by a term of the form

$$\frac{T^2}{16t_0}\left(s^3(t) - s(t)\right)'. \tag{7.7}$$

Note that the nonlinear distortion term here (proportional to $(s^3)'$) can be thought of as having a phase different from that inherent in the basic amplifier design (proportional to $(s^3)''$), so one cannot be used to cancel the other.

However, if we modify the design by adding to the comparator input a quantity

$$-\frac{1}{ct_0}h(t) \tag{7.8}$$

sampled at times

$$t = nT \quad \text{and} \quad \left(n + \frac{1}{2}\right)T, \tag{7.9}$$

in addition to the modification proposed in section 6, then the third-harmonic distortion term due to the carrier wave nonlinearity is canceled, and

$$\begin{aligned} g_a(t) &= -s(t) + c^{-1}(2+k)s'(t) \\ &\quad + c^{-2}\left[-\frac{1}{12}\left(12(2+k) + 6cT - c^2T^2\right)s''(t) + t_0^{-1}(2+k)s'(t)\right] + O(\epsilon^3). \end{aligned} \tag{7.10}$$

With an appropriate choice for k, this expression is essentially just a slightly delayed version of the input signal, i.e.,

$$g_a(t) = -s(t - t_1) + O(\epsilon^3). \tag{7.11}$$

To achieve this simplification we take

$$k = -1 + \left(1 + cT - \frac{1}{6}c^2T^2\right)^{1/2};$$

then the delay is

$$t_1 = \frac{2+k}{c}\left(1 + \frac{1}{ct_0}\right).$$

To the order calculated, there is no further distortion; the delay computed here is independent of the signal amplitude or frequency.

8. Conclusions. We have developed mathematical models for a variety of class-D amplifier designs. While models for the classical design with no negative feedback have been known for some time [1, 4, 5], the models presented here appear to be the first to treat in detail the negative-distortion design. One model describes the use of modulation to the carrier wave symmetry in order to reduce the intrinsic distortion of the negative-feedback design; another describes the use of a sample-and-hold device to achieve essentially the same ends. The mathematical analysis has, in each case, given theoretical backing to the design idea and quantitative statements about the parameter sets under which the designs are effective. A new means of reducing the intrinsic distortion has been proposed, on the basis of the mathematical models developed, which involves the use of a sample-and-hold device, but in a manner different from that in the existing design.

One of the authors (BHC) has tested all three designs (i.e., those described in sections 4, 5, and 6) in prototype and found them all to achieve significant reduction in harmonic distortion. The two designs involving a sample-and-hold unit are found not to work as well in practice as the carrier-wave modulation system, and are more expensive to produce. The carrier-wave modulation design is the basis of a successful commercial amplifier manufactured by Halcro.

It should be noted that the models developed here do not reflect a range of important practical issues, such as the noise and stability characteristics of the designs, nor their electromagnetic emissions. The models assume perfect components, an assumption that has particularly significant shortcomings in relation to sample-and-hold devices, for which the errors are relatively severe (in comparison with, say, integrators).

The models developed in this paper appear to be the first to provide an in-depth mathematical treatment of class-D amplifiers with negative feedback, and should be capable of extension to more complicated designs that reflect more accurately actual audio amplifiers.

REFERENCES

[1] H. S. Black, *Modulation Theory*, Van Nostrand, New York, 1953.

[2] P. H. Mellor, S. P. Leigh, and B. M. G. Cheetham, *Reduction of spectral distortion in class D amplifiers by an enhanced pulse width modulation sampling process*, IEE Proc. G, 138 (1991), pp. 441–448.

[3] L. M. Milne-Thomson, *The Calculus of Finite Differences*, Macmillan, London, 1933.

[4] C. Pascual, Z. Song, P. T. Krein, D. V. Sarwate, P. Midya, and W. J. Roeckner, *High-fidelity PWM inverter for digital audio amplification: Spectral analysis, real-time DSP implementation, and results*, IEEE Trans. Power Electronics, 18 (2003), pp. 473–485.

[5] Z. Song and D. V. Sarwate, *The frequency spectrum of pulse width modulated signals*, Signal Processing, 83 (2003), pp. 2227–2258.

[6] M. T. Tan, J. S. Chang, H. C. Chua, and B. H. Gwee, *An investigation into the parameters affecting total harmonic distortion in low-voltage low-power Class-D amplifiers*, IEEE Trans. Circuits Systems I, 50 (2003), pp. 1304–1315.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 489–509

ACOUSTIC PROPAGATION IN DISPERSIONS IN THE LONG WAVELENGTH LIMIT*

V. J. PINFIELD†, O. G. HARLEN‡, M. J. W. POVEY†, AND B. D. SLEEMAN‡

Abstract. The problem of scattering of ultrasound by particles in the long wavelength limit has a well-established solution in terms of Rayleigh expansions of the scattered fields. However, this solution is ill-conditioned numerically, and recent work has attempted to identify an alternative method. The scattered fields have been expressed as a perturbation expansion in the parameter Ka (the wavenumber multiplied by the particle radius), which is small in the long wavelength region. In the work reported here the problem has been formulated so as to be valid for all values of the thermal wavelength, which varies in order of magnitude, from much smaller to much larger than the particle size in the long wavelength region. Thus the present solution overlaps the limiting solutions for very small thermal wavelength (geometric theory) and very large thermal wavelength (low frequency) previously reported. Close agreement is demonstrated with the established Rayleigh expansion solution.

Key words. Helmholtz equation, scattering theory, ultrasound spectroscopy, dispersions

AMS subject classifications. 35C10, 35J05, 35P25, 76Q05

DOI. 10.1137/04061698X

1. Introduction.

Ultrasound spectroscopy is an increasingly popular technique for characterizing the physical properties of dispersions, emulsions, gels, and solutions of biomolecules. It is a noninvasive technique that can address the extensive range of particle sizes encountered in many particulate systems and can be used with optically opaque materials. The technique has been adopted in manufacturing processes in the food and chemical industries. Ultrasonic instruments may be used to determine particle size distribution and/or concentration of the dispersed particles. In order to do this, it is necessary to use a strong theoretical basis to relate the ultrasound properties, i.e., velocity and attenuation, to the particle size and physical properties of the materials.

The problem of scattering of sound waves by a single spherical object (a fluid droplet) was solved by Rayleigh [1] and later refined by Epstein and Carhart [2]. A similar problem, with solid particles, was addressed by Allegra and Hawley [3]. Their solution is referred to as ECAH. The scattered fields are expanded as spherical harmonics in order to allow the application of boundary conditions at the particle surface. Although the solution is analytically exact, its numerical solution can be troublesome, because the matrix equation is ill-conditioned, and the series does not converge uniformly. In addition, calculation of spherical Bessel functions at large complex arguments is imprecise, and at large distances the Hankel functions oscillate rapidly. Such numerical limitations cause difficulty in applying the ultrasound method in practical applications.

*Received by the editors October 14, 2004; accepted for publication (in revised form) April 27, 2005; published electronically December 30, 2005. This research was supported by the UK Engineering and Physical Research Council (EPSRC), grant GR/L/51034.
http://www.siam.org/journals/siap/66-2/61698.html
†Procter Department of Food Science, University of Leeds, Leeds LS2 9JT, UK (v.j.pinfield@leeds.ac.uk, m.j.w.povey@leeds.ac.uk).
‡Department of Applied Mathematics, University of Leeds, Leeds LS2 9JT, UK (o.g.harlen@leeds.ac.uk, b.d.sleeman@leeds.ac.uk).

The aim of the present work in this area has been to formulate a numerically stable solution to the single scatterer problem in such a way as to allow its potential extension to multiple scatterers and to nonspherical scatterers. The frequency range of interest is termed the *long wavelength limit*, in which the wavelength of the propagating acoustic wave is much larger than the size of the droplet ($Ka \ll 1$, where K is the wavenumber and a the radius of the particles). In this case, Kleinman's approach can be applied, in which the problem is reformulated to satisfy the radiation condition and uses a perturbation series solution in powers of Ka. At the lowest frequencies, at which the wavelength of the thermal waves produced by scattering is also much larger than the particle radius ($La \ll 1$, where L is the thermal wavenumber), the same technique can be applied to the thermal wave. The method was applied to single particle scattering in the *low-frequency potential scattering theory* (LFPST) previously published (Harlen et al. [4]). A later paper (Harlen et al. [5]) considered the case $La \gg 1$, where the thermal wavelength is much smaller than the particle size, and the Kleinman series expansion in La cannot be used. In this case, a geometric theory method was developed to approximate the solution for the thermal waves while retaining the Kleinman technique for the propagational waves.

In the work reported here, the Kleinman principles have again been used to separate the radiative terms of the waves and to define a series expansion which is convergent for the long wavelength limit ($Ka \ll 1$). The significant difference is that all wave modes are expanded as a series in Ka, leaving dependence on the thermal wavenumber implicit in the coefficients. This avoids assumptions on the size of La.

In the next section, the propagation of sound in fluids is considered in the context of a plane wave incident on a isolated spherical particle. The full ECAH method for the solution of the scattering problem is summarized and the perturbation expansion technique introduced. Sections 3 and 4 define the analytical forms of the wave potentials outside and inside the particle. Section 5 constructs the solution to the scattering problem in the perturbation method, showing the general solution and explicit results for the first few terms. Calculations are presented in section 6 to show that the method agrees with the full Rayleigh expansion method.

2. Sound fields in a fluid. The principles of sound propagation in homogeneous fluids are well documented, and only the most important results are given here. The equations of conservation of mass (continuity equation), momentum (Navier–Stokes equation), and energy, together with some thermodynamic relations, can be simplified by use of a velocity potential, ϕ, such that

$$\mathbf{u} = -\nabla\phi, \tag{1}$$

where $\mathbf{u}$ is the velocity of the fluid. The resulting biharmonic equation is further separated by defining two potentials, one for each of two types of wave mode (propagational and thermal). Propagational modes are the "usual" mode by which sound travels in a fluid. The thermal mode represents heat flow and is dissipative and therefore highly localized. There is an additional solution to the equations resulting from use of a vector potential, which corresponds to shear wave modes. Again these are dissipative with a very short decay length in fluids. In many practical applications of ultrasound, the shear wave modes resulting from scattering at dispersed particles are small. Hence for the subsequent analysis and in the previously published work (Harlen et al. [4], [5]), the vector potential solution is neglected. It was, however, included in the ECAH solutions.

Periodic solutions for the wave potentials have a time dependence defined by a factor $e^{-i\omega t}$, where ω is the angular frequency, which results in two separate Helmholtz equations, one each for the propagational and thermal modes, thus:

$$\left(\nabla^2 + K^2\right)\varphi = 0, \quad \left(\nabla^2 + L^2\right)\psi = 0. \tag{2}$$

The overall velocity potential in the fluid is the sum of the propagational (φ) and thermal (ψ) potentials.

The wavenumbers are given to a very good approximation in fluids by

$$K = \frac{\omega}{v}\left(1 + i\frac{(\gamma-1)\,\sigma\omega}{2v^2}\right), \quad L = \left(\frac{\omega}{2\sigma}\right)^{1/2}(1+i), \tag{3}$$

where γ is the ratio of specific heat capacities, v the speed of sound, and σ the thermal diffusivity, such that $\sigma = \tau/\rho C_p$, where τ is the thermal conductivity, ρ is the density, and C_p is the specific heat capacity at constant pressure.

The common expression for the wavenumber of the propagational mode has the form

$$K = \frac{\omega}{v} + i\alpha, \tag{4}$$

where α is the attenuation. There are many absorption effects in fluids which are not accounted for in the classical thermal and fluid momentum equations used to derive the wavenumber in (3). Hence the measured attenuation should be used instead, as in (4).

The pressure and temperature fluctuations which result from the wave motions are related to the velocity potentials as follows:

$$P = -i\omega\rho\,(\varphi + \psi), \quad T = \Gamma_c\varphi + \Gamma_t\psi, \tag{5}$$

where the thermal factor for each wave mode is

$$\Gamma_c = \frac{-iK^2\,(\gamma-1)}{\beta\,(\omega + i\gamma\sigma K^2)}, \quad \Gamma_t = \frac{-iL^2\,(\gamma-1)}{\beta\,(\omega + i\gamma\sigma L^2)}. \tag{6}$$

Subscript c is used to denote the compressional (or propagational) mode and t the thermal mode. Note that these temperature factors were quoted incorrectly in the previous paper (Harlen et al. [5]).

A useful thermodynamic relation is

$$\gamma - 1 = \frac{v^2\beta^2 T_0}{C_p}, \tag{7}$$

where β is the thermal expansivity and T_0 is the temperature of the system, not the small temperature changes caused by the wave motion.

2.1. Scattering of sound waves by particles. In order to calculate the ultrasound field produced by a dispersion of particles, it is first necessary to consider the effect on a sound wave of a single particle immersed in isolation in an infinite uniform fluid. The most relevant and simple system to study is that of a plane wave of angular frequency ω incident on a spherical particle of radius a. The fluid inside the particle has different physical properties and so will respond in a different way from the fluid surrounding it to the compression and rarefaction of the wave. Scattered waves of

each mode are produced inside and outside the particle. At the surface of the particle certain boundary conditions must be met. These include the requirement that the boundary not be disrupted, i.e., that material immediately inside and outside of the particle move at the same speed, and that temperature and heat flow be continuous (the same either side of the interface). In terms of the wave potentials, the boundary conditions are as follows.

The normal velocity of fluid on both sides of the boundary must be equal to avoid formation of a void:

$$\frac{\partial}{\partial r}\left(\varphi_0 + \varphi + \psi\right) = \frac{\partial}{\partial r}\left(\varphi' + \psi'\right). \tag{8}$$

The pressure must be equal on each side of the boundary:

$$\varphi_0 + \varphi + \psi = \hat{\rho}\left(\varphi' + \psi'\right). \tag{9}$$

The temperature must be equal on each side of the boundary:

$$\Gamma_c\varphi_0 + \Gamma_c\varphi + \Gamma_t\psi = \Gamma'_c\varphi' + \Gamma'_t\psi'. \tag{10}$$

The heat flux must be equal on each side of the boundary:

$$\Gamma_c\frac{\partial}{\partial r}\left(\varphi_0 + \varphi\right) + \Gamma_t\frac{\partial\psi}{\partial r} = \hat{\tau}\left(\Gamma'_c\frac{\partial\varphi'}{\partial r} + \Gamma'_t\frac{\partial\psi'}{\partial r}\right), \tag{11}$$

where primed quantities refer to the inside of the particle,

$$\hat{\rho} = \frac{\rho'}{\rho}, \quad \hat{\tau} = \frac{\tau'}{\tau}, \tag{12}$$

and φ_0 is the potential of the incident wave.

The objective is to determine the amplitude and phase of the scattered propagational mode, which is the only part which is still nonnegligible at a significant distance from the particle (the thermal field having decayed to zero). Other published work is used to determine the wavenumber for a dispersion of particles, by a multiple scattering approach, to obtain the net effect of many particles. For the present work, the aim is to calculate the scattered wave amplitude.

In order to obtain the solution to the scattering problem, a general form for each wave mode must be proposed. The potentials of the scattered fields must be solutions of the appropriate Helmholtz equation (2), whether inside or outside the particle. In addition, the waves inside the particle must be defined at the origin (the center of the particle), and those outside the particle must satisfy the radiation condition. Finally, the boundary conditions at the surface of the particle must be satisfied. In sections 3 and 4 appropriate forms for the solutions are constructed to allow the scattering problem to be resolved.

2.2. ECAH method. The Epstein and Carhart method [2] for the scattering problem expanded the solutions of the Helmholtz equation in spherical coordinates. The solutions are Rayleigh series in the spherical harmonics, that is, a combined series in the spherical Bessel functions (for the radial dependence) and Legendre polynomials (for the angular dependence). The Bessel functions are chosen appropriately for the region in which the wave exists; in the continuous phase the solution must be defined at large distances, so the Hankel function h_n is used, whereas inside the particle the

solution must be defined at the origin, so the j_n function is used. ECAH took the forms for each wave potential to be as follows:

$$\begin{aligned}
\varphi &= \sum_{n=0}^{\infty} i^n (2n+1) A_n h_n (Kr) P_n (\cos\theta), \\
\psi &= \sum_{n=0}^{\infty} i^n (2n+1) B_n h_n (Lr) P_n (\cos\theta), \\
\varphi' &= \sum_{n=0}^{\infty} i^n (2n+1) A'_n j_n (K'r) P_n (\cos\theta), \\
\psi' &= \sum_{n=0}^{\infty} i^n (2n+1) B'_n j_n (L'r) P_n (\cos\theta).
\end{aligned} \tag{13}$$

Similarly the incident field (a plane wave) can be expressed as

$$\varphi_0 = \sum_{n=0}^{\infty} i^n (2n+1) j_n (Kr) P_n (\cos\theta). \tag{14}$$

In the ECAH method, these functions and the relevant derivatives are directly evaluated at the surface of the particle and substituted into the set of boundary conditions. For a spherical particle of radius a the Bessel functions must be determined at $r = a$. In the long wavelength limit, $|Ka| \ll 1$ and $|K'a| \ll 1$, the Hankel and Bessel functions can cause difficulties, and the difference in scale of the values appearing in the boundary equations results in an ill-conditioned matrix equation which must be solved.

In order to avoid these problems, a solution is sought which uses alternative forms for the wave potentials, both to avoid the direct use of the Bessel functions and to produce a direct solution not relying on a matrix inversion for its solution.

2.3. Kleinman method and Poincaré series. In the long wavelength limit, the condition $|Ka| \ll 1$ applies. Kleinman developed a method for solving low-frequency scattering problems (see Harlen et al. [4]) in which he expressed the potentials as a perturbation expansion, i.e., as a series in powers of iKa:

$$\phi = \sum_{m=0}^{\infty} (iKa)^m \phi_m, \tag{15}$$

where ϕ is one of the wave potentials. The series is known to converge rapidly, with an error bounded by $\mathrm{O}\left(|Ka|^{m+1}\right)$ if the mth order solution is used. Thus the problem becomes one of finding the solution to a set of potential functions. Although this may seem to increase the number of equations which must be solved, it avoids the ill-conditioned numerical calculation suffered by the ECAH approach and allows the series to be terminated with some confidence that an accurate result has been obtained. Further details of the method are given in Harlen et al. [4] and [5].

In the present work, *all* potentials are expanded as a series in iKa. The previous low-frequency work (Harlen et al. [4]) expanded each wave potential as a series in its appropriate wavenumber; for example, the thermal wave mode was defined as a series in powers of L. Later, the work on the short thermal wavelength region (geometric theory [5]) introduced combined power series, with positive powers of the

propagational wavenumbers, and inverse powers of the thermal wavenumbers. In the frequency range considered in that work, the values of $|La|$ and $|L'a|$ were large, hence an inverse power series was appropriate. For example, for a scattered propagational mode potential,

$$\phi = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{(iKa)^n}{(iLa)^m} \phi_{nm}. \tag{16}$$

In order to achieve a method which is valid across the entire frequency range in the long wavelength limit, it is appropriate to use only a power series in iKa since this value is by definition small over the entire range, $|Ka| \ll 1$, thus assuring convergence of the series. Series in positive or inverse powers of La are limited in their scope because of the variation in magnitude of this parameter within the long wavelength region. Its value ranges from small at low frequency to very large at the upper frequency limit of the long wavelength region. Hence in the current work, the power series in iKa is applied to each wave potential, with any dependence on other wavenumbers being implicit in the rest of the potential.

3. Solution forms outside the particle. In the continuous phase outside the particle, the solutions of the Helmholtz equation must also satisfy the radiation condition, which restricts its form at large distances from the particle. The Sommerfield radiation condition is as follows:

$$\lim_{r \to \infty} \left[r \left(\frac{\partial \phi}{\partial r} - ik\phi \right) \right] = 0. \tag{17}$$

In physical terms, the condition means that there is no energy radiating inwards from infinity. Thus the solution appears as an outgoing spherical wave at large distances from the particle (cf. Colton and Kress [6, p. 21]).

The spherical Hankel function used in the ECAH method is one such solution, each $h_n(kr)$ including a factor e^{ikr}/r which represents a spherical wave. However, radiating solutions to the Helmholtz equation are not regular at infinity, and it is the exponential part of the function which caused numerical difficulties at large arguments (for the thermal waves).

In general terms, the form of the solution is

$$\phi = \frac{e^{ikr}}{r} \tilde{\phi} = \frac{e^{ikr}}{r} \sum_{l=0}^{\infty} \frac{f_l(\vartheta, \Omega)}{r^l}, \tag{18}$$

where f_l is the angular dependence (Harlen et al. [4]). The function $\tilde{\phi}$ does not suffer from the mathematical difficulties of the overall potential ϕ, and is regular at infinity.

Following the previous method (Harlen et al. [4]), it is therefore appropriate to introduce new potential functions, $\tilde{\varphi}$ and $\tilde{\psi}$, for the propagational and thermal modes, respectively, outside the particle such that

$$\varphi = e^{iK(r-a)} \tilde{\varphi}, \tag{19}$$

$$\psi = e^{iL(r-a)} \tilde{\psi}. \tag{20}$$

The exponential spherical-wave factors have been explicitly taken out, so that the remaining functions $\tilde{\varphi}$ and $\tilde{\psi}$ are regular and differentiable. In addition when applying

boundary conditions for a spherical particle at $r = a$ the exponential factors do not contribute.

Continuing the Kleinman approach, the next step is to express the new potential functions as a series in iKa (see previous section), thus

$$\left(\tilde{\varphi}, \tilde{\psi}\right) = \sum_m (iKa)^m \left(\tilde{\varphi}_m, \tilde{\psi}_m\right). \tag{21}$$

3.1. Propagational mode. Using the spherical harmonic solutions to the Helmholtz equation, the partial fields for the propagational mode can then be written as

$$\tilde{\varphi}_m = \sum_{n=0}^{\infty} \sum_{j=0} A_{nmj} \cdot \frac{r^j}{a^j} \cdot \frac{a^{n+1}}{r^{n+1}} \cdot P_n(\cos\vartheta). \tag{22}$$

This form of the solution is suggested by the results of the LFPST method (Harlen et al. [4]), although it is expressed here as a general series in r. The full wave potential can be constructed using (21) and (19). The Helmholtz equation (2) can then be shown to relate the potentials of consecutive order m by the equation

$$\nabla^2 \tilde{\varphi}_m = -\frac{2}{ar} \frac{\partial}{\partial r} (r\tilde{\varphi}_{m-1}). \tag{23}$$

By substituting the general solution (22) into this form of the Helmholtz equation, and matching powers of (iKa) for each (spherical harmonic) order n, it can be shown that the coefficients are related by the following recurrence relation:

$$A_{n,m,j} = -\frac{2(j-1-n)}{j(j-1-2n)} A_{n,m-1,j-1} \text{ for } j \geq 1. \tag{24}$$

Thus coefficients for the potential of order m are related to those for the previous order. Only the $j = 0$ coefficient remains to be solved from the boundary equations. By definition, all coefficients for orders $m < 0$ are zero. Note that the coefficients are zero for $j \geq n + 1$, and hence the coefficients may be nonzero up to and including $j = n$, i.e., it is a finite series. The solution for the lowest orders (see section 5.6) demonstrates that, excepting $n = 0$, the first (in m) nonzero coefficient is for $m = n$, which implies that at larger orders m the last nonzero term in the j-series will be for $j = m - n$. The result also shows that (22) does give a solution of the Helmholtz equation. Although the propagational mode solution (22) does not appear to have the same form as that required for a radiating solution (18), it is clear that since the coefficients are nonzero only for $j \leq n$, each term in l (18) includes contributions from different m, n, and j combinations, which together give the angular dependence f_l.

The ECAH method expresses the propagational scattered wave in terms of the spherical Hankel functions $h_n(Kr)$. Our result is not simply a power series expansion in Kr of the Hankel function. This is because the wave potential has been written as a power series in iKa, which removes all the K-dependence, leaving a series in r, whose coefficients are to be determined. Contributions from different orders m make up the overall potential series in Kr.

3.2. Thermal mode. In the ECAH method, the thermal wave potential in the continuous phase was based on the spherical Hankel function $h_n(Lr)$. Since the perturbation series expansion (21) is taken in powers of (iKa), rather than in powers

of L, the thermal potential can be expressed simply using the series expansion of the spherical Hankel function. Thus

$$h_n(x) = e^{ix} \sum_{j=1}^{n+1} \frac{h_{nj}}{x^j} \tag{25}$$

(see the appendix for the factors h_{nj}), so that the thermal wave potential takes the form

$$\tilde{\psi}_m = \sum_{n=0}^{\infty} \sum_{j=1}^{n+1} B_{nm} \cdot \frac{h_{nj}}{(Lr)^j} \cdot P_n(\cos\vartheta). \tag{26}$$

All factors of L are taken implicitly as part of the potential function. The spherical Hankel function, in fact, results in a finite series in inverse powers of Lr, whose coefficients are known. These are all included in the appropriate term in the series in powers of $(iKa)^m$.

3.3. The incident wave. The incident wave is a plane wave and can be expressed as a series of spherical harmonics, as in the ECAH method (see (14)). In order to follow the same method as used for the other waves, the spherical Bessel function can be expanded as a power series in (iKa), using the power series expansion of the spherical Bessel function

$$j_n(x) = 2^n x^n \sum_{s=0}^{\infty} \frac{(-1)^s (s+n)!}{s!\,(2s+2n+1)!} x^{2s} \tag{27}$$

(Arfken [7, p. 625]).

Thus the plane wave can be written

$$\varphi_0 = \sum_{n=0}^{\infty} \sum_{s=0}^{\infty} (iKa)^{n+2s} \left(\frac{r}{a}\right)^{n+2s} F_n(s)\, P_n(\cos\vartheta), \tag{28}$$

where

$$F_n(s) = \frac{2^n (2n+1)(s+n)!}{s!\,(2s+2n+1)!}, \tag{29}$$

where n and s are nonnegative integers (F is zero otherwise). For purposes of numerical calculation, the factorial functions suffer from overflow for all but very low orders (n, s). The following recurrence relations can be used for accurate calculation:

$$F_0(0) = 1, \tag{30}$$

$$\frac{F_n(0)}{F_{n-1}(0)} = \frac{1}{(2n-1)} \text{ for } n \geq 1, \tag{31}$$

$$\frac{F_n(s)}{F_n(s-1)} = \frac{1}{2s(2n+2s+1)} \text{ for } s \geq 1. \tag{32}$$

In the low-frequency scattering method (Harlen et al. [4]), the incident field was included in the form of (14), with the $j_n(Kr)$ function retained. Thus the contribution of the incident field was included entirely in the zeroth and first order terms of the perturbation series in (iKa). The later work for the higher frequency region

(Harlen et al. [5]) expressed the incident field through the power series expansion of the exponential form:

$$e^{iKz} = e^{iKr\cos\theta} = \sum_{n=0}^{\infty} \frac{(iKa\cos\theta)^n}{n!}. \tag{33}$$

However, this requires that each of the powers of $\cos\theta$ be expressed in terms of the associated Legendre polynomials $P_n(\cos\theta)$ which appear in the other wave potentials in order to match the angular dependence in the boundary conditions. Hence the expansion used here, (28), appears to be most consistent with the method, since the powers of iKa and the angular dependence $P_n(\cos\theta)$ can be matched directly in the boundary equations.

4. Solution forms inside the particle. Within the particle or droplet, the solutions to the Helmholtz equation need not satisfy the radiation condition, since the waves are in a bounded region. However, the solutions must be defined at the origin ($r = 0$). In spherical coordinates the appropriate solutions for the radial part of the potential are the spherical Bessel functions, j_n, rather than the spherical Hankel functions, h_n, which are not defined at the origin. When the boundary conditions are applied, the Bessel function for the thermal wave must be evaluated for an argument $L'a$ which can have a large imaginary component in the frequency range of interest. Hence, it is again desirable to avoid the use of Bessel functions.

4.1. Propagational mode. There are many different ways of expressing the spherical Bessel functions j_n—for example, as an infinite power series or as a combination of trigonometric functions *sin* and *cos*. For the propagational mode inside the particle, the power series form can be used, since the value $K'a$ (which is the argument of the function used in the boundary equations) is small in the long wavelength region. First, applying the perturbation series expansion as

$$\varphi' = \sum_m (iKa)^m \varphi'_m \tag{34}$$

and then expressing the potential as a series in powers of r gives

$$\varphi'_m = \sum_{n=0}^{\infty} \sum_{j=0} A'_{nmj} \cdot \frac{r^j}{a^j} \cdot \frac{r^n}{a^n} \cdot P_n(\cos\vartheta). \tag{35}$$

The Helmholtz equation (2) can again be used with (34) to relate potentials to those of a different order, and thus

$$\nabla^2 \varphi'_m = \frac{\hat{c}}{a^2} \varphi'_{m-2}, \tag{36}$$

where

$$\hat{c} = \frac{K'^2}{K^2}, \tag{37}$$

which is frequency independent to a very good approximation. Substituting the general solution, (35), into (36) and matching powers of iKa as before results in the following recurrence relation for the coefficients:

$$A'_{n,m,j} = \frac{\hat{c}}{j(2n+j+1)} A'_{n,m-2,j-2} \text{ for } j \geq 2, \tag{38}$$

$$A'_{n,m,j} = 0 \text{ for } j = 1 \text{ and all odd values of } j. \tag{39}$$

As in the case of the propagational mode in the continuous phase, the above equations show that the coefficients $A'_{n,m,j}$ for order m can be calculated from those of a previous order, given that all coefficients for $m < 0$ are zero. Only the $j = 0$ coefficient remains to be determined from the boundary conditions. In this case the limit of the series is determined by the number of nonzero coefficients for the previous order, producing an expanding triangle of coefficients. The solution of the boundary conditions shows that the first nonzero coefficient is for $n = m$, $j = 0$, so that the limit of the series in j would be $j = m - n$, where $m > n$.

4.2. Thermal mode. For the thermal wave, the argument of the function at the boundary, $L'a$, may be small or very large, depending on the frequency. The usual power series expansion of j_n would be inappropriate at large arguments. Similarly an infinite inverse power series in $L'r$, as used by Harlen et al. [5], would be unsuitable at small arguments and is not defined at the origin. Hence, either form is restricted in its frequency range.

The trigonometric form of the Bessel function j_n (e.g., $j_0(x) = \sin x/x$) can be modified by expressing the trigonometric functions as a sum or difference of two exponential terms e^{ix} and e^{-ix}, and thus

$$j_n(x) = e^{ix} \sum_{j=1}^{n+1} \frac{j_{nj+}}{x^j} - e^{-ix} \sum_{j=1}^{n+1} \frac{j_{nj-}}{x^j} \tag{40}$$

(see the appendix for the coefficients).

Hence, the wave potential for the thermal wave in the particle could be written as a sum of modified outward and inward spherical waves. Arfken [7, p. 627] states that "$j_n(x)$ and $n_n(x)$ are appropriate for a description of *standing spherical waves*; $h_n^1(x)$ and $h_n^2(x)$ correspond to *traveling spherical waves*." A standing wave results from a superposition of traveling waves in opposite directions.

Hence the thermal wave inside the particle can be written as

$$\psi' = e^{iL'(r-a)}\tilde{\psi}'_+ - e^{-iL'(r-a)}\tilde{\psi}'_-. \tag{41}$$

Following the previous perturbation series method with each of the new wave potentials,

$$\left\{\tilde{\psi}'_+, \tilde{\psi}'_-\right\} = \sum_m (iKa)^m \left\{\tilde{\psi}'_{+m}, \tilde{\psi}'_{-m}\right\}. \tag{42}$$

And each of the terms has the usual angular dependence; thus

$$\tilde{\psi}'_{+m} = \sum_{n=0}^{\infty} \sum_{j=1}^{n+1} B'_{nm} \cdot e^{2iL'a} \cdot \frac{j_{nj+}}{(L'r)^j} \cdot P_n(\cos\vartheta), \tag{43}$$

$$\tilde{\psi}'_{-m} = \sum_{n=0}^{\infty} \sum_{j=1}^{n+1} B'_{nm} \cdot \frac{j_{nj-}}{(L'r)^j} \cdot P_n(\cos\vartheta). \tag{44}$$

The factor $e^{2iL'a}$ results from the condition that the potential be defined at the origin. All inverse powers of r must cancel at the origin, leaving only a single term from $n = 0$.

The previous work on scattering at the high-frequency end of the long wavelength limit (Harlen et al. [5]) used only the second of the two terms given in (41). When a series solution is used in powers of $L'a$ the only solution for the zeroth order term is a/r (Harlen et al. [5, equation (4.9)]), which results in a function which is not defined at the origin. The result given above avoids this problem.

5. Construction of the solution.

5.1. Pressure and temperature factors. When applying the boundary conditions, terms in the same powers of iKa must be matched, as must the angular dependence $P_n(\cos\theta)$. For consistency, any other parameters which appear in the boundary conditions must be expressed as the appropriate power of iKa. The pressure and temperature changes caused by the different waves were defined in (5). The pressure is related to the wave potential by a factor which includes the frequency and density. Since the frequency is the same for all wave forms, this factor will cancel from the relevant boundary equation. However, the thermal factors (equation (6)) have different frequency dependence for the different wave modes, and these need to be defined in relation to powers of iKa.

For the propagational modes, the thermal factors can be simplified by using the relation

$$\left|\frac{K^2}{L^2}\right| \approx \frac{\omega\sigma}{v^2} \ll 1. \tag{45}$$

The approximation is not limited in frequency range, but relies on the small value of $\omega\sigma/v^2$ which is of order 10^{-5} at 100 MHz in water at 30° C, so that

$$\Gamma_c = \frac{-iK^2(\gamma-1)}{\beta(\omega+i\gamma\sigma K^2)} \approx \frac{K^2(\gamma-1)}{\beta\sigma L^2} \tag{46}$$

and

$$\Gamma'_c = \frac{-iK'^2(\gamma'-1)}{\beta'(\omega+i\gamma'\sigma' K'^2)} \approx \frac{\hat{c}K^2(\gamma'-1)}{\beta'\sigma' L'^2}. \tag{47}$$

The dependence on L is left implicit, whereas the power series in (iKa) requires explicit consideration of powers of K. Hence the K-dependence of thermal factors are expressed by two new parameters, and thus

$$\Gamma_c = (iKa)^2 g_c, \quad \Gamma'_c = (iKa)^2 g'_c. \tag{48}$$

The thermal factors for the thermal wave modes can be simplified,

$$\Gamma_t = \frac{-iL^2(\gamma-1)}{\beta(\omega+i\gamma\sigma L^2)} \approx -\frac{1}{\beta\sigma}, \tag{49}$$

and similarly in the dispersed phase. The temperature factors for the thermal waves can therefore be seen to be approximately independent of frequency, and hence independent of K, which is the power series being used.

5.2. Definitions. The application of the boundary conditions leads to some complicated equations, which can be made easier to read by using some further symbols to define collections of terms. In addition, in numerical calculation, greater accuracy is achieved (avoiding subtraction of nearly equal terms) by using the recurrence relation (24) to write

$$\sum_{j=0} A_{n,m-1,j} + \sum_{j=0}^{m-n} jA_{nmj} = \sum_{j=1}^{m-n} -\frac{j}{(j-2n)} A_{n,m-1,j} + \delta_{n0} A_{n,m-1,0}, \tag{50}$$

where δ_{n0} is a Kronecker delta. The second term on the right-hand side which affects only the $n = 0$ results was omitted in the reported LFPST solution (Harlen et al. [4]).

Other symbols are defined as follows:

$$S_h = \sum_{j=1}^{n+1} \frac{h_{nj}}{(La)^j}, \tag{51}$$

$$S_{dh} = \sum_{j=1}^{n+1} \frac{(iLa - j)\, h_{nj}}{(La)^j}, \tag{52}$$

$$S_j = e^{2iL'a} \sum_{j=1}^{n+1} \frac{j_{nj+}}{(L'a)^j} - \sum_{j=1}^{n+1} \frac{j_{nj-}}{(L'a)^j}, \tag{53}$$

$$S_{dj} = e^{2iL'a} \sum_{j=1}^{n+1} \frac{(iL'a - j)\, j_{nj+}}{(L'a)^j} + \sum_{j=1}^{n+1} \frac{(iL'a + j)\, j_{nj-}}{(L'a)^j}, \tag{54}$$

$$S_{A,m-s} = \sum_{j=0}^{n} A_{n,m-s,j} \text{ for } s = 1, 2, 3, \quad S_{A,m} = \sum_{j=1}^{n} A_{n,m,j}, \tag{55}$$

$$S_{jA,m-s} = \sum_{j=1}^{n} -\frac{j}{(j-2n)} A_{n,m-s,j} + \delta_{n0} A_{n,m-s,0} \text{ for } s = 0, 1, 2, 3, \tag{56}$$

$$S_{A',m-s} = \sum_{j=0}^{m-n} A'_{n,m-s,j} \text{ for } s = 1, 2, 3, \quad S_{A',m} = \sum_{j=1}^{m-n} A'_{n,m,j}, \tag{57}$$

$$S_{jA',m-s} = \sum_{j=1}^{m-n} j A'_{n,m-s,j} \text{ for } s = 0, 1, 2. \tag{58}$$

5.3. Boundary conditions. Having defined the wave potentials in a consistent form, as perturbation series in powers of iKa, and with the Legendre polynomials defining the angular dependence, the boundary conditions can now be applied at the surface of the spherical particle, $r = a$. Each boundary equation consists of summations over orders n, m. The spherical harmonic terms which define the angular dependence are independent and hence must be matched—so all terms in the same n must be matched. In addition, terms in powers of $(iKa)^m$ are matched on each side of the equation, which may arise from various orders of m. If each order m is determined in turn, all coefficients for previous orders, e.g., $m-1$, are already known. In addition, the propagational mode coefficients for order m for $j \geq 1$ can be calculated from the previous order results (see (24) and (38)). Hence, the boundary equations for the n, mth order include four unknowns:

$$A_{nm0},\ A'_{nm0},\ B_{nm},\ B'_{nm}.$$

The four boundary conditions (8)–(11) result in the equations below:

$$mF_n\left(\frac{m-n}{2}\right)+\sum_{j=0}^{m-1-n}A_{n,m-1,j}+\sum_{j=1}^{m-n}(j-n-1)\,A_{nmj}-(n+1)\,A_{nm0}+S_{dh}B_{nm}$$
$$=\sum_{j=1}^{m-n}(j+n)\,A'_{nmj}+nA'_{nm0}+S_{dj}B'_{nm}, \tag{59}$$

$$F_n\left(\frac{m-n}{2}\right)+\sum_{j=1}^{m-n}A_{nmj}+A_{nm0}+S_hB_{nm}$$
$$=\hat{\rho}\sum_{j=1}^{m-n}A'_{nmj}+\hat{\rho}A'_{nm0}+\hat{\rho}S_jB'_{nm}, \tag{60}$$

$$g_cF_n\left(\frac{m-n-2}{2}\right)+g_c\sum_{j=0}^{m-2-n}A_{n,m-2,j}+\Gamma_tS_hB_{nm}$$
$$=g'_c\sum_{j=0}^{m-2-n}A'_{n,m-2,j}+\Gamma'_tS_jB'_{nm}, \tag{61}$$

$$g_c\,(m-2)\,F_n\left(\frac{m-n-2}{2}\right)+g_c\sum_{j=0}^{m-3-n}A_{n,m-3,j}$$
$$+g_c\sum_{j=0}^{m-2-n}(j-n-1)\,A_{n,m-2,j}+\Gamma_tS_{dh}B_{nm}$$
$$=\hat{\tau}g'_c\sum_{j=0}^{m-2-n}(j+n)\,A'_{n,m-2,j}+\hat{\tau}\Gamma'_tS_{dj}B'_{nm}. \tag{62}$$

5.4. Solution. The solution proceeds by stepping through the orders of m starting at $m=0$. All coefficients are zero for $m<0$. For each m the recurrence relations are used to derive any nonzero propagational mode coefficients for order m (equations (24) and (38)). The two thermal boundary conditions, (61) and (62), for the n,mth order include only the unknown thermal coefficients; other terms, being from previous orders $m-2$ and $m-3$, are already known. Hence (61) and (62) can be solved for the thermal coefficients $B_{nm}$and B'_{nm}. These can be substituted into the other boundary equations, (59) and (60), in order to determine the remaining propagational mode coefficients $A_{nm0},\ A'_{nm0}$.

The thermal coefficients for n,m are

$$B_{nm}=\Big[-g_c\,(\hat{\tau}S_{dj}-(m-2)\,S_j)\,F_n\,((m-n-2)/2)$$
$$-g_c\,\{(\hat{\tau}S_{dj}+(n+1)\,S_j)\,S_{A,m-2}-S_jS_{jA,m-3}\}$$
$$+\hat{\tau}g'_c\,\{(S_{dj}-nS_j)\,S_{A',m-2}-S_jS_{jA',m-2}\}\Big]/\Gamma_t\,(\hat{\tau}S_{dj}S_h-S_jS_{dh})\,, \tag{63}$$

$$
\begin{aligned}
B'_{nm} = \Big[& -g_c \left(S_{dh} - (m-2)\, S_h\right) F_n\left((m-n-2)/2\right) \\
& - g_c \left\{\left(S_{dh} + (n+1)\, S_h\right) S_{A,m-2} - S_h S_{jA,m-3}\right\} \\
& + g'_c \left\{\left(S_{dh} - n\hat{\tau} S_h\right) S_{A',m-2} - \hat{\tau} S_h S_{jA',m-2}\right\}\Big] / \Gamma'_t \left(\hat{\tau} S_{dj} S_h - S_j S_{dh}\right).
\end{aligned} \tag{64}
$$

Substituting into the boundary conditions, (59) and (60), thus gives

$$
\begin{aligned}
A_{nm0} = \Big[& (m\hat{\rho} - n)\, F_n\left((m-n)/2\right) \\
& - (n + (n+1)\,\hat{\rho})\, S_{A,m} + \hat{\rho}\left(S_{jA,m-1} - S_{jA',m}\right) \\
& + \left(\hat{\rho} S_{dh} - n S_h\right) B_{nm} - \hat{\rho}\left(S_{dj} - n S_j\right) B'_{nm}\Big] / \left[n + (n+1)\,\hat{\rho}\right],
\end{aligned} \tag{65}
$$

$$
\begin{aligned}
A'_{nm0} = \Big[& (m+n+1)\, F_n\left((m-n)/2\right) \\
& - (n + (n+1)\,\hat{\rho})\, S_{A',m} + \left(S_{jA,m-1} - S_{jA',m}\right) \\
& + \left(S_{dh} + (n+1)\, S_h\right) B_{nm} - \left(S_{dj} + (n+1)\,\hat{\rho} S_j\right) B'_{nm}\Big] / \left[n + (n+1)\,\hat{\rho}\right].
\end{aligned} \tag{66}
$$

Since the thermal boundary equations (61) and (62) include terms in order $m-2$ and lower orders, this implies that the first nonzero thermal field contribution is at second order in iKa.

5.5. Multiple scattering. For practical application of the scattering results, it is necessary to relate the scattering from a single particle to the wavenumber (and corresponding velocity and attenuation) of a dispersion of such particles. The commonly used formulation for multiple scattering is that of Lloyd and Berry [8], whose result was derived in a different way by Waterman and Truell and later works [9], [10]. Taking the limiting form of the solution, (18), in the far field as r approaches infinity, the scattered field takes the form

$$
\varphi \to \frac{e^{iKr}}{r} f(\theta). \tag{67}
$$

In terms of the Legendre polynomials,

$$
f(\theta) = \frac{1}{iK} \sum_{n=0}^{\infty} (2n+1)\, T_n P_n(\cos\theta). \tag{68}
$$

The scattered propagational field, combining (19), (21), and (22), has the form

$$
\varphi = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \sum_{j=0} A_{nmj} (iKa)^m\, e^{iK(r-a)} \cdot \frac{r^j}{a^j} \cdot \frac{a^{n+1}}{r^{n+1}} \cdot P_n(\cos\vartheta), \tag{69}
$$

from which the far field coefficient T_n is

$$
T_n = \frac{e^{-iKa}}{(2n+1)} \sum_{m=0}^{\infty} (iKa)^{m+1} A_{nmn}, \tag{70}
$$

showing that only the terms $j = n$ contribute in the far field. For each spherical harmonic n (except for $n = 0$), the first nonzero term for the $j = 0$ coefficients (from the boundary equations) is for $m = n$ (see, for example, the results in Table 1).

Following the recurrence relation (equation (24)) through the orders m shows that the first nonzero term appearing in the far field (i.e., for $j = n$) will be at order $m = 2n$. So for the $n = 1$ far field coefficient, T_1, the first contribution results from A_{121}; for $n = 2$ the first nonzero term in T_2 corresponds to $m = 4$, i.e., A_{242}. It is very important to include sufficient orders $m \geq 2n$ when an accurate solution for the nth order far field coefficient is required.

The multiple scattering result for the wavenumber of the dispersion, B, is

$$\left(\frac{B}{K}\right)^2 = 1 + \frac{3\phi}{K^2a^3} f(0) + \frac{9\phi^2}{4K^4a^6}\left(f^2(\pi) - f^2(0) - \int_0^\pi d\theta \frac{1}{\sin(\theta/2)}\left(\frac{d}{d\theta} f^2(\theta)\right)\right), \tag{71}$$

which to second order gives

$$\left(\frac{B}{K}\right)^2 = 1 - \frac{3i\phi}{K^3a^3}(T_0 + 3T_1 + 5T_2) - \frac{27\phi^2}{K^6a^6}\left(T_0T_1 + \frac{10}{3}T_0T_2 + 2T_1^2 + 11T_1T_2 + \frac{230}{21}T_2^2\right). \tag{72}$$

Note that here the symbol ϕ refers to the volume fraction of the dispersed particles.

5.6. Explicit solutions for low orders. In order to demonstrate the method of solution and to derive explicit solutions which may be used instead of the general solution, the results are here derived for low orders of n and m. The following parameters are those which are used to obtain the low order solutions:

For $n = 0$

$$h_{01} = -i, \tag{73}$$

$$S_h = -i/(La), \quad S_{dh} = -i(iLa - 1)/(La), \tag{74}$$

$$j_{01+} = j_{01-} = 1/(2i), \tag{75}$$

$$S_j = \left(e^{2iL'a} - 1\right)\Big/(2iL'a), \tag{76}$$

$$S_{dj} = \left\{iL'a\left(e^{2iL'a} + 1\right) - \left(e^{2iL'a} - 1\right)\right\}\Big/(2iL'a), \tag{77}$$

$$F_0(0) = 1, \quad F_0(1) = 1/6, \tag{78}$$

and for $n = 1$

$$F_1(0) = 1. \tag{79}$$

All the nonzero coefficients are given in Table 1 for orders $n \leq 2$ and $m \leq 2$. The method for obtaining these solutions is summarized below.

TABLE 1
Explicit solutions for scattering coefficients at low orders.

	$n=0$	$n=1$	$n=2$
$m=0$	$A'_{000} = 1/\hat{\rho}$		
$m=1$		$A_{110} = \dfrac{(\hat{\rho}-1)}{(2\hat{\rho}+1)}$ $A'_{110} = \dfrac{3}{(2\hat{\rho}+1)}$	
$m=2$	$A'_{022} = \dfrac{\hat{c}}{6\hat{\rho}}$ $B_{02} = \dfrac{\hat{\tau}\,(g'_c - \hat{\rho} g_c)\,S_{dj}}{\hat{\rho}\Gamma_t\left(\hat{\tau} S_{dj} S_h - S_j S_{dh}\right)}$ $B'_{02} = \dfrac{(g'_c - \hat{\rho} g_c)\,S_{dh}}{\hat{\rho}\Gamma'_t\left(\hat{\tau} S_{dj} S_h - S_j S_{dh}\right)}$	$A_{120} = \dfrac{(\hat{\rho}-1)}{(2\hat{\rho}+1)}$ $A_{121} = -\dfrac{(\hat{\rho}-1)}{(2\hat{\rho}+1)}$	$A_{220} = \dfrac{2\,(\hat{\rho}-1)}{3\,(3\hat{\rho}+2)}$ $A'_{220} = \dfrac{5}{3\,(3\hat{\rho}+2)}$
	$A_{020} = \dfrac{(\hat{\rho}-\hat{c})}{3\hat{\rho}} + \dfrac{(g'_c - \hat{\rho} g_c)\left(\hat{\tau}\Gamma'_t - \Gamma_t\right) S_{dj} S_{dh}}{\hat{\rho}\Gamma_t\Gamma'_t\left(\hat{\tau} S_{dj} S_h - S_j S_{dh}\right)}$ $A'_{020} = \dfrac{1}{2\hat{\rho}} - \dfrac{\hat{c}\,(\hat{\rho}+2)}{6\hat{\rho}^2}$ $+\dfrac{(g'_c - \hat{\rho} g_c)\left[\hat{\tau}\Gamma'_t S_{dj}\,(S_{dh}+S_h) - \Gamma_t S_{dh}\left(S_{dj} + \hat{\rho} S_j\right)\right]}{\hat{\rho}^2\Gamma_t\Gamma'_t\left(\hat{\tau} S_{dj} S_h - S_j S_{dh}\right)}$		

By definition, all coefficients for orders $m < 0$ are zero; hence the recurrence relations (equations (24) and (38)) show that all propagational mode coefficients are zero for $j > 0$. There is no incident field contribution to the thermal boundary conditions (equations (61) and (62)) (since $s = (m-n-2)/2 = -1$), so the thermal coefficients are zero:

$$B_{00} = B'_{00} = 0.$$

The velocity and pressure boundary conditions (equations (59) and (60)) include a nonzero contribution from the incident field, such that the zeroth order of the incident field affects the zeroth order propagational mode. The resulting coefficients for the propagational modes are

$$A_{000} = 0, \quad A'_{000} = 1/\hat{\rho}. \tag{80}$$

For $m = 1$ the incident field makes no contribution at the boundary (since the arguments $s = (m-n-2)/2$ or $s = (m-n)/2$ are noninteger). No nonzero coefficients are found from the recurrence relations, so again the thermal field is zero. The propagational coefficients are also zero in this case (by substitution in the boundary equations):

$$A_{010} = A'_{010} = 0. \tag{81}$$

For $m = 2$, the recurrence relations now give a nonzero coefficient, A'_{022}, resulting from the A'_{000} coefficients, as shown in the table.

The summations over coefficients also now have nonzero terms:

$$S_{A',m} = \frac{\hat{c}}{6\hat{\rho}}, \quad S_{jA',m} = \frac{\hat{c}}{3\hat{\rho}}, \quad S_{A',m-2} = \frac{1}{\hat{\rho}}. \tag{82}$$

All coefficients ($j > 0$) of the continuous phase propagational mode are again zero.

The incident field contributes both to the thermal boundary conditions and to the velocity and pressure conditions producing the first nonzero thermal coefficients, B_{02}, B'_{02} (Table 1). These results are consistent with the results in the geometric theory paper (Harlen et al. [5]) for large values of La. The LFPST paper (Harlen et al. [4]) did not assign the incident field to the different orders of Ka, instead assigning it all to the zeroth and first order terms, so analytical comparison is not possible. The propagational coefficients, A_{020}, A'_{020} (Table 1) are found by substituting the thermal coefficients into the velocity and pressure boundary equations. For higher orders in n the process is exactly the same, but there are no thermal contributions up to order $m = 2$, resulting in coefficients which depend only on density.

To obtain the velocity and attenuation of a dispersion using the coefficients up to second order, the relevant far field coefficients, T_n, must be determined. Equation (70) shows that only coefficients for $j = n$ contribute to each T_n. Thus the far field coefficients to second order in m are

$$T_0 = e^{-iKa} (iKa)^3 A_{020}, \quad T_1 = \frac{e^{-iKa}}{3} (iKa)^3 A_{121}, \quad T_2 = 0. \tag{83}$$

Substituting these coefficients into the equation for the wavenumber of the dispersion, B (equation (72)), gives

$$\left(\frac{B}{K}\right)^2 = 1 - 3\phi e^{-iKa} (A_{020} + A_{121}) + 3\phi^2 e^{-2iKa} \left(3A_{020}A_{121} + 2A_{121}^2\right). \tag{84}$$

Since the exponential factors are near unity (Ka being small), and the coefficient A_{121} depends only on density, the particle size and frequency dependence appear almost entirely through the A_{020} coefficient. It is the parameters S_h, S_{dh}, S_j, S_{dj} (equations (74) and (76)) which define the dependence on particle size and frequency through the relationship between the thermal wavelengths and the particle size, expressed by the parameters La and $L'a$. Thus we have an analytical result for the wavenumber of the dispersion which is valid over the entire long wavelength region, and is simple enough to be calculated in a standard spreadsheet. It is, of course, an approximate result, but, as will be demonstrated in the next section, it is a good approximation unless the parameter $Ka > 0.01$. Visco-inertial scattering has not been included in the present theory, so for dispersions with a large density difference between the two components, the results will not be as accurate.

6. Results. Calculations have been carried out using MATLAB for a model system of sunflower oil in water at 30° C. The calculations are straightforward, and take only a few seconds to complete a spectrum of 50 frequency values. The physical properties of the two components are given in Table 2. A particle diameter of 1 μm was chosen so that a complete range of thermal wavelengths could be covered within the long wavelength limit. The concentration (by volume) was 20%. The attenuation of each material was not included in the calculation, so the attenuation determined by

TABLE 2
Physical properties of sunflower oil and water at 30° C.

	Water	Sunflower oil
Ultrasound velocity / m s^{-1}	1509.1	1437.9
Density / kg m^{-3}	995.7	912.9
Thermal expansivity / K^{-1}	0.00030	0.00073
Specific heat capacity / J kg^{-1} K^{-1}	4178.2	1980.0
Thermal conductivity / W m^{-1} K^{-1}	0.603	0.17

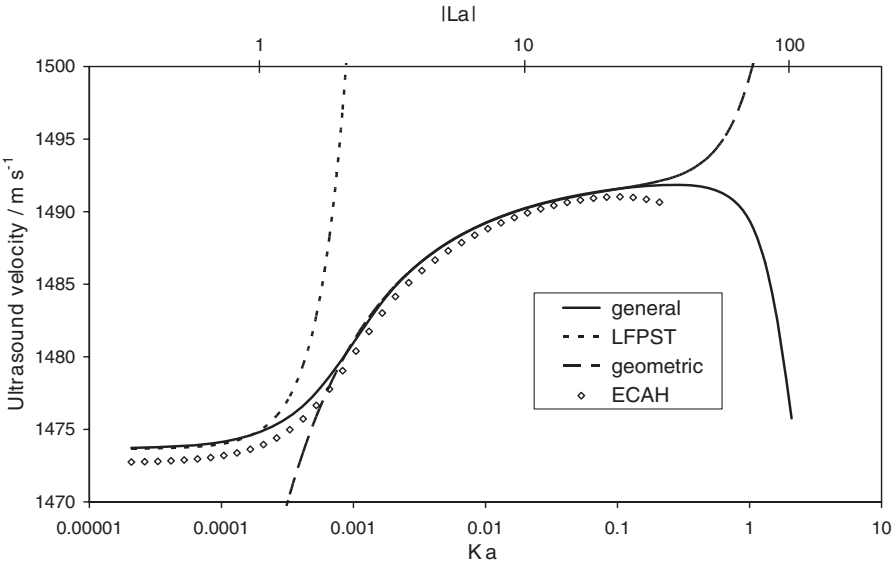


FIG. 1. *Ultrasound velocity as a function of the parameter Ka for* 20% *sunflower oil in water at* 30° *C with a particle diameter of* 1 μm. *Four different calculation methods are shown: the "general" method presented in this paper, the LFPST method (Harlen et al.* [4]*), the geometric theory method (Harlen et al.* [5]*), and the ECAH method (Epstein and Carhart* [2]*). The thermal parameter* $|La|$ *is shown above the plot. The parameters for the dispersed phase are* $|L'a| = 1.24\,|La|$ *and* $K'a = 1.05Ka$.

the scattering calculation is in addition to the nonscattering contribution. Figures 1 and 2 show the velocity and attenuation per wavelength ($\alpha\lambda$) as a function of frequency in the form of the parameters Ka and La. The ultrasound properties have been calculated by four different methods, including ECAH and the general theory results presented here.

It was found that the LFPST theory (Harlen et al. [4]) for low frequencies required orders up to $m = 8$ in order to obtain even the first part of the change in velocity and attenuation as the frequency increases. The theory is very much confined to the lowest frequency range. Similarly the geometric theory (Harlen et al. [5]) (which here was calculated only to second order in m and n) is valid only for high frequencies within the long wavelength region and deviates from the ECAH values as the frequency decreases.

The general theory presented in the current work is valid over the entire frequency range within the long wavelength region. The results for velocity and attenuation match closely those determined using the ECAH method. Visco-inertial scattering,

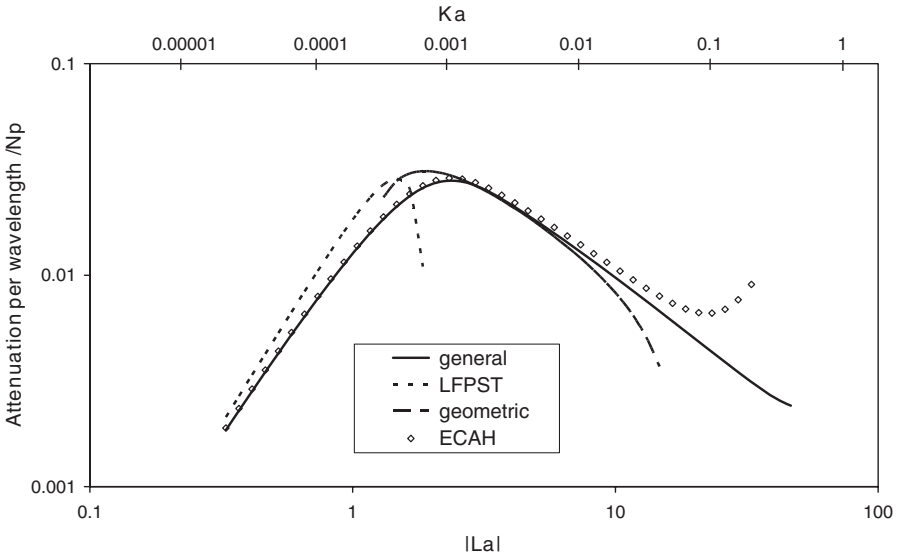


FIG. 2. *Attenuation per wavelength as a function of the parameter* $|La|$ *for* 20% *sunflower oil in water at* 30° *C. Four different calculation methods are shown: the "general" method presented in this paper, the LFPST method (Harlen et al.* [4]*), the geometric theory method (Harlen et al.* [5]*), and the ECAH method (Epstein and Carhart* [2]*). Corresponding values of Ka are shown above the plot. The parameters for the dispersed phase are* $|L'a| = 1.24\,|La|$ *and* $K'a = 1.05Ka$.

which relates to a difference in density between the two components and their viscosity, is not included in the general theory but is included in ECAH. The densities of the two components are similar, but the additional scattering accounts for the difference between the general theory and ECAH results. Terms for $m \leq 4$ were included for the general theory, as higher orders did not significantly change the result. At higher frequencies, the long wavelength criterion ($Ka \ll 1$) is no longer valid, and more and more terms are needed to obtain an accurate result. Figure 3 shows the contribution of including these terms. The second order solution is very accurate for frequencies below the point $Ka = 0.007$. The fourth order solution gives a more accurate result over a much wider frequency range.

7. Conclusions. A method has been presented for the solution of the ultrasound scattering problem in the long wavelength region. The work builds on previously published studies which covered only part of the frequency range, when the thermal wavelength is either much smaller or much larger than the particle size. The technique consists of expressing the scattered fields as perturbation series in the parameter Ka, which is always small in the long wavelength region, and explicitly removing the radiating field factor e^{iKr}. A result has been obtained which covers the complete long wavelength region. The calculation is much more straightforward than the widely used ECAH method, which relies on spherical harmonic expansions and suffers from numerical instability. A simplified analytical version of the result has been produced which enables calculation in an ordinary spreadsheet.

8. Appendix. The thermal wave solutions use expansions of the spherical Bessel functions which are not generally found in mathematical texts. The coefficients can be calculated in the order n by the formula below.

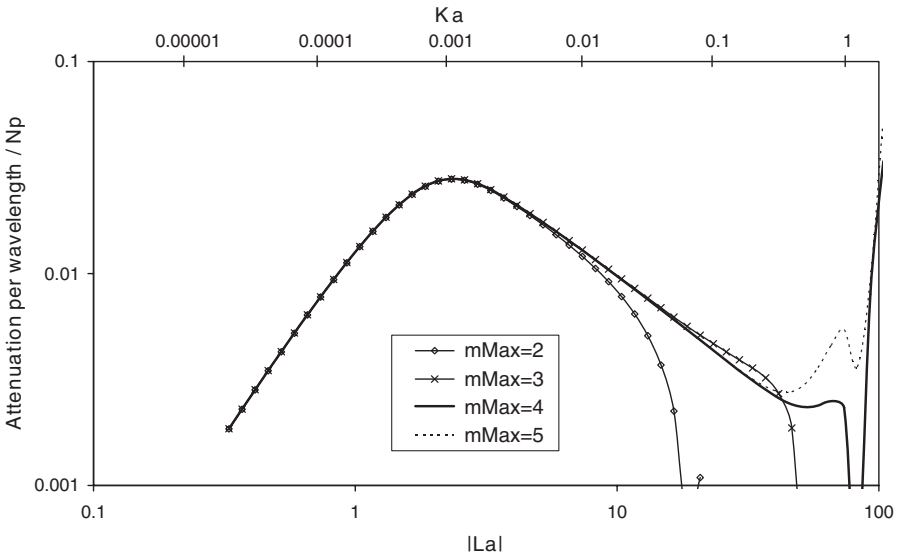


FIG. 3. *General theory results for* 20% *sunflower oil in water at* 30° *C showing the contributions of higher order terms in the series expansion, especially at larger values of* Ka. *The parameters for the dispersed phase are* $|L'a| = 1.24\,|La|$ *and* $K'a = 1.05Ka$.

For the spherical Hankel function,

$$h_{01} = -i \text{ for } n = 0,$$

$$\text{for } n > 0 \quad h_{n,j} = \begin{cases} -ih_{n-1,j} \text{ for } j = 1, \\ -ih_{n-1,j} + (n + j - 2)h_{n-1,j-1} \text{ for } 1 < j < n + 1, \\ (2n - 1)h_{n-1,j-1} \text{ for } j = n + 1. \end{cases}$$

Similarly for the spherical Bessel function, which was defined in two parts, an outgoing and an ingoing traveling wave:

$$j_{0,1+} = 1/2\text{i for } n = 0,$$

$$\text{for } n > 0 \quad j_{n,j+} = \begin{cases} -ij_{n-1,j+} \text{ for } j = 1, \\ -ij_{n-1,j+} + (n + j - 2)j_{n-1,j-1+} \text{ for } 1 < j < n + 1, \\ (2n - 1)j_{n-1,j-1+} \text{ for } j = n + 1, \end{cases}$$

$$j_{0,1-} = 1/2\text{i for } n = 0,$$

$$\text{for } n > 0 \quad j_{n,j-} = \begin{cases} ij_{n-1,j-} \text{ for } j = 1, \\ ij_{n-1,j-} + (n + j - 2)j_{n-1,j-1-} \text{ for } 1 < j < n + 1, \\ (2n - 1)j_{n-1,j-1-} \text{ for } j = n + 1. \end{cases}$$

REFERENCES

[1] J. W. STRUTT (BARON RAYLEIGH), *The Theory of Sound*, 2nd ed., Macmillan, London, 1896.
[2] P. S. EPSTEIN AND R. R. CARHART, *The absorption of sound in suspensions and emulsions.* I. *Water fog in air*, J. Acoust. Soc. Amer., 25 (1953), pp. 553–565.

[3] J. R. Allegra and S. A. Hawley, *Attenuation of sound in suspensions and emulsions: Theory and experiments*, J. Acoust. Soc. Amer., 51 (1972), pp. 1545–1564.
[4] O. G. Harlen, M. J. Holmes, M. J. W. Povey, Y. Qiu, and B. D. Sleeman, *A low frequency potential scattering description of acoustic propagation in dispersions*, SIAM J. Appl. Math., 61 (2001), pp. 1906–1931.
[5] O. G. Harlen, M. J. Holmes, M. J. W. Povey, and B. D. Sleeman, *Acoustic propagation in dispersions and the geometric theory of diffraction*, SIAM J. Appl. Math., 63 (2003), pp. 834–849.
[6] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer, Berlin, 1998.
[7] G. Arfken, *Mathematical Methods for Physicists*, 3rd ed., Academic Press, Orlando, FL, 1985.
[8] P. Lloyd and M. V. Berry, *Wave propagation through an assembly of spheres.* IV. *Relations between different multiple scattering theories*, Proc. Phys. Soc., 91 (1967), pp. 678–688.
[9] P. C. Waterman and R. Truell, *Multiple scattering of waves*, J. Math. Phys., 2 (1961), pp. 512–537.
[10] J. G. Fikioris and P. C. Waterman, *Multiple scattering of waves.* II. *"Hole corrections" in the scalar case*, J. Math. Phys., 5 (1964), pp. 1413–1420.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 510–532

THE INITIAL SURFACE TENSION–DRIVEN FLOW OF A WEDGE OF VISCOUS FLUID*

J. BILLINGHAM†

Abstract. In this paper, we consider the two-dimensional motion of a viscous, incompressible fluid with a free surface, initially lying inside a wedge. The fluid flows under the action of surface tension, and we analyze its small time motion using the method of matched asymptotic expansions. We show that, in contrast to the case where there is a surrounding fluid with viscosity [M. J. Miksis and J.-M. Vanden-Broeck, *Phys. Fluids*, 11 (1999), pp. 3227–3231], the initial motion is not self-similar but develops over two asymptotic regions: an inner, nonlinear, surface tension–driven Stokes flow region near the tip of the wedge, and an outer, linear, unsteady Stokes flow region, where inertia is important but surface tension is not. The initial velocity of the tip of the wedge is singular, of $O(\log t)$ as $t \to 0$. We calculate numerical solutions of both the inner and outer problem for a general wedge semiangle, α, and also construct asymptotic solutions in the limits $\alpha \to 0$ and $\alpha \to \pi$.

Key words. fluid mechanics, surface tension, matched asymptotic expansions

AMS subject classifications. 35C20, 76D45

DOI. 10.1137/050622419

1. Introduction.

In the time since Keller and Miksis [2] identified a similarity scaling, with lengths scaling like $t^{2/3}$, suitable for the solution of two- and three-dimensional inviscid, surface tension–driven flows in initial configurations with no geometrical lengthscale (for example, wedges and cones), many authors have studied related problems (for example, [3], [4], [5], [6], [7]). Such problems are of relevance to the recoil of fluid sheets and jets after rupture. However, at sufficiently small times, viscosity is always relevant in this type of problem. In order to investigate this, Miksis and Vanden-Broeck [1] studied the two-dimensional, surface tension–driven Stokes flow of two fluids initially lying in four wedges. This canonical problem is relevant to the rupture of liquid sheets (see, for example, [8]), and a similarity scaling is available, with lengths scaling like t. Miksis and Vanden-Broeck [1] solved this similarity problem numerically using the boundary integral method. However, they were unable to find a solution for the case of an inviscid outer fluid and hypothesized that no such solution exists.

In this paper, we consider the problem of the recoil of a single wedge of viscous, incompressible fluid under the action of surface tension. Noting that the solutions studied by Miksis and Vanden-Broeck [1] give the small time asymptotic behavior of the two-fluid problem, we study the small time solution of the single-fluid problem. Drawing on the results given in [4] for the case of an almost flat wedge, we show that inertia is never negligible, and the solution has a two-region asymptotic structure, consistent with the hypothesis of [1] that no similarity solution exists. In an outer region, with size $O(t^{1/2})$, surface tension is negligible and linearized inertia acts at leading order, while in an inner region, with size $O(t)$, inertia is negligible, and the leading order problem is a surface tension–driven Stokes flow, similar to that studied

*Received by the editors January 11, 2005; accepted for publication (in revised form) June 20, 2005; published electronically December 30, 2005. This work was supported by the Engineering and Physical Sciences Research Council through an Advanced Research Fellowship.

http://www.siam.org/journals/siap/66-2/62241.html

†School of Mathematical Sciences, The University of Nottingham, University Park, Nottingham NG7 2RD, UK (John.Billingham@Nottingham.ac.uk).

by Miksis and Vanden-Broeck [1]. The crucial difference is that the tip of the wedge recoils through a distance of $O(t \log t)$ in the single fluid problem that we study here, so that the fluid velocity is singular of $O(\log t)$ as $t \to 0$. The recoil distance is fixed by matching between the inner and outer solutions.

After setting up the initial/boundary value problem in section 2, we detail the asymptotic structure and matching conditions in section 3. We discuss numerical solutions of the outer problem in section 4 and of the inner problem in section 5. Finally, we consider the limiting cases of slender wedges, first of the interior, viscous fluid in section 6, and then of the exterior, inviscid fluid in section 7.

2. The initial/boundary value problem. The dimensionless initial/boundary value problem appropriate to the surface tension–driven recoil of a single wedge of viscous fluid consists of the Navier–Stokes equations for an incompressible, viscous fluid,

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = -\nabla p + \nabla^2 \mathbf{u} \quad \text{for } y > Y(x,t), \tag{2.1}$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{for } y > Y(x,t) \tag{2.2}$$

subject to the kinematic condition,

$$\frac{\partial Y}{\partial t} = u_y - u_x \frac{\partial Y}{\partial x} \quad \text{at } y = Y(x,t), \tag{2.3}$$

and the shear and normal stress continuity conditions,

$$\left\{1 - \left(\frac{\partial Y}{\partial x}\right)^2\right\}\left(\frac{\partial u_x}{\partial y} + \frac{\partial u_y}{\partial x}\right) - 4\frac{\partial Y}{\partial x}\frac{\partial u_x}{\partial x} = 0 \quad \text{at } y = Y(x,t), \tag{2.4}$$

$$\begin{aligned} -p - 2\left\{1 + \left(\frac{\partial Y}{\partial x}\right)^2\right\}^{-1} &\left[\left\{1 - \left(\frac{\partial Y}{\partial x}\right)^2\right\}\frac{\partial u_x}{\partial x} + \frac{\partial Y}{\partial x}\left(\frac{\partial u_x}{\partial y} + \frac{\partial u_y}{\partial x}\right)\right] \\ &= -\frac{\partial^2 Y}{\partial x^2}\left\{1 + \left(\frac{\partial Y}{\partial x}\right)^2\right\}^{-3/2} \quad \text{at } y = Y(x,t), \end{aligned} \tag{2.5}$$

where $\mathbf{u} = (u_x, u_y)$ is the velocity field, p is the pressure, and the free surface lies at $y = Y(x,t)$. The initial conditions are

$$\mathbf{u} = 0, \quad Y = \cot \alpha \, x \quad \text{when } t = 0, \tag{2.6}$$

so that the fluid is initially at rest in a wedge of semiangle α, and the far field conditions are

$$\mathbf{u} \to \mathbf{0}, \quad p \to 0 \quad \text{as } x^2 + y^2 \to \infty, \tag{2.7}$$

$$Y - \cot \alpha \, x \to 0 \quad \text{as } x \to \infty. \tag{2.8}$$

The problem has been made dimensionless using the length, time, velocity, and pressure scales,

$$l^* = \mu^2/\sigma\rho, \quad t^* = \mu^3/\sigma^2\rho, \quad u^* = \sigma/\mu, \quad p^* = \rho\sigma^2/\mu^2,$$

where μ, σ, and ρ are the constant viscosity, surface tension, and density of the fluid, respectively. These are the only physical quantities available, since the initial conditions provide no geometrical lengthscale. Note that (2.1) to (2.8) contain just one parameter: the wedge semiangle, α.

3. Asymptotic solution for $t \ll 1$: Structure and matching. We will now construct the small time asymptotic solution of this problem. This has already been done by Miksis and Vanden-Broeck [1] for the equivalent problem with an outer viscous fluid, where a similarity solution of the Stokes flow limit, with lengths scaling on t, provides the solution for $t \ll 1$ of the full Navier–Stokes problem. However, as we noted earlier, no such solution can be found when the viscosity of the exterior fluid is zero. The form of the asymptotic solution of (2.1) to (2.8) when $t \ll 1$ and $|\alpha - \pi/2| \ll 1$ is described in [4], where we found that the solution develops over two asymptotic regions: an inner, surface tension–driven Stokes flow region, where lengths scale with t, and an outer, unsteady Stokes flow region with no surface tension at leading order, where lengths scale with $t^{1/2}$. In addition, the tip of the wedge recoils through a distance of $O(t \log t)$, not simply of $O(t)$, as was the case for a viscous exterior fluid. An explanation for the presence of this logarithm is given by an argument equivalent to one made in [9], which analyzes the coalescence of viscous liquid drops. In [9, section 2], it is shown that the velocity field due to a point force at the tip of a wedge with free boundaries is singular at the tip, and that the velocity there must depend upon the logarithm of the size of an inner region, consistent with the scalings we use here.

The following are appropriate asymptotic scalings:

1. Outer region: $x,\ y = O(t^{1/2}),\ Y = \cot\alpha\, x + O(t),\ p = O(t^{-1/2}),\ u_x,\ u_y = O(1)$.

2. Inner region: $x = O(t),\ y = -Kt\log t + O(t),\ Y = -Kt\log t + O(t),\ p = O(t^{-1}),\ u_x = O(1),\ u_y = -K\log t + O(1)$.

The function $K \equiv K(\alpha)$ is an eigenvalue that must be determined as part of the solution for each value of α.

3.1. Leading order equations and matching conditions. In the outer region, we define the scaled variables

$$x = t^{1/2}\tilde{x},\quad y = t^{1/2}\tilde{y},\quad Y = \cot\alpha x + t\tilde{Y},\quad p = t^{-1/2}\tilde{p},\quad u_x = \tilde{u}_x,\quad u_y = \tilde{u}_y,$$

in terms of which, at leading order for $t \ll 1$, (2.1) to (2.8) become

$$-\frac{1}{2}\tilde{\mathbf{x}}\cdot\tilde{\nabla}\tilde{\mathbf{u}} = -\tilde{\nabla}\tilde{p} + \tilde{\nabla}^2\tilde{\mathbf{u}} \quad \text{for } \tilde{y} > \cot\alpha\,\tilde{x}, \tag{3.1}$$

$$\tilde{\nabla}\cdot\tilde{\mathbf{u}} = 0 \quad \text{for } \tilde{y} > \cot\alpha\,\tilde{x}, \tag{3.2}$$

$$\tilde{Y} - \frac{1}{2}\tilde{x}\frac{d\tilde{Y}}{d\tilde{x}} = \tilde{u}_y - \cot\alpha\,\tilde{u}_x \quad \text{at } \tilde{y} = \cot\alpha\,\tilde{x}, \tag{3.3}$$

$$\left(1 - \cot^2\alpha\right)\left(\frac{\partial\tilde{u}_x}{\partial\tilde{y}} + \frac{\partial\tilde{u}_y}{\partial\tilde{x}}\right) - 4\cot\alpha\,\frac{\partial\tilde{u}_x}{\partial\tilde{x}} = 0 \quad \text{at } \tilde{y} = \cot\alpha\,\tilde{x}, \tag{3.4}$$

$$-\tilde{p} + 2\cos 2\alpha\,\frac{\partial\tilde{u}_x}{\partial\tilde{x}} - \sin 2\alpha\,\left(\frac{\partial\tilde{u}_x}{\partial\tilde{y}} + \frac{\partial\tilde{u}_y}{\partial\tilde{x}}\right) = 0 \quad \text{at } \tilde{y} = \cot\alpha\,\tilde{x}, \tag{3.5}$$

$$\tilde{\mathbf{u}} \to \mathbf{0},\quad \tilde{p} \to 0 \quad \text{as } \tilde{x}^2 + \tilde{y}^2 \to \infty, \tag{3.6}$$

$$\tilde{Y} \to 0 \quad \text{as } \tilde{x} \to \infty. \tag{3.7}$$

These are the equations for unsteady Stokes flow in a wedge, with stress-free boundary conditions. Surface tension does not appear at leading order. The kinematic condition, (3.3), decouples from the other equations and allows $\tilde{Y}$ to be calculated once the velocity field is known. Note that uniform flow in the $\tilde{y}$-direction is a simple solution of these equations, although it does not satisfy the far field boundary condition. The system is forced by the matching conditions as $\tilde{x}^2 + \tilde{y}^2 \to 0$, and we shall determine these below.

In the inner region, we define the scaled variables

$$x = t\bar{x}, \quad y = -Kt\log t + t\bar{y}, \quad Y = -Kt\log t + t\bar{Y}, \quad p = t^{-1}\bar{p},$$

$$u_x = \bar{u}_x, \quad u_y = -K\log t + \bar{u}_y,$$

in terms of which, at leading order for $t \ll 1$, (2.1) to (2.8) become

$$0 = -\bar{\nabla}\bar{p} + \bar{\nabla}^2\bar{\mathbf{u}} \quad \text{for } \bar{y} > \bar{Y}(\bar{x}), \tag{3.8}$$

$$\bar{\nabla}\cdot\bar{\mathbf{u}} = 0 \quad \text{for } \bar{y} > \bar{Y}(\bar{x}), \tag{3.9}$$

$$\bar{Y} - \bar{x}\frac{d\bar{Y}}{d\bar{x}} = K + \bar{u}_y - \bar{u}_x\frac{d\bar{Y}}{d\bar{x}} \quad \text{at } \bar{y} = \bar{Y}(\bar{x}), \tag{3.10}$$

$$\left\{1 - \left(\frac{d\bar{Y}}{d\bar{x}}\right)^2\right\}\left(\frac{\partial\bar{u}_x}{\partial\bar{y}} + \frac{\partial\bar{u}_y}{\partial\bar{x}}\right) - 4\frac{d\bar{Y}}{d\bar{x}}\frac{\partial\bar{u}_x}{\partial\bar{x}} = 0 \quad \text{at } \bar{y} = \bar{Y}(\bar{x}), \tag{3.11}$$

$$\begin{aligned} -\bar{p} - 2\left\{1 + \left(\frac{d\bar{Y}}{d\bar{x}}\right)^2\right\}^{-1}&\left[\left\{1 - \left(\frac{d\bar{Y}}{d\bar{x}}\right)^2\right\}\frac{\partial\bar{u}_x}{\partial\bar{x}} + \frac{d\bar{Y}}{d\bar{x}}\left(\frac{\partial\bar{u}_x}{\partial\bar{y}} + \frac{\partial\bar{u}_y}{\partial\bar{x}}\right)\right] \\ &= -\frac{d^2\bar{Y}}{d\bar{x}^2}\left\{1 + \left(\frac{d\bar{Y}}{d\bar{x}}\right)^2\right\}^{-3/2} \quad \text{at } \bar{y} = \bar{Y}(\bar{x}). \end{aligned} \tag{3.12}$$

These are the equations for steady, surface tension–driven Stokes flow, but with the modified kinematic condition (3.10). We now need to consider the far field behavior of solutions of this inner problem so that we can determine appropriate matching conditions between the two asymptotic regions.

Since we must have $\bar{Y} \sim \cot\alpha\bar{x}$ as $\bar{x} \to \infty$, the flow in the far field lies within a wedge at leading order. By writing the equations in terms of polar coordinates, it is straightforward to show that

$$\bar{Y} \sim \cot\alpha\bar{x} - 2K\log\bar{x} + K\left(b_\infty - 3 + 2\log\sin\alpha\right) \quad \text{as } \bar{x} \to \infty, \tag{3.13}$$

$$\begin{aligned} \bar{p} &\sim 4K\frac{\bar{y}}{\bar{x}^2 + \bar{y}^2}, \quad \bar{u}_x \sim 2K\frac{\bar{x}\bar{y}}{\bar{x}^2 + \bar{y}^2}, \\ \bar{u}_y &\sim -K\log\left(\bar{x}^2 + \bar{y}^2\right) - 2K\frac{\bar{x}^2}{\bar{x}^2 + \bar{y}^2} + Kb_\infty \quad \text{as } \bar{x}^2 + \bar{y}^2 \to \infty. \end{aligned} \tag{3.14}$$

The function $b_\infty(\alpha)$ is, as yet, undetermined.

We can now write the matching condition for the outer problem,

$$\tilde{p} \sim 4K\frac{\tilde{y}}{\tilde{x}^2+\tilde{y}^2}, \quad \tilde{u}_x \sim 2K\frac{\tilde{x}\tilde{y}}{\tilde{x}^2+\tilde{y}^2},$$

$$\tilde{u}_y \sim -K\log\left(\tilde{x}^2+\tilde{y}^2\right) - 2K\frac{\tilde{x}^2}{\tilde{x}^2+\tilde{y}^2} + Kb_\infty \quad \text{as } \tilde{x}^2+\tilde{y}^2 \to 0. \tag{3.15}$$

4. Solution of the outer problem. Since the outer problem is linear and forced by the matching condition (3.15), we can scale the constant K out of the problem and solve to determine b_∞. Once we know b_∞, we have enough information to solve the inner problem. It is also convenient to rewrite the outer problem in streamfunction-vorticity form, so we define the scaled streamfunction ψ and vorticity ω using

$$\tilde{u}_x = K\frac{\partial\psi}{\partial\tilde{y}}, \quad \tilde{u}_y = -K\frac{\partial\psi}{\partial\tilde{x}}, \quad \omega = -\tilde{\nabla}^2\psi.$$

The outer problem then becomes, using the symmetry of the problem about the $\tilde{y}$-axis, in terms of polar coordinates

$$\tilde{\nabla}^2\psi + \omega = 0 \quad \text{for } \pi/2 < \theta < \pi/2+\alpha, \tag{4.1}$$

$$\tilde{\nabla}^2\omega + \frac{1}{2}\frac{\partial}{\partial\tilde{r}}\left(\tilde{r}\omega\right) = 0 \quad \text{for } \pi/2 < \theta < \pi/2+\alpha \tag{4.2}$$

subject to

$$\omega + \frac{\partial^2\psi}{\partial\tilde{r}^2} = 0 \quad \text{at } \theta = \pi/2+\alpha, \tag{4.3}$$

$$\frac{2}{\tilde{r}}\frac{\partial^3\psi}{\partial\tilde{r}^2\partial\theta} + \left(\frac{1}{2}-\frac{4}{\tilde{r}^2}\right)\left(\frac{\partial^2\psi}{\partial\tilde{r}\partial\theta} - \frac{1}{\tilde{r}}\frac{\partial\psi}{\partial\theta}\right) - \frac{1}{\tilde{r}}\frac{\partial\omega}{\partial\theta} = 0 \quad \text{at } \theta = \pi/2+\alpha, \tag{4.4}$$

the symmetry conditions

$$\psi = \omega = 0 \quad \text{at } \theta = \pi/2, \tag{4.5}$$

the far field conditions

$$\psi \to 0, \quad \omega \to 0 \quad \text{as } \tilde{r} \to \infty, \tag{4.6}$$

and the matching conditions

$$\psi \sim 2\tilde{r}\log\tilde{r}\cos\theta - b_\infty\tilde{r}\cos\theta, \quad \omega \sim -\frac{4\cos\theta}{\tilde{r}} \quad \text{as } \tilde{r} \to 0. \tag{4.7}$$

As we noted earlier, a constant flow in the $\tilde{y}$-direction, $\psi = k\tilde{r}\cos\theta$, $\omega = 0$, satisfies the equations and boundary conditions, except for the far field and matching conditions. We can think of b_∞ as the strength of the uniform component of the flow that must emerge from the inner region to ensure that there is no flow at infinity. Note that the singular part of the flow near the origin given by (4.7) is simply a Stokeslet, which

indicates that the outer flow sees a point force and a uniform flow at the tip of the wedge, which is provided by the surface tension–driven inner flow.

Although this is a linear boundary value problem in a wedge, there is no obvious way of solving it analytically using integral transforms. This remains true even when $\alpha = \pi$ and the domain of the solution is a half-plane. However, when $\alpha = \pi/2$, we can use the results given in [4], which show that $b_\infty(\pi/2) \approx 1.869$. We can also consider the limit of a slender wedge, $\alpha \ll 1$, for which a simple asymptotic solution is available, before we discuss the numerical solution of (4.1) to (4.7).

4.1. The slender wedge, $\alpha \ll 1$. When $\alpha \ll 1$, $\theta = \frac{\pi}{2} + O(\alpha)$, so we define the scaled variables

$$\theta = \frac{\pi}{2} - \alpha\bar{\theta}, \quad \psi = \alpha\bar{\psi}, \quad \omega = \alpha\bar{\omega}.$$

Equations (4.1) and (4.2) at leading order then show that $\bar{\psi}$ and $\bar{\omega}$ are linear in θ, and the symmetry conditions give

$$\bar{\psi} = A_0(\tilde{r})\bar{\theta}, \quad \bar{\omega} = B_0(\tilde{r})\bar{\theta}. \tag{4.8}$$

The boundary conditions (4.3) and (4.4) are unchanged by this rescaling and show that

$$B_0 + 2A_0'' = 0,$$

$$\frac{2}{\tilde{r}}A_0'' + \left(\frac{1}{2} - \frac{4}{\tilde{r}^2}\right)\left(A_0' - \frac{1}{\tilde{r}}A_0\right) - \frac{1}{\tilde{r}}B_0 = 0,$$

where a prime denotes a derivative. We can eliminate $B_0(\tilde{r})$ between these equations and arrive at

$$\frac{4}{\tilde{r}}A_0'' + \left(\frac{1}{2} - \frac{4}{\tilde{r}^2}\right)A_0' - \frac{1}{\tilde{r}}\left(\frac{1}{2} - \frac{4}{\tilde{r}^2}\right)A_0 = 0, \tag{4.9}$$

to be solved subject to

$$A_0 \sim 2\tilde{r}\log\tilde{r} - b_\infty(0)\tilde{r} \quad \text{as } \tilde{r} \to 0, \tag{4.10}$$

$$A_0 \to 0 \quad \text{as } \tilde{r} \to \infty. \tag{4.11}$$

Since $A_0 = \tilde{r}$ is an obvious solution of (4.9), which corresponds to the uniform flow solution of the original equations, we can use reduction of order to find the solution in the form

$$A_0 = k_1\tilde{r}\int_{1/16}^{\tilde{r}^2/16} \frac{e^{-u}}{u}du + k_2\tilde{r}.$$

The boundary condition (4.10) then shows that

$$k_1 = 1, \quad k_2 = -b_\infty(0) - \frac{1}{16},$$

and finally, (4.11) shows that

$$b_\infty(0) = \int_{1/16}^{\infty} \frac{e^{-u}}{u}du - \frac{1}{16} \approx 2.194. \tag{4.12}$$

4.2. Numerical solutions. Before attempting to solve (4.1) to (4.7) numerically, it is convenient to subtract out the singularity at the origin by defining

$$\psi = 2\tilde{r}\log\tilde{r}\cos\theta + \hat{\psi}, \quad \omega = -\frac{4\cos\theta}{\tilde{r}} + \hat{\omega}.$$

We then find that

$$\tilde{\nabla}^2\hat{\psi} + \hat{\omega} = 0 \quad \text{for } \pi/2 < \theta < \pi/2 + \alpha, \tag{4.13}$$

$$\tilde{\nabla}^2\hat{\omega} + \frac{1}{2}\frac{\partial}{\partial\tilde{r}}(\tilde{r}\hat{\omega}) = 0 \quad \text{for } \pi/2 < \theta < \pi/2 + \alpha \tag{4.14}$$

subject to

$$\hat{\omega} + \frac{\partial^2\hat{\psi}}{\partial\tilde{r}^2} = 0 \quad \text{at } \theta = \pi/2 + \alpha, \tag{4.15}$$

$$\frac{2}{\tilde{r}}\frac{\partial^3\hat{\psi}}{\partial\tilde{r}^2\partial\theta} + \left(\frac{1}{2} - \frac{4}{\tilde{r}^2}\right)\left(\frac{\partial^2\hat{\psi}}{\partial\tilde{r}\partial\theta} - \frac{1}{\tilde{r}}\frac{\partial\hat{\psi}}{\partial\theta}\right) - \frac{1}{\tilde{r}}\frac{\partial\hat{\omega}}{\partial\theta} = \cos\alpha \quad \text{at } \theta = \pi/2 + \alpha, \tag{4.16}$$

$$\hat{\psi} = \hat{\omega} = 0 \quad \text{at } \theta = \pi/2, \tag{4.17}$$

$$\hat{\psi} \sim -2\tilde{r}\log\tilde{r}\cos\theta, \quad \hat{\omega} \sim \frac{4\cos\theta}{\tilde{r}} \quad \text{as } \tilde{r} \to \infty, \tag{4.18}$$

$$\hat{\psi} \to 0, \quad \hat{\omega} \to 0 \quad \text{as } \tilde{r} \to 0, \tag{4.19}$$

$$\frac{\partial\hat{\psi}}{\partial\tilde{r}} \sim b_\infty\cos\theta, \quad \text{as } \tilde{r} \to 0. \tag{4.20}$$

Note that this rescaled problem contains a forcing term at the boundary in (4.16). We have also written the boundary conditions at the origin in a form that allows us to solve (4.13) to (4.19), and then use (4.20) to determine b_∞.

We solve (4.13) to (4.19) using finite differences on a polar grid. We discretize θ at constant intervals and use a nonuniform grid in the $\tilde{r}$-direction in order to accurately capture the behavior of the solution as $\tilde{r} \to \infty$. By considering solutions on progressively finer grids, we found that we could achieve converged solutions (b_∞ accurate to two decimal places) with 50 equally spaced grid points in the θ-direction, and 275 grid points in the $\tilde{r}$-direction with spacing gradually changing from 0.01 to 1 as $\tilde{r}$ increases to 100. We approximate derivatives using central differences and use three-point formulas at the boundaries. After solving, we obtained b_∞ from (4.20), using a three-point formula to calculate $\partial\hat{\psi}/\partial\theta$ at the origin, and taking the mean value of b_∞ calculated at the half of the discretized values of θ furthest from the boundaries.

Figure 4.1 shows the streamlines, in terms of the original variable, ψ, at various values of α. We can see that the solution has a local maximum of vorticity in the interior for sufficiently large values of α. Figure 4.2 shows the calculated value of b_∞. The numerical solution is in good agreement with the asymptotic solutions for $\alpha = 0$ and $\pi/2$. Note that the form of the scaled equations when $\alpha \ll 1$ indicates that $b_\infty = 2.194 + O(\alpha^2)$, consistent with Figure 4.2. The range of values of b_∞ is not large, with $2.194 > b_\infty > 1.714$. We fitted a cubic spline to the numerically calculated values of b_∞, and used this in the numerical solutions of the inner problem, which we describe below.

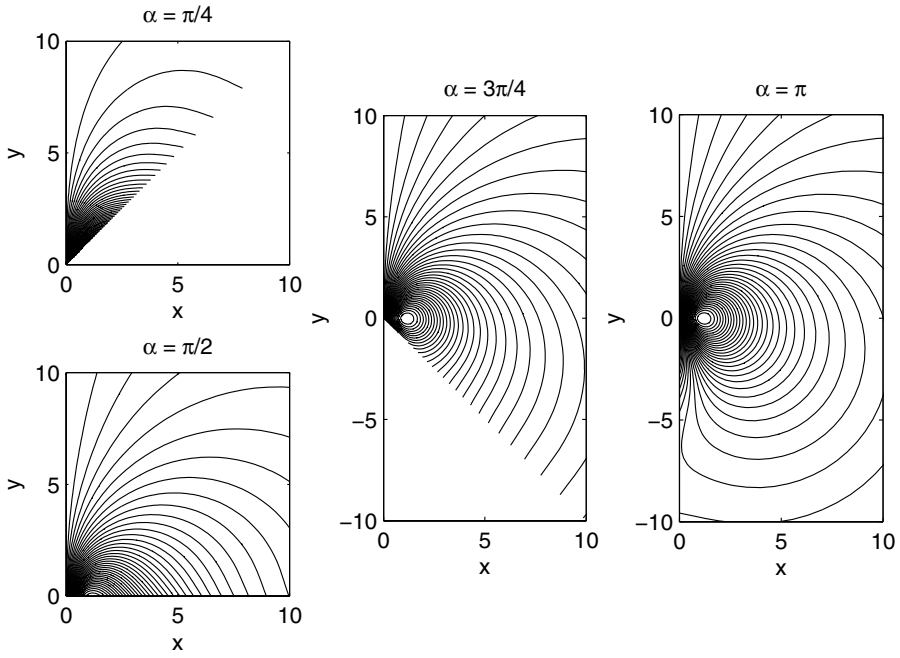


FIG. 4.1. *The streamlines in the outer region when* $\alpha = \pi/4$, $\pi/2$, $3\pi/4$, *and* π.

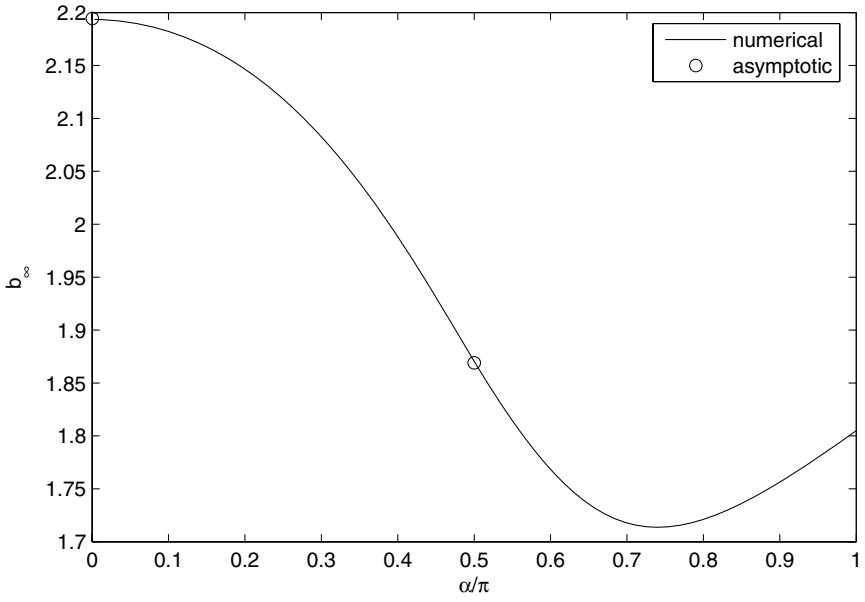


FIG. 4.2. *The numerically calculated value of* b_∞. *Also shown are the known values of* b_∞ *when* $\alpha = 0$ *and* $\alpha = \pi/2$.

5. Solution of the inner problem. Now that we know $b_\infty(\alpha)$, we are in a position to solve the inner problem (3.8) to (3.14). For this free boundary problem, it is natural to use the boundary integral method (see, for example, [10]). For Stokes flow in a bounded domain B,

$$\mathbf{u}(\mathbf{x}_0) = \frac{1}{2\pi}\int_{\partial B}\left\{-\mathbf{f}(\mathbf{x})\log\hat{r} + \frac{\hat{\mathbf{x}}\,(\mathbf{f}\cdot\hat{\mathbf{x}})}{\hat{r}^2} + \frac{4\,(\mathbf{u}\cdot\hat{\mathbf{x}})\,(\mathbf{n}\cdot\hat{\mathbf{x}})\,\hat{\mathbf{x}}}{\hat{r}^4}\right\} ds(\mathbf{x}), \tag{5.1}$$

where $\mathbf{n}$ is the outward unit normal, $\mathbf{f}$ is the force on the free surface, $\mathbf{x}_0$ lies on the free surface, s is arc length, $\hat{\mathbf{x}} = \mathbf{x} - \mathbf{x}_0$, and $\hat{r} = |\hat{\mathbf{x}}|$. We can now proceed by truncating our domain of solution with a circle of radius R_∞, centered on the origin, so that ∂B consists of the free surface, S, and the arc of the circle, S_∞. We discretize the free surface at n points an arc length $s = s_i$, $i = 1, 2, \ldots, n$, from the tip, with $s_1 = 0$ at the tip. At these points, $(\bar{x}, \bar{y}) = (X_i, Y_i) \equiv (X(s_i), Y(s_i))$. We also let $v(s)$ be the tangential fluid velocity at the free surface. We must therefore solve for the $3n + 1$ unknowns X_i, Y_i, $v_i \equiv v(s_i)$, and K.

Equations (3.10) to (3.12) show that

$$\mathbf{f} = (Y'X'' - X'Y'')\,\mathbf{n}, \quad \mathbf{u} = (\mathbf{n}\cdot\mathbf{X} + KX')\,\mathbf{n} + v\mathbf{t} \ \text{ on } S, \tag{5.2}$$

where $\mathbf{X} = (X, Y)$ and $\mathbf{t}$ is the unit tangent vector at the free surface. On S_∞ we assume that $\mathbf{f}$ and $\mathbf{u}$ take their far field values, so that

$$\begin{aligned}\mathbf{u} &= K\,(-2\log R_\infty + b_\infty)\sin\theta\,\mathbf{n} - K\,\{2(1+\log R_\infty) + b_\infty\}\cos\theta\,\mathbf{t},\\ \mathbf{f} &= -\frac{8K\sin\theta}{R_\infty}\mathbf{n} \ \text{ on } S_\infty.\end{aligned} \tag{5.3}$$

We represent the free surface, $(X(s), Y(s))$, and tangential velocity $v(s)$ using cubic splines. We evaluate the normal and tangential components of (5.1) using two-point Gaussian quadrature, collocating at the midpoint of each boundary element, using (5.2) and (5.3) to give the surface force and velocity, and making use of symmetry about the y-axis. We discretized most of S_∞ using equally spaced points. However, close to the point where S_∞ meets S in $x > 0$ we added extra points to resolve the rapid variation of the integrand when collocating at the midpoint of the final element of S. This provides $2n - 2$ nonlinear algebraic equations. We must also enforce the arc length condition

$$(X')^2 + (Y')^2 = 1. \tag{5.4}$$

We do this at the midpoint of each element, which provides a further $n - 1$ nonlinear algebraic equations. We now need four more conditions to close the problem. The far field conditions (3.13) and (3.14) give

$$Y_n = \cot\alpha\, X_n - 2K\log X_n + K(b_\infty - 3 + 2\log\sin\alpha), \tag{5.5}$$

$$v_n = -K\sin\theta_\infty(2\log R_\infty - b_\infty), \tag{5.6}$$

where θ_∞ is the value of θ at the point where S_∞ meets S. At $s = 0$, we imposed the symmetry conditions

$$X_1 = 0, \quad v_1 = 0. \tag{5.7}$$

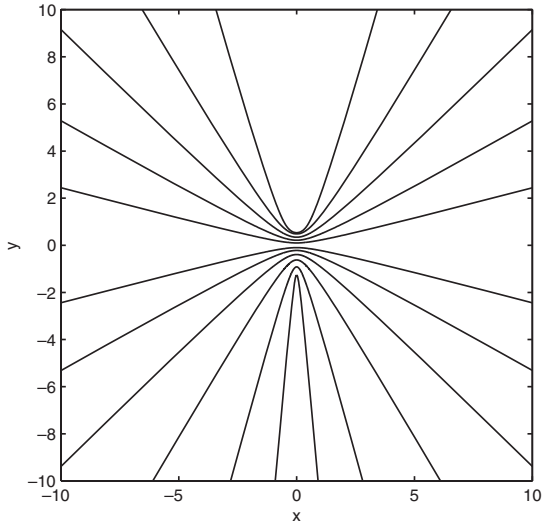


FIG. 5.1. *The numerical solution for* $\alpha = 15°$, $30°$, $45°$, $60°$, $75°$, $105°$, $120°$, $135°$, $150°$, $165°$, *and* $175°$.

We solved this system of $3n + 1$ nonlinear algebraic equations using Newton's method, calculating the Jacobian by numerical differentiation. We started with the known solution for $\alpha = \pi/2$, namely, the flat interface with $K = 0$, and progressively increased or decreased α. In this manner, we were able to obtain converged numerical solutions for $1° \leq \alpha \leq 178°$ or, in radians, $0.0175 \leq \alpha \leq 3.1067$. For $\alpha < \pi/2$, we used $n = 532$ and $s_n = 10^4$, with a grid for which Δs_i progressively increased from $\Delta s_1 = 10^{-2}$. This enabled us to compute down to small values of α for which, as we shall see in section 6, the curvature at the tip is bounded, and changes occur over a long lengthscale. Solutions on coarser grids indicate that the errors in our calculations of K are less than 1%. For $\alpha > \pi/2$, we used $n = 1074$ and $s_n = 50$, with a grid whose spacing became progressively finer for smaller s, with $\Delta s_1 = 5 \times 10^{-4}$. This allowed us to resolve the small region close to the tip, which we shall discuss in section 7. Solutions on coarser grids again indicate that the errors in our calculations of K are less than 1%.

Figure 5.1 shows the solution for various values of $\alpha \geq 15°$. We can see that for α close to π, the curvature of the tip becomes large, but that this does not occur as α approaches zero. The behavior for small α is illustrated in Figure 5.2, which shows that the curvature remains bounded, but that the position of the tip moves off in the negative y-direction as α decreases.

Figures 5.3, 5.4, and 5.5 show how K, the curvature at the tip, and the position of the tip vary with α. Also shown are the linearized predictions for $|\alpha - \pi/2| \ll 1$ given by Billingham [4]. To summarize, these are

$$K \sim -\frac{1}{2\pi}\left(\alpha - \frac{\pi}{2}\right), \quad \kappa(0) \sim -\frac{4}{\pi}\left(\alpha - \frac{\pi}{2}\right),$$

$$Y(0) \sim 0.359\left(\alpha - \frac{\pi}{2}\right) \quad \text{for } |\alpha - \pi/2| \ll 1. \tag{5.8}$$

The numerical and asymptotic solutions are in excellent agreement. We can also see that K is singular as $\alpha \to 0$, as is $Y(0)$, while the curvature remains bounded. As

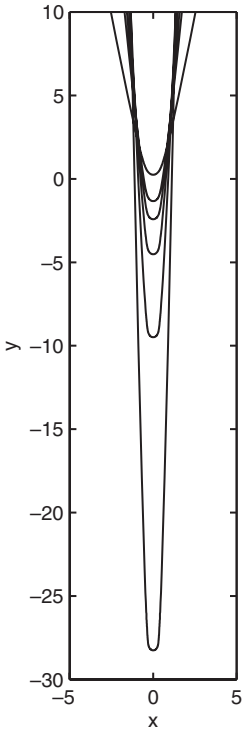


FIG. 5.2. *The numerical solution for* $\alpha = 10°$, $5°$, $4°$, $3°$, $2°$, *and* $1°$.

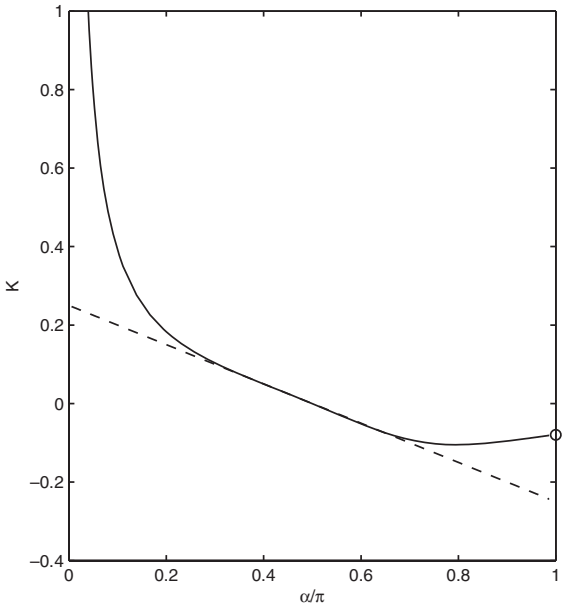


FIG. 5.3. *The numerically calculated value of* K, *along with the asymptotic solution for* $|\alpha - \pi/2| \ll 1$ *(broken line). The value of* K *when* $\alpha = \pi$, *calculated in section 7* ($K = K_0 = -1/4\pi$), *is marked with a circle.*

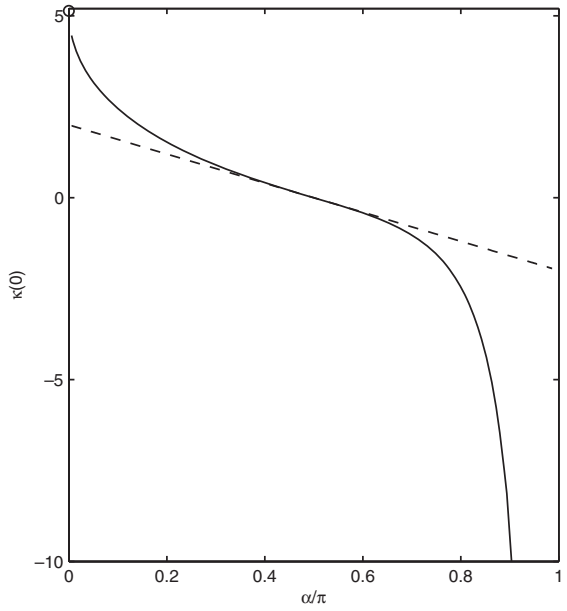


FIG. 5.4. *The numerically calculated value of the curvature at the tip of the recoiling wedge, along with the asymptotic solution for* $|\alpha - \pi/2| \ll 1$ *(broken line). The curvature when* $\alpha = 0$*, calculated in section* 6*, is marked with a circle.*

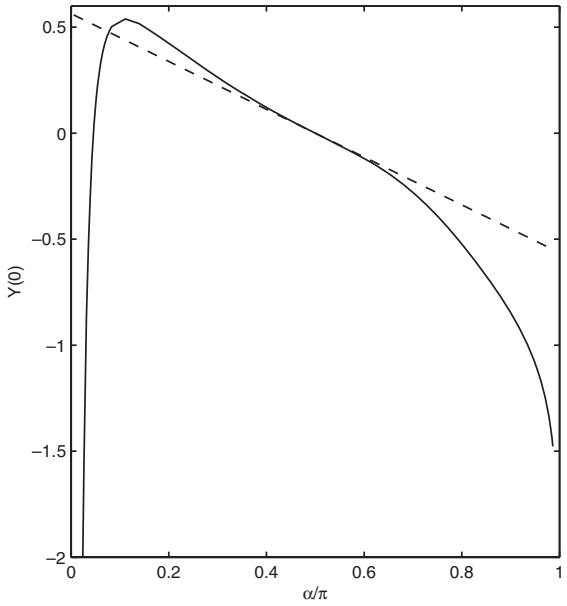


FIG. 5.5. *The numerically calculated value of the position of the tip of the recoiling wedge, along with the asymptotic solution for* $|\alpha - \pi/2| \ll 1$ *(broken line).*

$\alpha \to \pi$, K is bounded, while the curvature is strongly singular, and $Y(0)$ is weakly singular. We will now investigate these cases further using the method of matched asymptotic expansions.

6. Asymptotic solution of the inner problem for $\alpha \ll 1$. In this lubrication limit, it is more convenient to parameterize the free surface as lying at $\bar{x} = \bar{X}(\bar{y})$, in terms of which (3.8) to (3.12) become

$$0 = -\bar{\nabla}\bar{p} + \bar{\nabla}^2\bar{\mathbf{u}} \quad \text{for } 0 < \bar{x} < \bar{X}(\bar{y}), \tag{6.1}$$

$$\bar{\nabla}\cdot\bar{\mathbf{u}} = 0 \quad \text{for } 0 < \bar{x} < \bar{X}(\bar{y}), \tag{6.2}$$

$$\bar{y}\frac{d\bar{X}}{d\bar{y}} - \bar{X} = K\frac{d\bar{X}}{d\bar{y}} + \bar{u}_y\frac{d\bar{X}}{d\bar{y}} - \bar{u}_x \quad \text{at } \bar{x} = \bar{X}(\bar{y}), \tag{6.3}$$

$$\left\{\left(\frac{d\bar{X}}{d\bar{y}}\right)^2 - 1\right\}\left(\frac{\partial\bar{u}_x}{\partial\bar{y}} + \frac{\partial\bar{u}_y}{\partial\bar{x}}\right) - 4\frac{d\bar{X}}{d\bar{y}}\frac{\partial\bar{u}_x}{\partial\bar{x}} = 0 \quad \text{at } \bar{x} = \bar{X}(\bar{y}), \tag{6.4}$$

$$\begin{aligned} -\bar{p} - 2\left\{1 + \left(\frac{d\bar{X}}{d\bar{y}}\right)^2\right\}^{-1}&\left[\left\{\left(\frac{d\bar{X}}{d\bar{y}}\right)^2 - 1\right\}\frac{\partial\bar{u}_x}{\partial\bar{x}} + \frac{d\bar{X}}{d\bar{y}}\left(\frac{\partial\bar{u}_x}{\partial\bar{y}} + \frac{\partial\bar{u}_y}{\partial\bar{x}}\right)\right] \\ &= \frac{d^2\bar{X}}{d\bar{y}^2}\left\{1 + \left(\frac{d\bar{X}}{d\bar{y}}\right)^2\right\}^{-3/2} \quad \text{at } \bar{x} = \bar{X}(\bar{y}), \end{aligned} \tag{6.5}$$

along with the symmetry conditions

$$\bar{u}_x = \frac{\partial\bar{u}_y}{\partial\bar{x}} = \frac{\partial\bar{p}}{\partial\bar{x}} = 0 \quad \text{at } \bar{x} = 0 \tag{6.6}$$

subject to the matching conditions (3.14) and

$$\bar{X} \sim \tan\alpha\left\{\bar{y} + 2K\log\left(\frac{\bar{y}}{\cos\alpha}\right) - K\left(b_\infty(0) - 3\right)\right\} \quad \text{as } \bar{y} \to \infty, \tag{6.7}$$

where $b_\infty(0)$ is given by (4.12).

In order to be able to match the solutions in the two regions that we describe below, we need $K = \hat{K}/\alpha$ with $\hat{K} = O(1)$ as $\alpha \to 0$. We then use a simple lubrication scaling that produces a leading order balance in (6.2), namely,

$$\bar{y} = \hat{y}/\alpha, \quad \bar{x} = \hat{x}, \quad \bar{X} = \hat{X}, \quad \bar{u}_x = \hat{u}_x, \quad \bar{u}_y = \hat{u}_y/\alpha, \quad \bar{p} = \hat{p}.$$

On substituting these scalings into (6.1) to (6.7), we find that, at leading order, $\hat{u}_y \equiv \hat{u}_y(\hat{y})$, $\hat{u}_x = -\hat{x}\hat{u}_y'$, and $\hat{p} = -2\hat{u}_y$, where a prime denotes $d/d\hat{y}$. Equation (6.3) then shows that

$$\frac{\hat{X}'}{\hat{X}} = \frac{1 + \hat{u}_y'}{\hat{y} - \hat{K} - \hat{u}_y}. \tag{6.8}$$

In order to close the problem, we need to consider the tangential stress condition, (6.4), at $O(\alpha^2)$, which shows that

$$\hat{u}_y' = -\frac{2\hat{K}}{\hat{X}}. \tag{6.9}$$

By eliminating $\hat{u}_y$ between (6.8) and (6.9), we obtain an equation for $\hat{X}$, which can be solved analytically to give the implicit solution that satisfies (6.7),

$$\begin{aligned} \hat{y} &= \hat{X} - 2\hat{K}\log\hat{X} + 2\hat{K}\log\alpha - (3 - b_\infty(0))\,\hat{K}, \\ \hat{u}_y &= -2\hat{K}\log\hat{X} - \frac{4\hat{K}^2}{\hat{X}} + 2\hat{K}\log\alpha + \hat{K}b_\infty(0). \end{aligned} \tag{6.10}$$

We will see below why we can justify treating the term of $O(\log\alpha)$ as a constant in this procedure. Note that the leading order shape of the free surface just reproduces the far field boundary condition.

Since $\hat{X}' = \hat{X}/(\hat{X} - 2\hat{K})$, the slope of the free surface becomes unbounded as $\hat{X} \to 2\hat{K}$, and we need an inner-inner region in the neighborhood of $\hat{X} = 2\hat{K}$ in order to complete the solution. Note that $\hat{y} \to \hat{y}_0$ and $\hat{u}_y \to \hat{u}_{y0}$ as $\hat{X} \to 2\hat{K}$, where

$$\hat{y}_0 = (b_\infty(0) - 1)\,\hat{K} - 2\hat{K}\log\left(2\hat{K}/\alpha\right), \quad \hat{u}_{y0} = \hat{y}_0 - \hat{K}. \tag{6.11}$$

By determining the form of the next correction to the asymptotic expansion in the neighborhood of $\hat{y} = \hat{y}_0$, we find that there is a nonuniformity in the expansion when $\hat{y} - \hat{y}_0 = O(\alpha)$.

6.1. Inner-inner region. Appropriate scaled variables in the inner-inner region are[1]

$$\hat{y} = \hat{y}_0 + \alpha\tilde{y}, \quad \hat{x} = \tilde{x}, \quad \hat{X} = \tilde{X}, \quad \hat{u}_x = \tilde{u}_x, \quad \hat{u}_y = \hat{u}_{y0} + \alpha\tilde{u}_y, \quad \hat{p} = \tilde{p}.$$

In terms of the original inner variable defined in section 3, this rescaling just represents a shift of $O(\alpha^{-1})$ in the y-direction along with the removal of a uniform flow of $O(\alpha^{-1})$ in the y-direction. As we would expect, we therefore recover the equations for surface tension–driven Stokes flow. However, the kinematic condition becomes

$$(\tilde{y} - \tilde{u}_y)\,\frac{d\tilde{X}}{d\tilde{y}} = \tilde{X} - \tilde{u}_x, \tag{6.12}$$

which specifies the normal component of the fluid velocity at the free surface, and the far field conditions are given by matching as

$$\tilde{u}_x \sim \tilde{x}, \quad \tilde{u}_y \sim -\tilde{y}, \quad \tilde{p} \to 2, \quad \tilde{X} \to 2\hat{K} \ \text{ as } \tilde{y} \to \infty. \tag{6.13}$$

We can solve this boundary value problem using the boundary element method described in section 5. Note that in this case it is easier to truncate the domain of solution using the straight line $\tilde{y} = \tilde{y}_\infty$, and that we can evaluate the contribution to the integral in (5.1) along this line analytically for the simple far field given by (6.13).

We found the solution by introducing an artificial continuation parameter, β, modifying the far field conditions to be

$$\tilde{X}'\cos\beta - \tilde{Y}'\sin\beta = 0, \quad \tilde{u}_y = -\tilde{y}_\infty\cos\beta \ \text{ at } \tilde{y} = \tilde{y}_\infty.$$

When the continuation parameter β is $\pi/2$, we can converge to a solution from an initial guess with $\tilde{X} = \tilde{u}_y = \tilde{u}_x = 0$, and when $\beta = 0$, we recover the problem whose

[1] We have used tildes here for notational convenience, and note that these variables should not be confused with those used in the outer region.

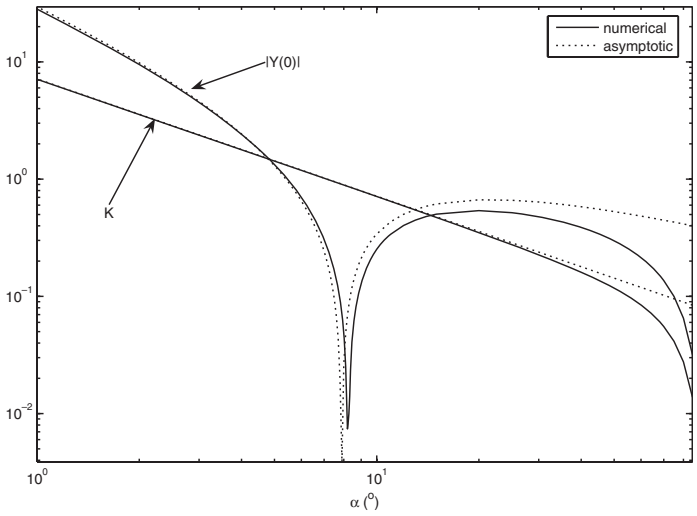


FIG. 6.1. *A comparison between the numerical solution in the inner region and the asymptotic solution for* $\alpha \ll 1$.

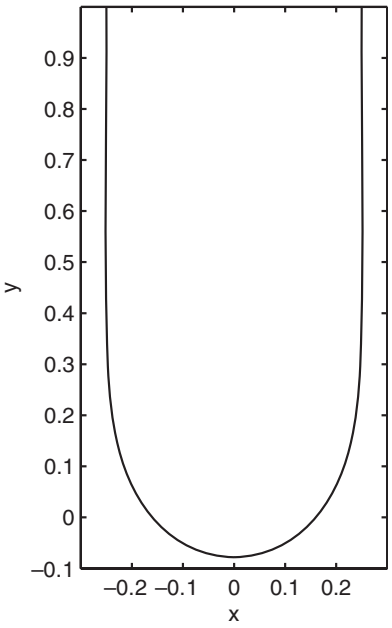


FIG. 6.2. *The asymptotic solution in the inner-inner region for* $\alpha \ll 1$.

solution we are interested in. Starting from $\beta = \pi/2$ and successively solving for smaller values of β, we were able to obtain a converged solution with $\beta = 0$. From the condition that $\tilde{X} \to 2\hat{K}$ as $\tilde{y} \to \infty$, we find that $\hat{K} \approx 0.125$, and hence $K \approx 0.125/\alpha$, and that the curvature at the tip of the fluid is $\kappa(0) \approx 5.13$, consistent with the behavior of the full numerical solution shown in Figure 5.4. Figure 6.1 shows the asymptotic and numerical values of K and $\bar{Y}(0)$, which are in very good agreement.

Note that

$$\bar{Y}(0) = (b_\infty(0) - 1)\, K - 2K \log 2K,$$

so that the logarithmic constant that appears in (6.11) is absorbed into the definition of K. Figure 6.2 shows the asymptotic solution in the inner-inner region, and should be compared to the solutions shown in Figure 5.2. Note that, although $\bar{Y}(0) < 0$ in Figure 5.2, the tip lies at

$$Y(0) \equiv -Kt \log t + t\bar{Y}(0) \sim \frac{\hat{K}t}{\alpha} \log\left(\frac{\alpha^2 e^{b_\infty(0)-1}}{\hat{K}^2 t}\right) \quad \text{as } \alpha \to 0 \text{ for } t \ll 1. \tag{6.14}$$

We conclude that in the double limit $\alpha \to 0$, $t \to 0$ we require that $t \ll \alpha^2$ or $t = O(\alpha^2)$ for this solution to be valid.

7. Asymptotic solution of the inner problem for $\pi - \alpha \ll 1$. As we saw in the previous section, it is more convenient to parameterize the free surface as lying at $\bar{x} = \bar{X}(\bar{y})$. By balancing terms in the equations, we find that the richest asymptotic balance arises when we leave K and all of the variables unscaled, except for $\bar{X}$, which we write as $\bar{X} = \epsilon\bar{\chi}$, where $\chi = O(1)$ for $\epsilon = \pi - \alpha \ll 1$. The region D, where the fluid lies, is given by $\bar{x} > 0$ for $\bar{y} \geq \bar{y}_0$, and $\bar{x} > \epsilon\bar{\chi}(\bar{y})$ for $\bar{y} \leq \bar{y}_0$. The free surface meets the $\bar{y}$-axis at $\bar{y} = \bar{y}_0$, which we need to determine. We must therefore solve

$$0 = -\bar{\nabla}\bar{p} + \bar{\nabla}^2\bar{\mathbf{u}} \ \text{ in } D, \tag{7.1}$$

$$\bar{\nabla} \cdot \bar{\mathbf{u}} = 0 \ \text{ in } D, \tag{7.2}$$

$$\bar{y}\frac{d\bar{\chi}}{d\bar{y}} - \bar{\chi} = K\frac{d\bar{\chi}}{d\bar{y}} + \bar{u}_y\frac{d\bar{\chi}}{d\bar{y}} - \epsilon^{-1}\bar{u}_x \ \text{ at } \bar{x} = \epsilon\bar{\chi}(\bar{y}) \text{ for } \bar{y} < \bar{y}_0, \tag{7.3}$$

$$\left\{\epsilon^2\left(\frac{d\bar{\chi}}{d\bar{y}}\right)^2 - 1\right\}\left(\frac{\partial\bar{u}_x}{\partial\bar{y}} + \frac{\partial\bar{u}_y}{\partial\bar{x}}\right) - 4\epsilon\frac{d\bar{\chi}}{d\bar{y}}\frac{\partial\bar{u}_x}{\partial\bar{x}} = 0 \ \text{ at } \bar{x} = \epsilon\bar{\chi}(\bar{y}) \text{ for } \bar{y} < \bar{y}_0, \tag{7.4}$$

$$\begin{aligned} -\bar{p} - 2\left\{1 + \epsilon^2\left(\frac{d\bar{\chi}}{d\bar{y}}\right)^2\right\}^{-1}\left[\left\{\epsilon^2\left(\frac{d\bar{\chi}}{d\bar{y}}\right)^2 - 1\right\}\frac{\partial\bar{u}_x}{\partial\bar{x}} + \epsilon\frac{d\bar{\chi}}{d\bar{y}}\left(\frac{\partial\bar{u}_x}{\partial\bar{y}} + \frac{\partial\bar{u}_y}{\partial\bar{x}}\right)\right] \\ = \epsilon\frac{d^2\bar{\chi}}{d\bar{y}^2}\left\{1 + \left(\frac{d\bar{\chi}}{d\bar{y}}\right)^2\right\}^{-3/2} \quad \text{at } \bar{x} = \epsilon\bar{\chi}(\bar{y}) \text{ for } \bar{y} < \bar{y}_0, \end{aligned} \tag{7.5}$$

$$\bar{u}_x = \frac{\partial\bar{u}_y}{\partial\bar{x}} = \frac{\partial\bar{p}}{\partial\bar{x}} = 0 \ \text{ at } \bar{x} = 0 \text{ for } \bar{y} > \bar{y}_0 \tag{7.6}$$

subject to the matching conditions (3.14) and

$$\bar{\chi} \sim -\frac{\tan\epsilon}{\epsilon}\left\{\bar{y} + 2K\log\left(\frac{-\bar{y}}{\cos\epsilon}\right) - K\left(b_\infty(\pi) - 3\right)\right\} \quad \text{as } \bar{y} \to -\infty. \tag{7.7}$$

The numerical results presented in section 4 show that $b_\infty(\pi) \approx 1.82$.

We expand $\bar{u}_x = \bar{u}_{x0} + \epsilon \bar{u}_{x1} + O(\epsilon^2)$, and similarly for the other variables. The far field solution, (3.14), is a solution of the full problem at leading order. This solution is singular, which indicates that an inner-inner region will be needed. Since the singularity must be where the free surface meets the $\bar{y}$-axis, the appropriate leading order solution is

$$\bar{p}_0 = 4K_0 \frac{\bar{y} - \bar{y}_0}{\bar{x}^2 + (\bar{y} - \bar{y}_0)^2}, \quad \bar{u}_{x0} = 2K_0 \frac{\bar{x}(\bar{y} - \bar{y}_0)}{\bar{x}^2 + (\bar{y} - \bar{y}_0)^2},$$

$$\bar{u}_{y0} = -K_0 \log\left\{\bar{x}^2 + (\bar{y} - \bar{y}_0)^2\right\} - 2K_0 \frac{\bar{x}^2}{\bar{x}^2 + (\bar{y} - \bar{y}_0)^2} + K_0 b_\infty. \tag{7.8}$$

Note that $\bar{\chi}_0$, the leading order position of the free surface, is not determined by the leading order problem, and is coupled to the solution at $O(\epsilon)$, which satisfies, in terms of a streamfunction in polar coordinates,

$$\bar{\nabla}^4 \psi = 0 \quad \text{for } 0 < \theta < \pi, \tag{7.9}$$

$$\begin{aligned} &- \left[K_0 \left\{b_\infty(\pi) + 1 - 2\log r\right\} - \bar{y}_0 + r\right] \frac{d\bar{\chi}_0}{dr} \\ &+ \left(1 + \frac{2K_0}{r}\right) \bar{\chi}_0 = \frac{\partial \psi}{\partial r} \quad \text{at } \theta = \pi, \end{aligned} \tag{7.10}$$

$$\frac{\partial^2 \psi}{\partial r^2} - \frac{1}{r}\frac{\partial \psi}{\partial r} - \frac{1}{r^2}\frac{\partial \psi}{\partial \theta^2} = 8K_0 \frac{d}{dr}\left(\frac{\bar{\chi}_0}{r}\right) \quad \text{at } \theta = \pi, \tag{7.11}$$

$$-\frac{3}{r}\frac{\partial^3 \psi}{\partial r^2 \partial \theta} + \frac{3}{r^2}\frac{\partial^2 \psi}{\partial r \partial \theta} - \frac{4}{r^3}\frac{\partial \psi}{\partial \theta} - \frac{1}{r^3}\frac{\partial^3 \psi}{\partial \theta^3} = \frac{d^3 \bar{\chi}_0}{dr^3} \quad \text{at } \theta = \pi, \tag{7.12}$$

$$\psi = \frac{\partial^2 \psi}{\partial \theta^2} = 0 \quad \text{at } \theta = 0, \tag{7.13}$$

$$\psi \sim -2K_1 r \log r \sin\theta + \left(K_1 b_\infty(\pi) + K_0 b'_\infty(\pi)\right) r \sin\theta \quad \text{as } r \to \infty, \tag{7.14}$$

$$\bar{\chi}_0 \sim r - \bar{y}_0 - 2K_0 \log r + (b_\infty(\pi) - 3)K_0 \quad \text{as } r \to \infty. \tag{7.15}$$

Note that we have chosen a polar coordinate system with the $\bar{y}$-axis at $\theta = 0$, which is more convenient for the following calculations. We must therefore solve for Stokes flow in the upper half-plane, driven by a stress and normal velocity prescribed on the negative $\bar{y}$-axis, and coupled linearly to the unknown position of the free surface, $\bar{\chi}_0$.

Clearly, the behavior of the solution as $r \to 0$ is crucial, and we investigate this first. If we assume that $\bar{\chi}_0 \sim kr^n$ as $r \to 0$ for some constant k, then the streamfunction must take the form $\psi \sim r^n \log r\, f(\theta) + r^n\, g(\theta)$. If we first consider the terms of $O(r^n \log r)$ and make use of the symmetry condition (7.13), we find that a biharmonic streamfunction has

$$f(\theta) = A_0 \sin n\theta + B_0 \sin(n-2)\theta$$

for some constants A_0 and B_0. On substituting this form into the boundary conditions (7.10) to (7.12), we obtain three linear equations in these two constants. The only

way that we can satisfy these is if n is a half integer. In order for this to be able to match the solution in the inner-inner region, $\bar{\chi}_0$ must be bounded, with nonzero derivative as $r \to 0$, which means that $n = 1/2$ and $\bar{\chi}_0 \sim kr^{1/2}$ as $r \to 0$. This then allows us to fix A_0 and B_0 and obtain

$$f(\theta) = \frac{1}{2}K_0 k \left(3\sin\frac{1}{2}\theta - \sin\frac{3}{2}\theta\right). \tag{7.16}$$

At $O(r^{1/2})$ we obtain

$$g(\theta) = A_1 \sin\frac{1}{2}\theta + B_1 \sin\frac{3}{2}\theta + \frac{1}{2}K_0 k\theta\left(3\cos\frac{1}{2}\theta + \cos\frac{3}{2}\theta\right).$$

On substituting this into the normal stress boundary condition, (7.12), we find that A_1 and B_1 do not appear, so that K_0 is determined at this point in the analysis, with $K_0 = -1/4\pi$. This is in excellent agreement with numerical solutions of the full inner problem, as shown in Figure 5.3. The boundary conditions (7.10) and (7.11) then fix A_1 and B_1, so that

$$\begin{aligned} g(\theta) = k\Bigg[&-\frac{1}{4}\left\{K_0\left(3b_\infty(\pi) - 1\right) - 3\bar{y}_0\right\}\sin\frac{1}{2}\theta + \frac{1}{4}\left\{K_0\left(b_\infty(\pi) + 5\right) - \bar{y}_0\right\}\sin\frac{3}{2}\theta \\ &+\frac{1}{2}K_0\theta\left(3\cos\frac{1}{2}\theta + \cos\frac{3}{2}\theta\right)\Bigg]. \end{aligned} \tag{7.17}$$

It now remains to determine k as a function of $\bar{y}_0$. We could proceed numerically, but, as we shall see in section 7.2, $\bar{y}_0$ is large and negative for $\epsilon \ll 1$, consistent with the behavior shown in Figure 5.5.

7.1. Inner solution for $-\bar{y}_0 \gg 1$. We seek to solve (7.9) to (7.15) when $\bar{y}_0$ is large and negative. Note that since this problem is linear, the results that we obtained above for $r \ll 1$ remain valid when $-\bar{y}_0$ is large. In particular, we must have $\bar{\chi}_0 \sim kr^{1/2}$ as $r \to 0$ and $K_0 = -1/4\pi$. Note also that arbitrary multiples of the solutions of the homogeneous problem, $r\log r\sin\theta$ and $r\sin\theta$, can be added to ψ without affecting the equations satisfied by $\bar{\chi}_0$. We therefore subtract the far field behavior, given by (7.14), from ψ, so that we require $\psi = o(r)$ as $r \to \infty$.

If we define $\delta = -\bar{y}_0^{-1} \ll 1$, we find that we can obtain a suitable leading order balance by defining scaled variables

$$r = \delta^{-1}\tilde{r}, \quad \bar{\chi}_0 = \delta^{-1}\tilde{\chi}, \quad \psi = \delta^{-2}\tilde{\psi}, \quad k = \delta^{-1/2}\tilde{k},$$

in terms of which (7.9) to (7.15) become

$$\bar{\nabla}^4\tilde{\psi} = 0 \quad \text{for } 0 < \theta < \pi, \tag{7.18}$$

$$\begin{aligned} &-\left[K_0\left\{b_\infty(\pi) + 1 - 2\log\tilde{r} + 2\log\delta\right\}\delta + 1 + \tilde{r}\right]\frac{d\tilde{\chi}}{d\tilde{r}} \\ &\qquad + \left(1 + \frac{2K_0\delta}{\tilde{r}}\right)\tilde{\chi} = \frac{\partial\tilde{\psi}}{\partial\tilde{r}} \quad \text{at } \theta = \pi, \end{aligned} \tag{7.19}$$

$$\frac{\partial^2\tilde{\psi}}{\partial\tilde{r}^2} - \frac{1}{\tilde{r}}\frac{\partial\tilde{\psi}}{\partial\tilde{r}} - \frac{1}{\tilde{r}^2}\frac{\partial\tilde{\psi}}{\partial\theta^2} = 8K_0\delta\frac{d}{d\tilde{r}}\left(\frac{\tilde{\chi}}{\tilde{r}}\right) \quad \text{at } \theta = \pi, \tag{7.20}$$

$$-\frac{3}{\tilde{r}}\frac{\partial^3\tilde{\psi}}{\partial\tilde{r}^2\partial\theta}+\frac{3}{\tilde{r}^2}\frac{\partial^2\tilde{\psi}}{\partial\tilde{r}\partial\theta}-\frac{4}{\tilde{r}^3}\frac{\partial\tilde{\psi}}{\partial\theta}-\frac{1}{\tilde{r}^3}\frac{\partial^3\tilde{\psi}}{\partial\theta^3}=-\delta\frac{d^3\tilde{\chi}}{d\tilde{r}^3}\quad\text{at }\theta=\pi, \tag{7.21}$$

$$\tilde{\psi}=\frac{\partial^2\tilde{\psi}}{\partial\theta^2}=0\quad\text{at }\theta=0, \tag{7.22}$$

$$\tilde{\psi}=o(\tilde{r})\quad\text{as }\tilde{r}\to\infty, \tag{7.23}$$

$$\tilde{\chi}\sim\tilde{r}+1-2K_0\delta\log\tilde{r}+2K_0\delta\log\delta+(b_\infty(\pi)-3)K_0\delta\quad\text{as }\tilde{r}\to\infty, \tag{7.24}$$

$$\tilde{\chi}\sim\tilde{k}\tilde{r}^{1/2}\quad\text{as }\tilde{r}\to 0, \tag{7.25}$$

$$\begin{aligned}\tilde{\psi}\sim&-\frac{1}{4}\tilde{k}\left(3\sin\frac{1}{2}\theta-\sin\frac{3}{2}\theta\right)+\frac{1}{2}K_0\tilde{k}\delta\left(3\sin\frac{1}{2}\theta-\sin\frac{3}{2}\theta\right)\tilde{r}^{1/2}\left(\log\tilde{r}-\log\delta\right)\\&+\tilde{k}K_0\delta\left\{-\frac{1}{4}\left(3b_\infty(\pi)-1\right)\sin\frac{1}{2}\theta+\frac{1}{4}\left(b_\infty(\pi)+5\right)\sin\frac{3}{2}\theta\right.\\&\left.+\frac{1}{2}\theta\left(3\cos\frac{1}{2}\theta+\cos\frac{3}{2}\theta\right)\right\}\tilde{r}^{1/2}\quad\text{as }\tilde{r}\to 0.\end{aligned} \tag{7.26}$$

We now expand $\tilde{\psi}=\tilde{\psi}_0+\delta\log\delta\tilde{\psi}_1+\delta\tilde{\psi}_2+o(\delta)$, and similarly for $\tilde{\chi}$ and $\tilde{k}$. At leading order and at $O(\delta\log\delta)$, analytical solutions are available, with

$$\begin{aligned}\tilde{\psi}_0&=-\frac{1}{4}\tilde{k}_0\left(3\sin\frac{1}{2}\theta-\sin\frac{3}{2}\theta\right)\tilde{r}^{1/2},\\\tilde{\chi}_0&=\frac{1}{2}\tilde{k}_0\left\{(\tilde{r}+1)\tan^{-1}\tilde{r}^{1/2}+\tilde{r}^{1/2}\right\},\quad\tilde{k}_0=\frac{4}{\pi},\end{aligned} \tag{7.27}$$

$$\begin{aligned}\tilde{\psi}_1&=-\left(\frac{1}{4}\tilde{k}_1+\frac{1}{2}K_0\tilde{k}_0\right)\left(3\sin\frac{1}{2}\theta-\sin\frac{1}{2}\theta\right)\tilde{r}^{1/2},\\\tilde{\chi}_1&=\tilde{k}_1\tan^{-1}\tilde{r}^{1/2},\quad\tilde{k}_1=\frac{4K_0}{\pi}.\end{aligned} \tag{7.28}$$

At $O(\delta)$ the problem is forced by the leading order solution, and no analytical solution is available directly. Although the problem can be solved, in principle, using Mellin integral transforms, this is not a practical approach, since the Mellin transform of $\tilde{\chi}_0$ is not available analytically. The problem at $O(\delta)$ can be solved numerically, using a technique similar to that used in section 4.2, but, as we shall see, we do not actually need it in what follows. Finally, note that the weak nonuniformity between the terms in the expansion of $\tilde{\chi}$ for $\tilde{r}\gg 1$ is just a reordering and can be ignored.

7.2. The inner-inner region for $\pi-\alpha=\epsilon\ll 1$. Since $\bar{\chi}_0\sim 4\delta^{-1/2}r^{1/2}/\pi$ as $r\to 0$ for $\delta=-\bar{y}_0^{-1}\ll 1$, and we expect a nonuniformity when $\epsilon d\bar{\chi}_0/dr=O(1)$, the inner-inner region has $\bar{y}=\bar{y}_0+O(\epsilon^2\delta^{-1})$, $\bar{x}=O(\epsilon^2\delta^{-1})$. We therefore define scaled variables

$$\bar{x}=\epsilon^2\delta^{-1}\tilde{\bar{x}},\quad\bar{y}=\bar{y}_0+\epsilon^2\delta^{-1}\tilde{\bar{y}},\quad\bar{X}=\epsilon\bar{\chi}=\epsilon^2\delta^{-1}\tilde{\bar{X}},\quad\bar{\mathbf{u}}=\tilde{\bar{\mathbf{u}}},\quad\bar{p}=\epsilon^{-2}\delta\tilde{\bar{p}}.$$

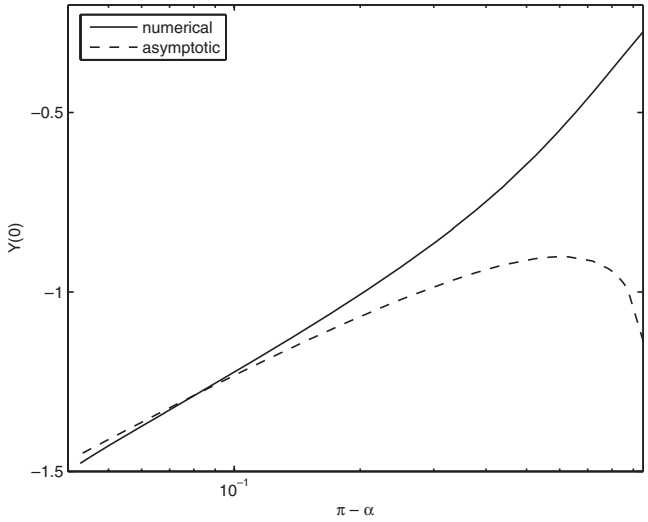


FIG. 7.1. *The numerically calculated value of the position of the tip of the recoiling wedge and the asymptotic approximation given by* (7.30).

At leading order we obtain surface tension–driven Stokes flow, but with kinematic condition

$$\tilde{\bar{u}}_y \frac{d\tilde{\bar{X}}}{d\tilde{\bar{y}}} - \tilde{\bar{u}}_x = (\bar{y}_0 - K_0)\frac{d\tilde{\bar{X}}}{d\tilde{\bar{y}}}. \tag{7.29}$$

By writing the inner solution in terms of these inner-inner variables, we can obtain the matching condition, which gives the inner-inner behavior as $\tilde{\bar{x}}^2 + \tilde{\bar{y}}^2 \to \infty$, in the usual way. Note, however, that we must retain the contribution from the $O(\epsilon)$ correction to $\bar{u}_x$ in the inner region, since it was this that determined the scalings for the inner region. Moreover, we have to be careful with logarithmic terms. The matching condition for $\tilde{\bar{u}}_y$ is

$$\tilde{\bar{u}}_y \sim -2K_0 \log\left(\epsilon^2\delta^{-1}\right) - K_0 \log\left(\tilde{\bar{x}}^2 + \tilde{\bar{y}}^2\right) - \frac{2K_0\tilde{\bar{x}}^2}{\tilde{\bar{x}}^2 + \tilde{\bar{y}}^2} + K_0 b_\infty(\pi) + o(1) \text{ as } \tilde{\bar{x}}^2 + \tilde{\bar{y}}^2 \to \infty.$$

If we define $\tilde{\bar{U}}_y = \tilde{\bar{u}}_y + 2K_0 \log\left(\epsilon^2\delta^{-1}\right)$ and $\tilde{y}_0 = \bar{y}_0 + 2K_0 \log\left(\epsilon^2\delta^{-1}\right) = \bar{y}_0 + 2K_0 \log\left(-\epsilon^2\bar{y}_0\right)$, with $\tilde{\bar{U}}_y$, $\tilde{y}_0 = O(1)$, we can remove the logarithm from the problem. This means that $\bar{y}_0$, the position of the tip, is weakly singular, with

$$\begin{aligned} \bar{y}_0 &\equiv -\delta^{-1} = -4K_0 \log\epsilon - 2K_0 \log\left(4K_0 \log\epsilon\right) + \tilde{y}_0 + o(1) \\ &\equiv \frac{1}{\pi}\log\epsilon + \frac{1}{2\pi}\log\left(-\frac{1}{\pi}\log\epsilon\right) + \tilde{y}_0 + o(1). \end{aligned} \tag{7.30}$$

Figure 7.1 shows that numerical solutions of the full inner problem are consistent with this weak singularity in the position of the tip of the wedge. It now remains to determine $\tilde{y}_0$ by solving the inner-inner problem numerically.

Note that the kinematic condition (7.29) is now

$$\tilde{\bar{U}}_y \frac{d\tilde{\bar{X}}}{d\tilde{\bar{y}}} - \tilde{\bar{u}}_x = (\tilde{y}_0 - K_0)\frac{d\tilde{\bar{X}}}{d\tilde{\bar{y}}}. \tag{7.31}$$

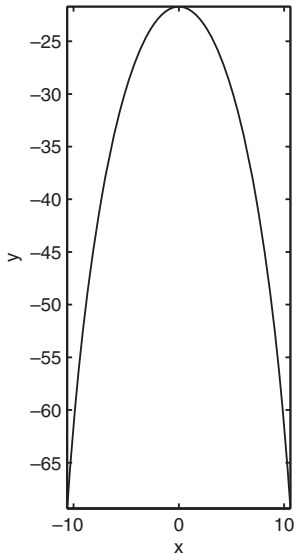


FIG. 7.2. *The numerical solution of the inner-inner problem.*

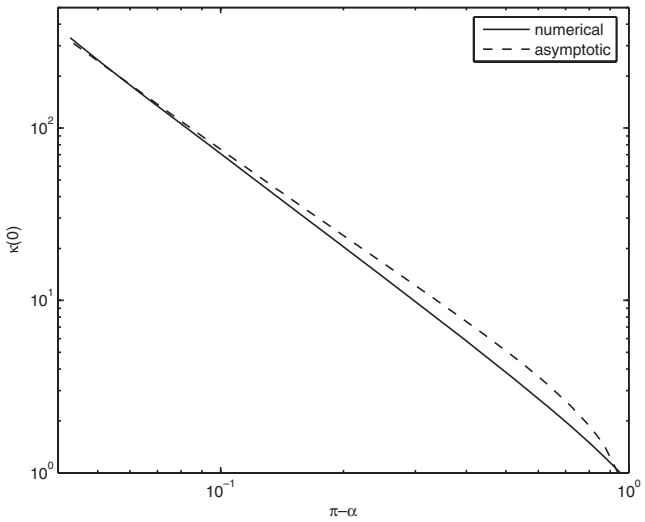


FIG. 7.3. *The numerically calculated value of the curvature at the tip of the recoiling wedge, and our asymptotic estimate for* $\pi - \alpha \ll 1$.

We also have the matching conditions

$$\tilde{\tilde{X}} \sim \frac{4}{\pi}\left(-\tilde{\tilde{y}}\right)^{1/2} \quad \text{as } \tilde{\tilde{y}} \to -\infty, \tag{7.32}$$

and, in terms of a streamfunction and polar coordinates,

$$\tilde{\tilde{\psi}} \sim \frac{1}{2\pi}\tilde{\tilde{r}}\log\tilde{\tilde{r}}\sin\theta - \frac{b_\infty(\pi)}{4\pi}\tilde{\tilde{r}}\sin\theta - \frac{1}{\pi}\left(3\sin\frac{1}{2}\theta - \sin\frac{3}{2}\theta\right)\tilde{\tilde{r}}^{1/2} \text{as } \tilde{\tilde{r}} \to \infty. \tag{7.33}$$

We can solve this boundary value problem using the numerical method described in section 5. Since we were unable to find a suitable artificial continuation parameter,

analogous to β in section 6.1, we used the outer asymptotic solution, (7.27), as our initial guess, from which we were able to obtain a converged solution. We found that $\bar{y}_0 \approx -0.381$, which gives the asymptotic solution for the position of the tip of the wedge shown in Figure 7.1. The numerical solution for the free surface is shown in Figure 7.2, which should be compared to the solutions shown in Figure 5.1. From this, we can calculate the curvature at the tip of the wedge, which is of $O(\delta/\epsilon^2)$. This asymptotic estimate is shown in Figure 7.3.

8. Conclusions. In this paper, we have presented an analysis of the initial motion of a viscous fluid with a free surface under the action of surface tension when the fluid lies in a wedge at $t = 0$. We showed that this is an example of a problem in which inertia is never negligible as $t \to 0$, even though an obvious viscous-dominated similarity scaling exists, since the solution develops over two asymptotic regions: a nonlinear inner region a distance of $O(t \log t)$ from the origin with size $O(t)$ dominated by surface tension and viscous forces, and a linear outer region with size $O(t^{1/2})$ dominated by viscous and inertial forces. The results of [1] show that the addition of an exterior fluid with viscosity much less than that of the interior fluid is a singular perturbation, and the solution can then be described in terms of a region with size $O(t)$, where the flow is dominated by surface tension and viscosity. It would be interesting to investigate the structure of the initial flow in the two-fluid problem when the viscosity ratio is small. Another interesting problem is the case of an initial cone of fluid. The inviscid version of this problem has been studied in [6] and [7].

One reason for attempting this analysis was that it was hoped that the limiting cases presented in sections 6 and 7 would provide some insight into the initial stages of the coalescence of droplets and bubbles. This analogy has proved to be useful in the case of inviscid coalescence (see [11]). However, as discussed in the appendix, this idea does not work here. An analysis of this problem, which shows that $t \log t$-dependence does occur in these coalescence problems, but with coefficients different than those that arise in the wedge problem, is given in [9].

Appendix. Inapplicability to the coalescence of drops and bubbles. Since the asymptotic structure of the solution of the wedge problem indicates that the inner region is small, and that the tip of the wedge lies a distance from the origin much greater than the size of the inner region, we can postulate that the inner flow sees only the local slope of the interface, even if the free surface is actually curved on a much longer lengthscale, an idea that works well for inviscid coalescence (see [11]). Unfortunately, because the separation of scales between the size of the inner region and its distance from the origin is only of $O(\log t)$, we find that this simple idea is not applicable.

A.1. Drops. The wedge solution with $\alpha \ll 1$ indicates that

$$Y(0) \sim \frac{\hat{K}t}{\alpha}\left\{-\log t + \log\left(\frac{\alpha^2 e^{b_\infty(0)-1}}{\hat{K}^2}\right)\right\} \quad \text{as } \alpha \to 0 \text{ for } t \ll 1. \tag{A.1}$$

If initially $x = \epsilon y^2$ with $\epsilon \ll 1$, then, locally,

$$\alpha \sim \frac{dx}{dy} \sim 2\epsilon y \sim -2\epsilon \frac{\hat{K}}{\alpha} t \log t,$$

and hence $\alpha \sim \sqrt{-2\epsilon \hat{K} t \log t}$. However, this means that the two terms inside the braces in (A.1) are of comparable size, and there is no separation of scales.

A.2. Bubbles. The wedge solution with $\pi - \alpha \ll 1$ indicates that

$$Y(0) \sim -\frac{1}{2\pi} t \log t + t \left\{ \frac{1}{\pi} \log(\pi - \alpha) + \frac{1}{2\pi} \log \left(-\frac{1}{\pi} \log (\pi - \alpha) \right) + \tilde{y}_0 \right\} \tag{A.2}$$
$$\text{as } \alpha \to \pi \text{ for } t \ll 1.$$

The same argument as given in the previous subsection indicates that, if initially $x = \epsilon y^2$ with $\epsilon \ll 1$, then $\pi - \alpha \sim -\epsilon t \log t / 2\pi$. The second term in (A.2) is then comparable to the first, and, again, there is no separation of scales.

REFERENCES

[1] M. J. Miksis and J.-M. Vanden-Broeck, *Self-similar dynamics of a viscous wedge of fluid*, Phys. Fluids, 11 (1999), pp. 3227–3231.

[2] J. B. Keller and M. J. Miksis, *Surface tension driven flow*, SIAM J. Appl. Math., 43 (1983), pp. 268–277.

[3] J. B. Lawrie, *Surface tension driven flow in a wedge*, Quart. J. Mech. Appl. Math., 43 (1990), pp. 251–273.

[4] J. Billingham, *Surface-tension-driven flow in fat fluid wedges and cones*, J. Fluid Mech., 397 (1999), pp. 45–71.

[5] J. B. Keller, P. A. Milewski, and J.-M. Vanden-Broeck, *Merging and wetting driven by surface tension*, Eur. J. Mech. B Fluids, 19 (2000), pp. 491–502.

[6] S. P. Decent and A. C. King, *The recoil of a broken liquid bridge*, in Proceedings of the IUTAM Symposium on Free Surface Flows, A. C. King and Y. D. Shikhmurzaev, eds., Kluwer Academic, Dordrecht, 2001, pp. 81–88.

[7] A. Sierou and J. R. Lister, *Self-similar recoil of inviscid drops*, Phys. Fluids, 15 (2004), pp. 1379–1394.

[8] L. Ting and J. B. Keller, *Slender jets and thin sheets with surface tension*, SIAM J. Appl. Math., 50 (1990), pp. 1533–1546.

[9] J. Eggers, J. R. Lister, and H. A. Stone, *Coalescence of liquid drops*, J. Fluid Mech., 401 (1999), pp. 293–310.

[10] C. Pozrikidis, *Boundary Integral and Singularity Methods for Linearized Viscous Flow*, Cambridge University Press, Cambridge, UK, 1992.

[11] J. Billingham and A. C. King, *Surface tension-driven flow outside a slender wedge with an application to the inviscid coalescence of drops*, J. Fluid Mech., 533 (2005), pp. 193–221.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 533–553

KINETICS OF MARTENSITIC PHASE TRANSITIONS: LATTICE MODEL*

LEV TRUSKINOVSKY† AND ANNA VAINCHTEIN‡

Abstract. Martensitic phase transitions are often modeled by mixed-type hyperbolic-elliptic systems. Such systems lead to ill-posed initial-value problems unless they are supplemented by an additional kinetic relation. In this paper we explicitly compute an appropriate closing relation by replacing the continuum model with its natural discrete prototype. The procedure can be viewed as either regularization by discretization or a physically motivated account of underlying discrete microstructure. We model phase boundaries by traveling wave solutions of a fully inertial discrete model for a bi-stable lattice with harmonic long-range interactions. Although the microscopic model is Hamiltonian, it generates macroscopic dissipation which can be specified in the form of a relation between the velocity of the discontinuity and the conjugate configurational force. This kinetic relation respects entropy inequality but is not a consequence of the usual Rankine–Hugoniot jump conditions. According to the constructed solution, the dissipation at the macrolevel is due to the induced radiation of lattice waves carrying energy away from the propagating front. We show that sufficiently strong nonlocality of the lattice model may be responsible for the multivaluedness of the kinetic relation and can quantitatively affect kinetics in the near-sonic region. Direct numerical simulations of the transient dynamics suggest stability of at least some of the computed traveling waves.

Key words. martensitic phase transitions, lattice models, nonlocal interactions, driving force, lattice waves, radiative damping

AMS subject classifications. 37K60, 74N10, 74N20, 74H05

DOI. 10.1137/040616942

1. Introduction.

A characteristic feature of martensitic phase transitions in active materials is the energy dissipation leading to experimentally observed hysteresis. The dissipation is due to propagating phase boundaries that can be represented at the continuum level as surfaces of discontinuity. Classical elastodynamics admits nonzero dissipation on moving discontinuities but provides no information about its origin and kinetics. Although the arbitrariness of the rate of dissipation does not create problems in the case of classical shock waves, it is known to be the cause of nonuniqueness in the presence of subsonic phase boundaries (see [7, 13, 20] for recent reviews).

The ambiguity at the macroscale reflects the failure of the continuum theory to describe phenomena inside the narrow transition fronts where dissipation actually takes place. The missing closing relation can be found by analyzing a regularized theory which describes the fine structure of the transition front. When the local curvature effects can be neglected, the problem reduces to the study of a one-dimensional steady-state problem. To formulate the simplest problem of this type it is sufficient to consider longitudinal motions of a homogeneous elastic bar. The total energy of

*Received by the editors October 13, 2004; accepted for publication (in revised form) July 22, 2005; published electronically December 30, 2005.
http://www.siam.org/journals/siap/66-2/61694.html
†Laboratoire de Mechanique des Solides, CNRS-UMR 7649, Ecole Polytechnique, 91128, Palaiseau, France (trusk@lms.polytechnique.fr). The work of this author was supported by NSF grant DMS-0102841.
‡Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (aav4@pitt.edu). The work of this author was supported by NSF grant DMS-0137634.

such bar is the sum of kinetic and potential contributions

$$\mathcal{E} = \int \left[\frac{\rho \dot{u}^2}{2} + \phi(u_x) \right] dx, \tag{1.1}$$

where $u(x,t)$ is the displacement field, $\dot{u} \equiv \partial u/\partial t$ is the velocity, $u_x \equiv \partial u/\partial x$ is the strain, ρ is the constant mass density, and $\phi(u_x)$ is the elastic energy density. The function $u(x,t)$ satisfies the nonlinear wave equation

$$\rho \ddot{u} = (\sigma(u_x))_x, \tag{1.2}$$

where $\sigma(u_x) = \phi'(u_x)$ is the stress-strain relation.

Although in classical elastodynamics (1.2), which is often presented as a first-order p-system, is assumed to be hyperbolic, the hyperbolicity condition $\sigma'(u_x) > 0$ is violated for martensitic materials with nonmonotone stress-strain relation $\sigma(u_x)$ [8]. This makes the initial-value problem associated with the mixed-type equation (1.2) ill-posed; in particular, it leads to the appearance of discontinuities violating the Lax condition (subsonic phase boundaries or kinks, e.g., [16, 17, 31]).

To be more specific, consider a discontinuity moving with velocity V. Let f_- and f_+ denote the limiting values of a function $f(x)$ to the left and to the right of the interface, and introduce the notations $[\![f]\!] \equiv f_+ - f_-$ for the jump and $\{f\} \equiv (f_+ + f_-)/2$ for the average of f across the discontinuity. The parameters on a discontinuity must satisfy both the classical Rankine–Hugoniot jump conditions,

$$[\![\dot{u}]\!] + V[\![u_x]\!] = 0, \quad \rho V[\![\dot{u}]\!] + [\![\sigma(u_x)]\!] = 0, \tag{1.3}$$

and the entropy inequality $\mathcal{R} = GV \geq 0$, where

$$G = [\![\phi]\!] - \{\sigma\}[\![u_x]\!] \tag{1.4}$$

is the configurational (driving) force. Contrary to conventional shock waves, the martensitic phase boundaries usually fail to satisfy the Lax condition $c_+ < V < c_-$, where $c_\pm$ are the values of the sound velocity in front and behind the discontinuity.

One way to remedy the resulting nonuniqueness is to supplement (1.3) by a *kinetic relation* specifying the dependence of the configurational force on the velocity of the phase boundary $G = G(V)$ [1, 30]. Since the nonlinear wave equation (1.2) provides no information about the kinetic relation, the dependence $G(V)$ has often been modeled phenomenologically [1, 29, 30]. An alternative approach has been to derive the kinetic relation from an augmented model incorporating regularizing terms. A typical example is the viscosity-capillarity model, accounting for both dispersive and dissipative corrections [21, 29]. The problem with both approaches is that they introduce into the theory parameters of unclear physical origin.

The aim of the present paper is to obtain the kinetic relation without any phenomenological assumptions at the macroscale by means of direct replacement of the continuum model (1.2) with its natural discrete prototype. Such procedure of going back from continuum to discrete level can be viewed as either regularization by discretization or as a physically motivated account of underlying atomic or mesoscopic microstructure. It is clear that the discrete model must be Hamiltonian to reproduce the conservative structure of the smooth solutions of (1.2). The energy dissipation on the discontinuities can then be interpreted as the nonlinearity-induced radiation of lattice-scale waves which takes the energy away from the long-wave continuum level. This phenomenon is known in physics literature as radiative damping (e.g., [11, 12]).

To regularize (1.2) from "first principles," we consider in this paper fully inertial dynamics of a one-dimensional lattice with bi-stability and long-range interactions. Following some previous work on cracks [22] and dislocations [2], we assume piecewise linear approximation of nonlinearity and construct an explicit traveling wave solution of the discrete problem. There exists an extensive literature on shock waves and solitons in the local and nonlocal discrete systems with convex interatomic potentials (e.g., [9, 10, 19, 27]) and on the semilinear prototypes of the present bi-stable system (e.g., [3, 5, 18]). A discrete quasilinear problem for martensitic phase transitions and failure waves in the chains with nearest-neighbor (NN) interactions was considered in [24, 25, 26, 35]. In the present paper we extend these results to the case of harmonic interactions of finite but arbitrary long range. We show that the local (NN) model is degenerate, find a general solution of the nonlocal model, and provide detailed illustrations for a particular case.

Our analytic solutions demonstrate that the nonlocal model generates a much broader class of admissible solutions than the local model; in particular, it allows the possibility of radiation both in front and behind the moving discontinuity. We also show that sufficiently strong nonlocality may be responsible for the multivaluedness of the kinetic relation and can quantitatively affect kinetics in the near-sonic region. The advantage of the explicit formulas obtained in the paper is that they capture certain details that are difficult or even impossible to detect in numerical simulations, such as singular behavior of solutions near static-dynamic bifurcation and around resonances.

The paper is organized as follows. The piecewise linear discrete model with long-range interactions and the associated dynamical system are introduced in section 2. In section 3 we formulate the dimensionless equations for the traveling waves, the boundary conditions, and the admissibility conditions. An explicit solution for the steady state motion of an isolated phase boundary is obtained by Fourier transform in section 4. In section 5 we obtain static solutions describing lattice-trapped phase boundaries and link them to a nontrivial limit of the dynamic solutions. The energy transfer from long to short waves is studied in section 6, where we obtain a closed-form kinetic relation. In section 7 we illustrate the general theory via the case when the only long-range interactions are due to the second nearest neighbors. Numerical simulations of the transient problem suggesting stability of at least sufficiently fast traveling waves are described in section 8. The last section contains our conclusions.

2. Discrete model. The simplest lattice structure can be modeled as a chain of point masses connected by elastic springs. Suppose that the interactions are of long-range type and that every particle interacts with its q neighbors on each side. If $u_n(t)$ is the displacement of the nth particle, the total energy of the chain can be written as

$$\mathcal{E} = \varepsilon \sum_{n=-\infty}^{\infty} \left[\frac{\rho \dot{u}_n^2}{2} + \sum_{p=1}^{q} p\phi_p\left(\frac{u_{n+p} - u_n}{p\varepsilon} \right) \right], \tag{2.1}$$

where ε is the reference interparticle distance and $\phi_p(w)$ is the energy density of the interaction between pth nearest neighbors. The dynamics of the chain with energy (2.1) is governed by an infinite system of ordinary differential equations:

$$\rho \ddot{u}_n = \frac{1}{\varepsilon} \sum_{p=1}^{q} \left[\phi_p'\left(\frac{u_{n+p} - u_n}{p\varepsilon} \right) - \phi_p'\left(\frac{u_n - u_{n-p}}{p\varepsilon} \right) \right]. \tag{2.2}$$

A continuum system, formally obtained by identifying $u(x,t)$ with a limit of $u(n\varepsilon,t) = u_n(t)$ as $\varepsilon \to 0$, reduces to the nonlinear wave equation (1.2) with a specific macroscopic stress-strain relation

$$\sigma(w) = \sum_{p=1}^{q} p\phi_p'(w). \tag{2.3}$$

When the function $\sigma(w)$ is nonmonotone, (1.2) constitutes an incomplete description of the limit. As we show, in this case the correct limit procedure starting from the discrete problem (2.1), (2.2) must also produce a specific kinetic relation $G = G(V)$. This relation, expressed exclusively in terms of the elastic potentials entering (2.1), provides the desired closure for the macroscopic problem (1.2), (1.3). Here we do not consider the issue of a nucleation criterion, whose discrete prototype was studied in [17].

To obtain analytical results, we consider the simplest potentials allowing for a possibility of a phase transitions: bi-quadratic for local interactions (NN) and quadratic for nonlocal interactions (NNN, NNNN, etc.). Specifically we define

$$\phi_1(w) = \begin{cases} \dfrac{1}{2}\Psi(1)w^2, & w \le w_c, \\ \dfrac{1}{2}\Psi(1)(w-a)^2 + a\Psi(1)\left(w_c - \dfrac{a}{2}\right), & w \ge w_c, \end{cases} \tag{2.4}$$

and

$$\phi_p(w) = \frac{1}{2}p\Psi(p)w^2, \quad p = 2, \dots, q. \tag{2.5}$$

One can see that the nonlinear springs representing NN interactions can be found in two different states depending on whether the strain w is below (phase I) or above (phase II) the critical value w_c. Parameter a defines the microscopic transformation strain (distance between the two linear branches); note that a and w_c are in general independent. For simplicity we assume that the two energy wells of the bi-stable NN potential have equal curvatures $\Psi(1) > 0$.

It is convenient to reformulate the problem using dimensionless variables:

$$\bar{t} = t(\Psi(1)/\rho)^{1/2}/\varepsilon, \quad \bar{u_n} = u_n/(a\varepsilon), \quad \bar{w_c} = w_c/a, \quad \bar{\Psi}(p) = \Psi(p)/\Psi(1), \quad p = 1, \dots, q. \tag{2.6}$$

In terms of these variables with the bars dropped, the energy (2.1) becomes

$$\mathcal{E} = \sum_{n=-\infty}^{\infty}\left[\frac{\dot{u}_n^2}{2} - \frac{1}{2}\sum_{|k-n|\le q} u_n\Psi(k-n)u_k - (u_n - u_{n-1} - w_c)\theta(u_n - u_{n-1} - w_c)\right]. \tag{2.7}$$

By introducing the strain variables $w_n = u_n - u_{n-1}$, we can rewrite the governing equations (2.2) in the form

$$\ddot{w}_n - \sum_{|k-n|\le q}\Psi(k-n)w_k = 2\theta(w_n - w_c) - \theta(w_{n+1} - w_c) - \theta(w_{n-1} - w_c), \tag{2.8}$$

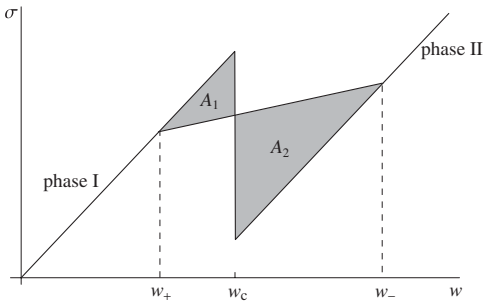


FIG. 2.1. *The bi-linear macroscopic stress-strain law and the Rayleigh line connecting the states at infinity for a traveling wave solution describing an isolated phase boundary. The difference between the shaded areas $A_2 - A_1$ represents the configurational force.*

where

$$\Psi(0) = -2\sum_{p=1}^{q} \Psi(p), \quad \Psi(-p) = \Psi(p), \tag{2.9}$$

and $\theta(w)$ is a unit step function. The macroscopic stress-strain relation (2.3) takes the form

$$\sigma(w) = c^2 w - \theta(w - w_c), \tag{2.10}$$

where

$$c = \left(\sum_{p=1}^{q} p^2 \Psi(p)\right)^{1/2} \tag{2.11}$$

is the dimensionless macroscopic sound speed. The microscopic elastic moduli $\Psi(p)$ must be chosen to ensure that the uniform deformation $w_n = w$ is stable in each of the phases. For this it is necessary and sufficient that all phonon frequencies $\omega^2(k) > 0$, with $\omega(k)$ defined in (3.6) and $k \in (0, \pi]$, are real. This implies, in particular, that the square of the macroscopic sound speed (2.11) is positive. The resulting macroscopic stress-strain relation (2.10) is shown in Figure 2.1.

3. Traveling waves. An isolated phase boundary moving with a constant velocity V can be obtained as a traveling wave solution of (2.8) with $w_n(t) = w(\xi)$, $\xi = n - Vt$. We assume further that in the moving coordinate system all springs in the region $\xi > 0$ are in phase I ($w_n < w_c$), and all springs with $\xi < 0$ are in phase II. The system (2.8) can then be replaced by a single nonlinear advance-delay differential equation:

$$V^2 w'' - \sum_{|p| \le q} \Psi(p) w(\xi + p) = 2\theta(-\xi) - \theta(-\xi - 1) - \theta(1 - \xi). \tag{3.1}$$

The configurations at $\xi = \pm\infty$ must correspond to stable homogeneous equilibria plus superimposed short-wave oscillations with zero average; the averaging is over the largest period of oscillations but can be also defined as

$$\langle w(\xi) \rangle = \lim_{s \to \infty} \frac{1}{s} \int_{\xi}^{\xi + s} w(\zeta)\, d\zeta. \tag{3.2}$$

In terms of the averaged quantities we obtain the following boundary conditions:

$$\langle w(\xi)\rangle \to w_\pm \quad \text{as } \xi \to \pm\infty. \tag{3.3}$$

The nonlinearity of the problem is in the switching condition

$$w(0) = w_c. \tag{3.4}$$

We assume that a solution is admissible if the NN springs in front of the moving interface are still in phase I and behind it already in phase II. This implies that

$$w(\xi) < w_c \quad \text{for } \xi > 0, \qquad w(\xi) > w_c \quad \text{for } \xi < 0. \tag{3.5}$$

Consequently, the mathematical problem reduces to solving (3.1) subject to (3.3), (3.4), and (3.5).

Observe first that the equation (2.8) is linear in each phase ($\xi < 0$ and $\xi > 0$), which means that the solution can be represented as a superposition of linear waves $w_n = \exp(i(kn - \omega t))$. Since the elastic moduli are equal, the dispersion relation

$$\omega^2(k) = 4\sum_{p=1}^{q} \Psi(p) \sin^2 \frac{pk}{2} \tag{3.6}$$

is the same in both phases. For the linear modes to be compatible with the traveling wave ansatz, their phase velocity $V_p(k) = \omega/k$ must be equal to V. This gives the restriction on the admissible wave numbers in the form

$$L(k, V) = 0, \tag{3.7}$$

where

$$L(k, V) = 4\sum_{p=1}^{q} \Psi(p) \sin^2 \frac{pk}{2} - V^2 k^2. \tag{3.8}$$

Among the modes selected by (3.7), the ones with complex wave numbers must be exponentially decaying on both sides of the front. They describe the core structure of the phase boundary. The modes with nonzero real wave numbers correspond to radiation. The waves with $k = 0$ are naturally associated with the macroscopic part of the solution.

4. Exact solution. We solve (3.1) by writing $w(\xi) = h(\xi) + w_-$ and applying the complex Fourier transform

$$\hat{h}(k) = \int_{-\infty}^{\infty} h(\xi) e^{i(k+i\alpha)\xi} d\xi, \quad h(\xi) = \frac{1}{2\pi} \int_{-\infty+i\alpha}^{\infty+i\alpha} \hat{h}(k) e^{-ik\xi} dk,$$

where $\alpha > 0$ is a small parameter which guarantees convergence of the integrals. After inverting the Fourier transform and letting $\alpha \to 0$, we obtain

$$w(\xi) = w_- - \frac{2}{\pi i} \int_\Gamma \frac{\sin^2(k/2) e^{ik\xi} dk}{kL(k, V)}, \tag{4.1}$$

where the contour Γ coincides with the real axis passing the singular point $k = 0$ from below. The singularities associated with nonzero real roots of $L(k, V) = 0$ must

comply with the radiation conditions. Specifically, the modes with group velocity $V_g = \partial\omega/\partial k$ larger than V can appear only in front, while the modes with $V_g < V$ can appear only behind the phase boundary [24]. Using the relation

$$V_g = V + \frac{L_k(k,V)}{2Vk}, \tag{4.2}$$

where $L_k(k,V) = \partial L/\partial k$ and assuming $V > 0$, we obtain that $V_g \gtrless V$ whenever $kL_k(k,V) \gtrless 0$. Therefore, to satisfy the radiation conditions, we need to dent the integration contour in (4.1) in such a way that it passes below the singularities on the real axis if $kL_k(k,V) > 0$ and above if $kL_k(k,V) < 0$.

To compute the integral (4.1) explicitly, we use the residue method closing the contour in the upper half-plane when $\xi > 0$ and in the lower half-plane when $\xi < 0$. The solutions look different in the generic case $q > 1$ and the degenerate case $q = 1$.

For $q > 1$ the Jordan lemma can be applied directly, and by separating the macroscopic part of the solution from the microscopic one, we obtain

$$w(\xi) = \begin{cases} w_- + \displaystyle\sum_{k\in M^-(V)} \frac{4\sin^2(k/2)e^{ik\xi}}{kL_k(k,V)} & \text{for } \xi < 0, \\ w_- - \dfrac{1}{c^2 - V^2} - \displaystyle\sum_{k\in M^+(V)} \frac{4\sin^2(k/2)e^{ik\xi}}{kL_k(k,V)} & \text{for } \xi > 0. \end{cases} \tag{4.3}$$

Here

$$M^{\pm}(V) = \{k : L(k,V) = 0,\ \mathrm{Im}k \gtrless 0\} \bigcup N^{\pm}(V) \tag{4.4}$$

are all roots of the dispersion relation contributing to the solution on either side of the front, with

$$N^{\pm}(V) = \{k : L(k,V) = 0,\ \mathrm{Im}k = 0,\ kL_k(k,V) \gtrless 0\} \tag{4.5}$$

denoting the sets of real roots describing radiation.

For $q = 1$ (NN interactions only) the contribution from a semi-arch at infinity does not vanish at $\xi = \pm 0$ and relations (4.3) must be supplemented by the following limiting conditions:

$$w(\xi) = \begin{cases} w_- + \displaystyle\sum_{k\in M^-(V)} \frac{4\sin^2(k/2)e^{ik\xi}}{kL_k(k,V)} - \frac{1}{2} & \text{for } \xi = -0, \\ w_- - \dfrac{1}{c^2 - V^2} - \displaystyle\sum_{k\in M^+(V)} \frac{4\sin^2(k/2)e^{ik\xi}}{kL_k(k,V)} + \frac{1}{2} & \text{for } \xi = +0. \end{cases} \tag{4.6}$$

In both cases, by applying the boundary conditions (3.3) at infinity we obtain

$$w_+ = w_- - \frac{1}{c^2 - V^2}. \tag{4.7}$$

It is easy to see that (4.7) coincides with the Rankine–Hugoniot relation $V^2[\![w]\!] = [\![\sigma]\!]$, computed for the macroscopic stress-strain relation (2.10). The continuity of $w(\xi)$ at $\xi = 0$ implies that

$$\frac{1}{c^2 - V^2} + \sum_{k\in M(V)} \frac{4\sin^2(k/2)}{kL_k(k,V)} = \begin{cases} 1, & q = 1, \\ 0, & q > 1, \end{cases} \tag{4.8}$$

where $M(V) = M^+(V) \bigcup M^-(V)$. Condition (4.8) is automatically satisfied for $q > 1$ since the sum of residues at all poles (including $k = 0$) equals zero; for $q = 1$ and $\xi = 0$ the integral over a contour at infinity contributes additional unity in the right-hand side of (4.8). The switching condition (3.4) together with (4.8) requires that

$$w_\pm = w_c \mp \frac{1}{2(c^2 - V^2)} + \sum_{k \in N_{\rm pos}(V)} \frac{4\sin^2(k/2)}{|kL_k(k,V)|}, \tag{4.9}$$

where $N_{\rm pos}(V) = \{k : L(k,V) = 0,\ {\rm Im}k = 0, k > 0\} \subset M(V)$ is the set of positive real roots of the dispersion relation. By virtue of (4.7), the two conditions (4.9) are dependent and can be replaced by a single condition:

$$\frac{1}{2}(w_- + w_+) - w_c = \sum_{k \in N_{\rm pos}(V)} \frac{4\sin^2(k/2)}{|kL_k(k,V)|}. \tag{4.10}$$

As we show in section 6, (4.10) represents the desired kinetic relation.

We can use the explicit formulas for $w(\xi)$ to reconstruct the particle velocity profile from $v(\xi) = -Vu'(\xi)$ if we assume that $V \neq 0$. The relation between the velocity and the strain fields reads

$$v(\xi) - v(\xi - 1) = -Vw'(\xi), \tag{4.11}$$

and the right-hand side is already known from (4.3). Solving (4.11) by Fourier transform, we obtain

$$v(\xi) = \begin{cases} v_+ - \dfrac{V}{c^2 - V^2} - 2V \displaystyle\sum_{k \in M^-(V)} \frac{\sin(k/2)e^{ik(\xi + \frac{1}{2})}}{L_k(k,V)} & \text{for } \xi < -\dfrac{1}{2}, \\ v_+ + 2V \displaystyle\sum_{k \in M^+(V)} \frac{\sin(k/2)e^{ik(\xi + \frac{1}{2})}}{L_k(k,V)} & \text{for } \xi > -\dfrac{1}{2}. \end{cases} \tag{4.12}$$

It is easy to check that the average velocities at infinity satisfy the remaining Rankine–Hugoniot condition $(1.3)_1$, which in our case takes the form

$$v_+ - v_- = \frac{V}{c^2 - V^2}. \tag{4.13}$$

Notice that the obtained set of traveling wave solutions is parametrized by the velocity V and the boundary value data $w_\pm$ and $v_\pm$. The average particle velocity v_+ in front can always be set equal to zero due to the Gallilean invariance. If the strain in front of the discontinuity is also prescribed, the remaining three macroscopic parameters are fully constrained by the two classical Rankine–Hugoniot conditions (4.7) and (4.13), plus the nonclassical admissibility condition (4.10).

5. Static solutions. A special consideration is needed when $V = 0$. In this case continuous variable $\xi = n - Vt$ takes integer values, and the strain profile becomes discontinuous at every $\xi = n$. The differential equation reduces to a system of finite-difference equations, and we can replace the continuous Fourier transform by its discrete analogue (see also [6, 11, 23]). First observe that for a piecewise continuous function $w(\xi)$ with discontinuities at integer ξ we have

$$\hat{w}(k) = \int_{-\infty}^{\infty} w(\xi)e^{ik\xi}d\xi = \sum_{n=-\infty}^{\infty} \int_n^{n+1} w(\xi)e^{ik\xi}d\xi.$$

Therefore, assuming that the strain profile $w(\xi)$ converges to $w_0(n)$ as $V \to 0$, we obtain

$$\hat{w}_0(k) = \sum_{n=-\infty}^{\infty} w_0(n) \frac{e^{ik(n+1)} - e^{ikn}}{ik} = \frac{e^{ik} - 1}{ik} \hat{w}_0^{\mathrm{D}}(k), \tag{5.1}$$

where $\hat{w}_0^{\mathrm{D}}(k) = \sum_{n=-\infty}^{\infty} w_0(n) e^{ikn}$ is the discrete Fourier transform of $w_0(n)$. Now we can use (4.1) to obtain

$$\hat{w}_0(k) = 2\pi\delta(k) w_- + \frac{4\sin^2(k/2)}{ik\omega^2(k)},$$

where $\delta(k)$ is the Dirac delta function and $\omega^2(k)$ is given by (3.6). Using (5.1) we can then find $\hat{w}_0^{\mathrm{D}}(k)$ and, applying inverse discrete Fourier transform, obtain a representation of the discrete solution:

$$w_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{w}_0^{\mathrm{D}}(k) e^{-ikn} dk = w_- - \frac{1}{\pi i} \int_{-\pi}^{\pi} \frac{\sin(k/2) e^{ik(n+1/2)} dk}{\omega^2(k)}.$$

To avoid the singularity at $k = 0$ we must pass it from below; all other roots of the equation $\omega^2(k) = 0$ inside the strip $-\pi \le \mathrm{Re} k \le \pi$ have nonzero imaginary parts. Closing the contour of integration in the upper half-plane for $n \ge 0$ and lower half-plane for $n < 0$, we obtain by residue theorem

$$w_n = \begin{cases} w_- + \displaystyle\sum_{k \in F^-} \frac{\sin(k/2) e^{ik(n+1/2)}}{\omega(k)\omega'(k)}, & n < 0, \\ w_- - \dfrac{1}{c^2} - \displaystyle\sum_{k \in F^+} \frac{\sin(k/2) e^{ik(n+1/2)}}{\omega(k)\omega'(k)}, & n \ge 0, \end{cases} \tag{5.2}$$

where $F^{\pm} = \{k : \omega^2(k) = 0, \mathrm{Im} k \gtrless 0, -\pi \le \mathrm{Re} k \le \pi\}$. Solutions satisfying the admissibility constraints

$$w_n \ge w_c \quad \text{for } n \le -1, \qquad w_n \le w_c \quad \text{for } n \ge 0 \tag{5.3}$$

form a family of lattice-trapped equilibria parametrized by the total stress in the chain $\sigma = c^2 w_- - 1$; the set of such stresses constitutes the *trapping region*.

To specify the trapping region we observe that in our static solutions the phase boundary is pinned at the site $n = -1$. If the strain profile (5.2) is monotone, which occurs, for example, when all long-range interactions are repulsive ($\Psi(p) < 0$ for $p \ge 2$[1]), the constraints (5.3) can be replaced by $w_0 \le w_c$ and $w_{-1} \ge w_c$. The trapping region can then be described explicitly:

$$\sigma_{\mathrm{M}} - \sigma_{\mathrm{P}} \le \sigma \le \sigma_{\mathrm{M}} + \sigma_{\mathrm{P}}, \tag{5.4}$$

where $\sigma_{\mathrm{M}} = c^2 w_c - 1/2$ is the Maxwell stress and

$$\sigma_{\mathrm{P}} = \frac{1}{2} + c^2 \sum_{k \in F^+} \frac{\sin(k/2) e^{ik/2}}{\omega(k)\omega'(k)} \tag{5.5}$$

[1]Since $\Psi(1) = 1 > 0$, the homogeneous phases can still be stable if the negative long-range moduli are sufficiently small.

is the *Peierls stress* (see also [4, 32]). The phase boundary remains trapped until the stress reaches one of the limiting values: $\sigma = \sigma_{\rm M} - \sigma_{\rm P}$, corresponding to $w_{-1} = w_c$ when the interface starts moving to the left ($V < 0$), or $\sigma = \sigma_{\rm M} + \sigma_{\rm P}$, corresponding to $w_0 = w_c$ when the interface starts moving to the right ($V > 0$). The two limiting solutions represent unstable equilibria from which the dynamic solution bifurcates.

6. Kinetic relation. The waves generated in the core of the moving phase boundary carry the energy away from the front without changing the average values of parameters at infinity. At the continuum level these lattice waves are invisible and therefore the associated energy transfer is perceived as dissipation. To evaluate the rate of dissipation, we start with the microscopic energy balance

$$\frac{d\mathcal{E}}{dt} = \mathcal{A}(t),$$

where $\mathcal{E}$ is the total energy of the chain and $\mathcal{A}(t)$ is the power supplied by the external loads. Since the solution of the discrete problem at infinity can be represented as a sum of the macroscopic contribution and the superimposed oscillations, we can split the averaged power accordingly. We obtain

$$\langle \mathcal{A} \rangle = \mathcal{P} - \mathcal{R}, \tag{6.1}$$

where $\mathcal{P} = \sigma_+ v_+ - \sigma_- v_-$ is the macroscopic rate of work and $\mathcal{R}$ is the energy release due to radiated waves which is invisible at the macroscale. While in the general case the expression for $\mathcal{R}$ may contain coupling terms, in the piecewise linear model adopted in this paper, the macroscopic and microscopic contributions decouple (see also [11, 24]). The dissipation rate $\mathcal{R}$ can be written as the sum of the contributions from the areas ahead and behind the front:

$$\mathcal{R}(V) = \mathcal{R}_+(V) + \mathcal{R}_-(V). \tag{6.2}$$

To specify the entries in the right-hand side, we observe that due to the exponential decay of the modes with complex wave numbers, the strain and velocity fields given by (4.3) and (4.12) have the following asymptotic representation at $\xi = \pm\infty$:

$$v(\xi) \approx v_0(\xi) + \sum_{k \in N_{\rm pos}(V)} v_k(\xi), \quad w(\xi) \approx w_0(\xi) + \sum_{k \in N_{\rm pos}(V)} w_k(\xi),$$

where

$$v_0(\xi) = \begin{cases} v_-, & \xi < 0, \\ v_+, & \xi > 0, \end{cases} \qquad w_0(\xi) = \begin{cases} w_-, & \xi < 0, \\ w_+ & \xi > 0, \end{cases}$$

are the homogeneous components and

$$v_k(\xi) = \begin{cases} -\dfrac{4V \sin(k/2)\cos(k(\xi - 1/2))}{L_k(k,V)}, & \xi < 0,\ k \in N^-_{\rm pos}(V), \\ \dfrac{4V \sin(k/2)\cos(k(\xi - 1/2))}{L_k(k,V)}, & \xi > 0,\ k \in N^+_{\rm pos}(V), \end{cases}$$

$$w_k(\xi) = \begin{cases} \dfrac{8\sin^2(k/2)\cos k\xi}{kL_k(k,V)}, & \xi < 0,\ k \in N^-_{\rm pos}(V), \\ -\dfrac{8\sin^2(k/2)\cos k\xi}{kL_k(k,V)}, & \xi > 0,\ k \in N^+_{\rm pos}(V), \end{cases} \tag{6.3}$$

are the oscillatory components. Here $N^{\pm}_{\rm pos}(V) \equiv \{k \in N^{\pm}(V) : k > 0\}$ (recall (4.5)), $N^{+}_{\rm pos}(V) \bigcup N^{-}_{\rm pos}(V) = N_{\rm pos}(V)$. Due to the asymptotic orthogonality of the linear modes, the terms in the right-hand side of (6.2) can be expressed as contributions due to individual modes. Since the energy flux associated with the linear mode k is the product of the average energy density $\langle \mathcal{G}_k \rangle$ and the relative velocity $|V_g - V|$ of the energy transport with respect to the moving front, we can write

$$\mathcal{R}_+(V) = \sum_{k \in N^{+}_{\rm pos}(V)} \langle \mathcal{G}_k \rangle_+ (V_g - V), \qquad \mathcal{R}_-(V) = \sum_{k \in N^{-}_{\rm pos}(V)} \langle \mathcal{G}_k \rangle_- (V - V_g). \tag{6.4}$$

Here $\langle \mathcal{G}_k \rangle_{\pm}$ is the average energy density carried by the wave with the wave number $k \in N^{\pm}_{\rm pos}(V)$. It is given by

$$\langle \mathcal{G}_k \rangle_{\pm} = \lim_{n \to \pm\infty} \frac{1}{2V\tau(k)} \int_{n - V\tau(k)}^{n} \Bigg[v_k^2(\xi) + c^2 (w_k(\xi))^2 - \sum_{p=1}^{q-1} B(p) \{ (w_k(\xi + p) - w_k(\xi))^2 + (w_k(\xi) - w_k(\xi - p))^2 \} \Bigg] d\xi,$$

where $B(p) = \frac{1}{2} \sum_{l=1}^{q-p} l \Psi(l + p)$ and $\tau(k) = 2\pi/\omega(k) = 2\pi/(Vk)$. Using (6.3), we obtain

$$\langle \mathcal{G}_k \rangle_{\pm} = \frac{8V^2 \sin^2(k/2)}{(L_k(k, V))^2},$$

which gives for the total energy flux

$$\mathcal{R}(V) = \sum_{k \in N^{+}_{\rm pos}(V)} \frac{4V \sin^2(k/2)}{k L_k(k, V)} - \sum_{k \in N^{-}_{\rm pos}(V)} \frac{4V \sin^2(k/2)}{k L_k(k, V)} = \sum_{k \in N_{\rm pos}(V)} \frac{4V \sin^2(k/2)}{|k L_k(k, V)|}.$$

Recalling the definition $\mathcal{R}(V) = G(V)V$ we can write the microscopic expression for the configurational force:

$$G(V) = \sum_{k \in N_{\rm pos}(V)} \frac{4 \sin^2(k/2)}{|k L_k(k, V)|}. \tag{6.5}$$

The function $G(V)$ is well-defined since both $L(k, V)$ and $N_{\rm pos}(V)$ depend on V in a known way. Comparing (6.5) with the macroscopic definition of the configurational force (1.4) we obtain

$$G = \frac{1}{2}(w_- + w_+) - w_c, \tag{6.6}$$

which can be interpreted geometrically as the area difference between two shaded triangles in Figure 2.1. Combining (6.5) and (6.6), we obtain exactly (4.10), which shows that (4.10) is indeed the desired kinetic relation and that micro and macro assessments of dissipation are compatible.

7. An example. To illustrate the general solution, consider a special case $q = 2$ (see Figure 7.1). The model is then fully characterized by a single dimensionless parameter,

$$\beta = 4\Psi(2)/\Psi(1),$$

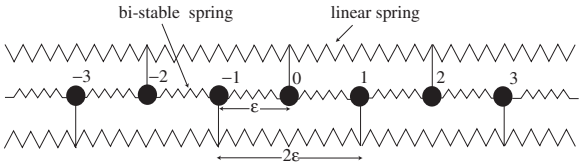


FIG. 7.1. *Discrete chain with nearest and next-to-nearest-neighbor interactions* ($q = 2$).

measuring the relative strength of NNN and NN interactions. The stability constraints (e.g., [36]) give $-1 < \beta \leq \infty$. Recalling that the inequality $\Psi(2) < 0$ is suggested by the linearization of the potentials of the Lennard–Jones type [32, 33], we can further restrict the admissible interval to

$$-1 < \beta \leq 0. \tag{7.1}$$

The total energy of the system can be written as

$$\mathcal{E} = \sum_{n=-\infty}^{\infty} \left[\frac{v_n^2}{2} + \frac{1+\beta}{2} w_n^2 - \theta(w_n - w_c)(w_n - w_c) - \frac{\beta}{4}(w_{n+1} - w_n)^2 \right]. \tag{7.2}$$

One can see that parameter $\beta/(1+\beta)$ characterizes the effect of discreteness: $\beta \sim 0$ corresponds to weak and $\beta \sim -1$ to strong coupling. This identification is compatible with the fact that at $\beta = 0$ the Peierls stress characterizing the size of the lattice-trapping domain takes the largest value (equal to the spinodal limit), while at $\beta = -1$ the Peierls stress is equal to zero.[2]

The energy (7.2) produces the following equation for the traveling waves:

$$\begin{aligned} V^2 w'' - \frac{\beta}{4}\Big(w(\xi+2) - 2w(\xi) + w(\xi-2) \Big) - w(\xi+1) + 2w(\xi) - w(\xi-1) \\ = 2\theta(-\xi) - \theta(-\xi-1) - \theta(1-\xi). \end{aligned} \tag{7.3}$$

The formal solution of this equation has been obtained in section 4. Below we provide detailed illustrations for the physically relevant range of parameters β.

7.1. Dispersion relation. To compute the strain and velocity profiles at a given V we need to find the nonzero roots k of the dispersion relation

$$L(k, V) = 4\sin^2(k/2) + \beta \sin^2 k - V^2 k^2 = 0. \tag{7.4}$$

It is convenient to present the complex roots explicitly as $k = k_1 + ik_2$ and divide them into three categories: real, responsible for radiation; purely imaginary, providing the monotone structure of the core region; and complex with nonzero real part, describing oscillatory contributions to the core.

Since $L(k, V)$ is an even function of k, the real roots appear in pairs $k = \pm k_1$. Assuming positive V, we obtain

$$V(k_1) = \frac{\sqrt{4\sin^2(k_1/2) + \beta \sin^2 k_1}}{|k_1|}.$$

This function is plotted in Figure 7.2(a). An infinite number of local maxima on

[2]The picture emerging in our piecewise linear model is somewhat obscured by the fact that in the limit of strong coupling the macroscopic sound speed tends to zero.

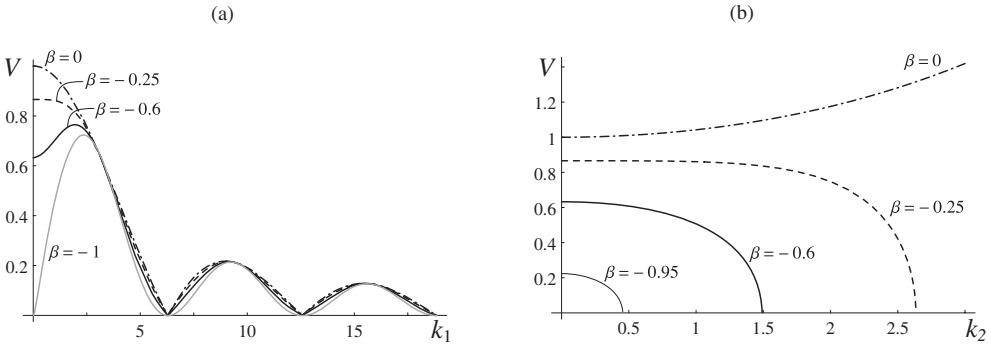


FIG. 7.2. *Real* (a) *and imaginary* (b) *roots of the dispersion relation* $L(k, V) = 0$ *at different* β.

this graph, denoted by $V = V_i$, correspond to resonance velocities: at these points $L_k(k, V) = 0$ and the sums in (4.3), (4.9), and (4.12) diverge. Between the resonance velocities, (7.4) possesses a finite number of positive real roots corresponding to propagating waves. To determine whether these waves propagate ahead or behind the front, we need to check whether $kL_k(k, V) = 2k^3V(k)V'(k)$ is positive or negative. At $V > 0$ the radiation conditions say that the waves with $kV'(k) > 0$ propagate in front of the phase boundary, while the waves with $kV'(k) < 0$ propagate behind.

Changing β affects the function $V(k)$ noticeably only at long waves (small k). A straightforward computation shows that $V(0)$ is equal to the macroscopic sound speed $c = (1 + \beta)^{1/2}$. We also obtain that $V'(0) = 0$ and

$$V''(0) = -\frac{1 + 4\beta}{12\sqrt{1 + \beta}}.$$

At $-1/4 < \beta \le 0$, the function $V(k)$ has a maximum at $k = 0$ while at $-1 < \beta < -1/4$ it has a local minimum implying that sufficiently strong coupling ($\beta < -1/4$) creates the possibility for the lattice waves to move faster than the macroscopic sound speed. The range of supersonic speeds increases as $\beta \to -1$, and in the limiting case $\beta = -1$ all propagating waves are macroscopically supersonic. It is interesting that the critical value $\beta = -1/4$ also emerges in the strain-gradient approximation of the energy (7.2), where it corresponds to the change of sign of the coefficient in front of the strain gradient term [15, 28]. In this approximation the dispersion relation $V(k)$ is replaced by a parabola: for $\beta > -1/4$ (weak nonlocality) the parabola is directed downward and the strain-gradient coefficient is negative while for $\beta < -1/4$ (strong nonlocality) the parabola is upward and the strain-gradient contribution to the energy is positive definite. The latter implies that subsonic phase boundaries can only be dissipation free. To yield a nontrivial kinetic relation in this range of parameters the quasicontinuum model must be augmented by higher-order terms [37].

The purely imaginary roots of (7.4) appear in symmetric pairs and correspond to nonoscillatory modes exponentially decreasing away from the front. By solving $L(ik_2, V) = 0$ for V we obtain

$$V(k_2) = \frac{\sqrt{4\sinh^2(k_2/2) + \beta \sinh^2 k_2}}{|k_2|}.$$

This function is shown in Figure 7.2b. One can show that $V(0) = c$, $V'(0) = 0$, and $V''(0) = (1 + 4\beta)/(12\sqrt{1 + \beta})$. For $-1 < \beta < -1/4$ the maximum of the curve

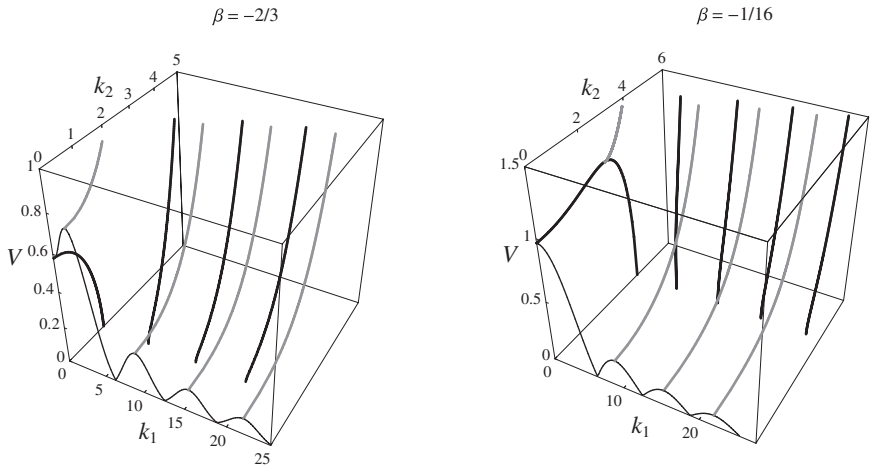


FIG. 7.3. *The structure of nonzero roots of* $L(k,V)=0$ *in the cases of strong* ($\beta=-2/3$) *and weak* ($\beta=-1/16$) *nonlocality. Thin lines: real roots, set N; thick lines: P-roots; gray lines: Q-roots.*

$V(k_2)$ is reached at $k_2=0$, which means that in the case of strong coupling only macroscopically subsonic phase boundaries have monotone contribution to the core structure. Both the value $V(0)$ and the range of available wave numbers decrease as $\beta\to-1$, so that in the limit purely imaginary roots disappear. In the case $\beta=0$ the function $V(k_2)$ is convex and no imaginary roots contribute to the subsonic solution. This is compatible with the fact that in the degenerate NN limit the static interface ($V=0$) is atomically sharp.

Complex roots with nonzero real part contribute to the oscillatory structure of the core region. For the given V the real (k_1) and imaginary (k_2) parts of the relevant wave numbers satisfy the system of two equations: $\mathrm{Re}L(k_1+ik_2,V)=0$ and $\mathrm{Im}L(k_1+ik_2,V)=0$. The set of complex roots contains infinitely many branches that come in symmetric quadruples. The first quadrant of the complex plane is shown in Figure 7.3. The complex roots can be divided into two sets: Q and P. The set Q (thick gray lines), has a purely dynamic nature and contributes to the boundary layers around the front only at nonzero V. The set P, shown in Figure 7.3 by thick black lines, contains purely imaginary branches which intersect the plane $V=0$ and contribute to the static solution: at $V=0$ they are given by

$$k=2\pi n\pm i\lambda,\qquad \lambda=2\mathrm{arccosh}\left[\frac{1}{\sqrt{|\beta|}}\right], \tag{7.5}$$

where n is an integer [33]. As β tends to zero, the imaginary parts of P-roots approach $\pm\infty$; the eventual disappearance of these roots in the limit $\beta\to0$ is responsible for the sharpening of the front in the NN approximation.

7.2. Strain and velocity profiles. Typical profiles of strain $w(\xi)$ and velocity $v(\xi)$ computed for the NNN model from (4.3), (4.12) are shown in Figure 7.4, where $\beta=-0.2$. For this case the first two resonance velocities are $V_1=0.2164$ and $V_2=0.1282$. Accordingly, at $V=0.5>V_1$ we see only one radiative mode propagating behind the phase boundary; at $V_2<V=0.16<V_1$ the solution exhibits two additional radiative modes, one propagating behind and one in front of the phase boundary.

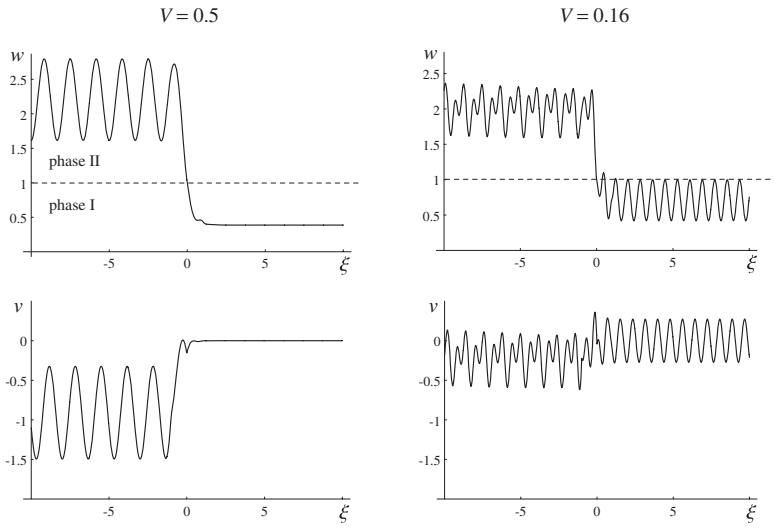


FIG. 7.4. *Strain and velocity profiles at $V > V_1$ and $V_2 < V < V_1$. Here $\beta = -0.2$, $w_c = 1$, $v_+ = 0$.*

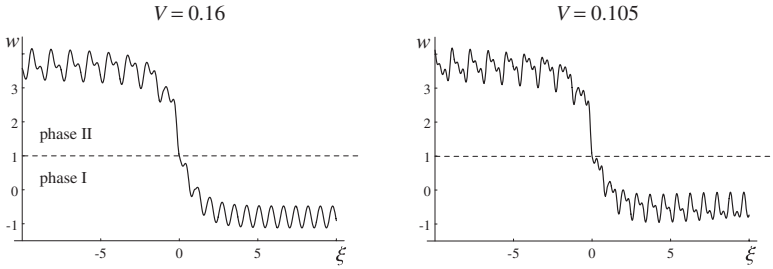


FIG. 7.5. *Strain profiles at $V_2 < V < V_1$ and $V_3 < V < V_2$. Here $\beta = -0.75$, $w_c = 1$.*

A closer inspection of the solutions at $V < 0.266$ reveals a violation of the constraints (3.5): in the strain profile corresponding to $V = 0.16$ in Figure 7.4 the threshold $w = w_c$ is crossed at both $\xi = 0$ and $\xi > 0$. Moreover, for this value of β our numerical computations suggest that the entire velocity interval $(0, 0.266)$ around the resonances has to be excluded as inadmissible. Similar "velocity gaps" also have been detected in [11, 12, 14] for the semilinear Frenkel–Kontorova problem.

At larger β steady interface propagation becomes possible in certain subcritical velocity intervals. For instance, at $\beta = -0.75$ we found admissible traveling wave solutions in the intervals: $[0.24, 0.5]$ (between $V_1 = 0.215$ and $c = 0.5$); $[0.142, 0.19]$ (between V_1 and $V_2 = 0.1279$); $[0.1, 0.11]$ (between V_2 and $V_3 = 0.0912$); $[0.078, 0.08]$ (between V_3 and $V_4 = 0.0708$); and possibly in some shorter intervals at smaller V. Two such solutions are shown in Figure 7.5. The first one corresponds to $V = 0.16$, which is between the first and second resonances; unlike its counterpart at $\beta = -0.2$, this solution is admissible. The second admissible profile corresponds to the value of velocity $V = 0.105$ located between the second and third resonances. In this case there are five radiative modes, two in front and three behind the phase boundary. The appearance of the small-velocity intervals of existence of the traveling wave solutions

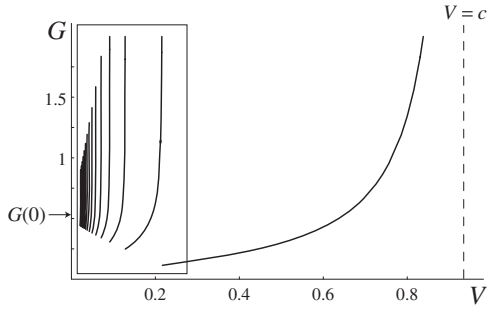


FIG. 7.6. *Kinetic relation $G(V)$ for $\beta = -1/8$. The region inside the rectangle should be excluded because the corresponding solutions violate admissibility constraints* (3.5).

at nonzero β is due to the presence of P-roots at nonzero β: as β grows in absolute value, these roots move closer to the real axis, widening the transition layer and suppressing oscillations due to the Q-roots (see Figure 7.3).

7.3. Kinetic relation. Using (6.5) and the known dispersion spectrum, we can now explicitly evaluate the kinetic relation $G(V)$. A representative example is shown in Figure 7.6. At resonance velocities the configurational force required to move the interface tends to infinity. The singularities are due to equal curvatures of the energy wells ensuring that the energy transport is simultaneously blocked in both phases. The main physical reasons are related to low dimensionality of the model and the absence of microscopic dissipation.

As we discussed above, at sufficiently small β the entire region around the small-velocity resonances has to be excluded because the corresponding solutions violate the admissibility constraints (3.5). With β increasing, some of the small-velocity solutions between the resonances become admissible, as shown in Figure 7.7(b)–(d). Observe also that there is an infinite number of β at which the sonic speed c coincides with one of the resonance velocity. Thus, at $\beta = -0.9539$ we have $c = V_1 = 0.2147$, implying that for $\beta \le -0.9539$ the subsonic region lies below the first resonance (see Figure 7.7(d)). Overall, the total domain of existence of the traveling wave solutions shrinks as $\beta \to -1$, while the domain of admissible traveling waves between the resonances expands.

Zero-velocity limit. To check the compatibility of static and dynamic branches of solutions it is instructive to trace the zero velocity limit of our dynamic theory. At $V = 0$ we can use (5.2) with $F^{\pm} = \{\pm i\lambda\}$, where λ is defined in (7.5). After some algebraic manipulations, the family of lattice-trapped equilibria (5.2) can be represented in the form

$$w_n = \begin{cases} \dfrac{\sigma+1}{1+\beta} - \dfrac{e^{\lambda(n+1/2)}}{2(1+\beta)\cosh(\lambda/2)}, & n < 0, \\ \dfrac{\sigma}{1+\beta} + \dfrac{e^{-\lambda(n+1/2)}}{2(1+\beta)\cosh(\lambda/2)}, & n \ge 0, \end{cases} \tag{7.6}$$

where σ lies in the region (5.4). The solutions (7.6) can be shown to be metastable [32]. The expression for the Peierls stress (5.5) marking the threshold of metastability can now be written explicitly as

$$\sigma_{\mathrm{P}} = \frac{1}{2}\sqrt{1+\beta}.$$

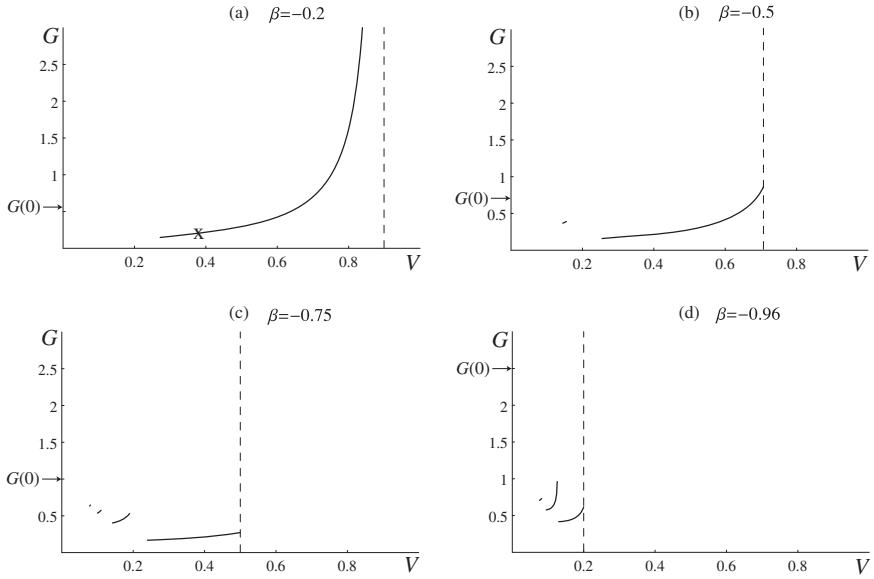


FIG. 7.7. *Kinetic relations $G(V)$ for different β. The point marked by* x *in* (a) *corresponds to the numerical results in Figures* 8.1 *and* 8.2.

Observe that at $\beta = 0$ the Peierls stress coincides with the spinodal stress $\sigma_{\rm S} = 1/2$. As β grows, the trapping region becomes narrower and eventually disappears at $\beta = -1$ ($\sigma_{\rm P} = 0$). The upper boundary of the trapping region (5.4) corresponds to the case when $w_0 = w_c$, which is exactly the condition (3.4). The corresponding saddle-point configuration is given by

$$w_n = \lim_{V\to 0} w(n - Vt) = \begin{cases} w_c + \dfrac{e^{\lambda/2} - e^{\lambda(n+1/2)}}{2(1+\beta)\cosh(\lambda/2)}, & n < 0, \\ w_c + \dfrac{e^{-\lambda(n+1/2)} - e^{-\lambda/2}}{2(1+\beta)\cosh(\lambda/2)}, & n \geq 0. \end{cases} \tag{7.7}$$

Using this solution, we can obtain the value of the configurational force at the depinning limit:

$$G(0) = \frac{1}{2}(w_- + w_+) - w_c = \frac{1}{2\sqrt{1+\beta}}. \tag{7.8}$$

Notice that although the Peierls stress tends to zero when $\beta \to -1$, the corresponding value of the configurational force $G(0)$ tends to infinity. This is, of course, an artifact of our specific assumptions concerning the elastic moduli and is due to the divergence of the macroscopic transformation strain in the limit.

Sonic limit. The qualitative behavior of the function $G(V)$ in the limit $V \to c$ depends on β. If $-1/4 < \beta \leq 0$, and $V \lesssim c$, the wave spectrum contains a single wave number k which approaches zero as $V \to c$. Expanding the expression for the configurational force (6.5) at small k we obtain

$$G \sim \frac{6}{(1+4\beta)k^2},$$

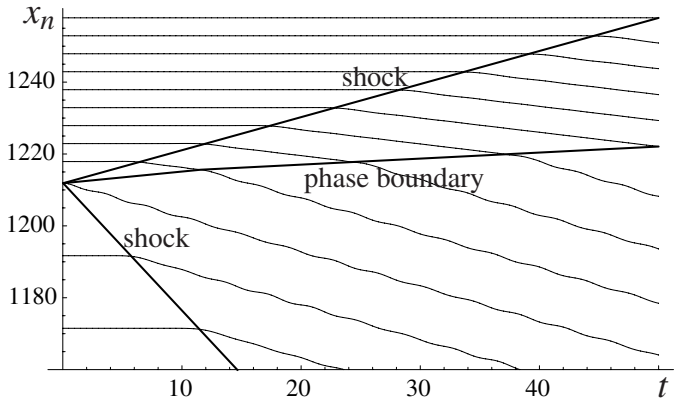


FIG. 8.1. *Positions $x_n(t)$ of every fifth particle in the interval $290 \le n \le 345$ in numerical solution of the Riemann problem with initial data compatible with the traveling wave at $V = 0.375$ at $\beta = -0.2$. The phase boundary is initially placed at $n_0 = 300$, and the problem is solved on the interval $1 \le n \le 600$. The corresponding point on the kinetic relation $G(V)$ is marked by* x *in Figure* 7.7a.

which implies that $G(V) \to \infty$ as $V \to c$ (see Figures 7.6 and 7.7(a)). The picture is qualitatively different when $-1 < \beta < -1/4$. In this case as V approaches c from below, the limit of the corresponding real wave number k is nonzero k_s, and therefore configurational force $G(V)$ remains finite (see Figure 7.7(b)–(d)).

8. Stability of the traveling waves. We now present some numerical experiments aimed at accessing stability of the admissible traveling wave solutions. To simplify the consideration of the transient regimes for the system (2.8), we consider Riemann initial data of the form

$$(w_n, v_n)\big|_{t=0} = \begin{cases} (w_-^0, 0), & n < n_0, \\ (w_c, 0), & n = n_0, \\ (w_+^0, 0), & n > n_0. \end{cases} \tag{8.1}$$

We assume that $w_+^0 < w_c$ and $w_-^0 > w_c$. The analysis of the corresponding continuum problem for the p-system (1.2) suggests the formation of a phase boundary with two shocks in front and behind (e.g., [17]). This is indeed what we see in Figure 8.1, which shows a typical numerical solution of (2.8) obtained using the Verlet algorithm on a large domain.

After a transient period, the phase boundary starts moving with a constant speed. To check convergence of the non-steady-state problem to the admissible traveling wave solution we used the following algorithm. For a traveling wave solution with velocity V, (4.9) provides the average strains at both sides of the discontinuity $w_\pm(V)$. We can then use (4.13) and the Rankine–Hugoniot jump conditions across the shocks to compute the corresponding initial strains $w_\pm^0$ in terms of V and v_+. Without loss of generality, we choose v_+ so that $w_+^0 = 0$ and obtain an explicit formula relating the Riemann data with the observed phase boundary velocity:

$$w_-^0(V) = w_-(V) + w_+(V) + \frac{V}{c(c^2 - V^2)} = 2\left(w_c + \sum_{k \in N_{\rm pos}(V)} \frac{4\sin^2(k/2)}{|kL_k(k,V)|}\right) + \frac{V}{c(c^2 - V^2)}.$$

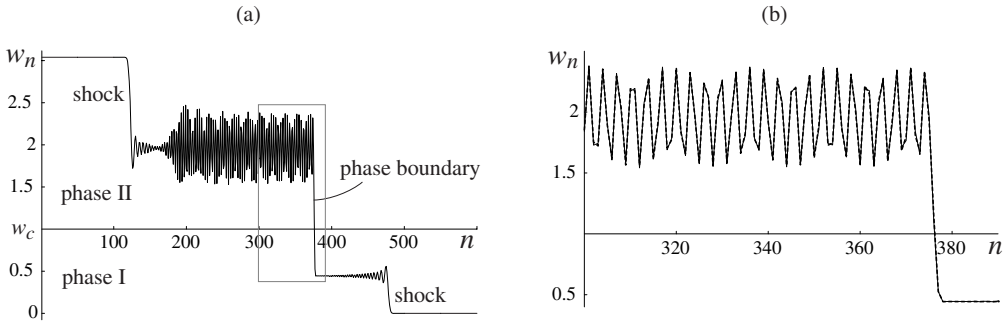


FIG. 8.2. (a) *Strain profile* $w_n(200)$ *for the numerical solution shown in Figure* 8.1. (b) *The same solution (solid line) zoomed in around the phase boundary (inside the rectangle in* (a)*) and compared to the analytical traveling wave solution (thick dashed line).*

Computations based on this relation consistently indicate that the fast branch of the kinetic relation ($V > V_1$) corresponds to traveling wave solutions with a finite domain of attraction. An example corresponding to $\beta = -0.2$ and $V = 0.375$ is shown in Figures 8.1 and 8.2. One can see not only that the generated phase boundary propagates steadily with the predicted velocity but that the strain profile $w_n(t)$ zoomed around the phase boundary (solid line in Figure 8.2(b)) compares perfectly with the analytical solution (4.3) (thick dashed line). The above analysis suggests that the fast branch of the kinetic relation with $V > V_1$, corresponding to traveling wave solutions with a single oscillatory mode behind and monotonic leading edge, is locally stable; proving this conjecture rigorously is highly nontrivial.

Unfortunately, we were not able to find similar evidence of numerical stability for admissible traveling waves with $V < V_1$ exhibiting oscillations both behind and in front of the phase boundary. Numerical simulation for the data expected to converge to the particular traveling wave lead instead to a solution which does not agree with the traveling wave ansatz although the phase boundary propagates steadily (with a slightly smaller velocity). The macroscopic solution is very close to the corresponding traveling wave; the analytical formula captures the core structure but not the oscillations. One can conjecture that while an admissible traveling wave with radiation on both sides of the front is not a global attractor, it is surprisingly close to one.

Finally, we refer to [17] for a related study of the phase boundary stability in the context of a different discretization of the mixed-type p-system.

9. Conclusions. In this paper we used a physically motivated discretization of the p-system to derive an explicit kinetic relation for a one-dimensional theory of martensitic phase transitions. The macroscopic dissipation was interpreted as the energy of the lattice waves emitted by a moving macroscopic discontinuity. By using the simplest piecewise linear model, we obtained an explicit formula for the continuum rate of entropy production which depends only on interatomic potentials. We showed that despite the difference in the structure of micro and macro theories, the assessments of dissipation at different scales are fully compatible. The present study complements previous analyses of related systems in fracture and plasticity framework by including general harmonic long-range interactions.

More specifically, we showed that contrary to the simplest theory with NN interactions, which has a mean field character and is therefore degenerate, strongly nonlocal models produce multivalued kinetic relations with several admissible branches and

rich variety of configurations of emitted lattice waves. In addition to enlarging the domain of existence of steady-state regimes, sufficiently strong long-range interactions significantly alter the structure of mobility curves near sonic speeds. The nonlocality also affects the size of lattice trapping: as long-range interactions become stronger, the trapping region reduces in size in terms of stress. At the same time, it widens in our model in terms of driving forces, which emphasizes an important difference between the physical and configurational descriptions. Although the main effects of nonlocality were illustrated in the paper by the explicit computations for the NNN model, we have also conducted a similar study of the NNNN model which showed qualitatively similar behavior.

The present work was motivated by similar studies of the semilinear discrete Frenkel–Kontorova model (e.g., [11, 12]). While the two models turn out to be equivalent in static and overdamped limits [32, 34], the fully inertial versions are quite different. An additional level of complexity in the quasilinear model considered here is associated with a different structure of nonlinearity that results in the presence of the limiting characteristic velocity, microscopic and macroscopic particle velocities, and the discrete Rankine–Hugoniot jump conditions.

REFERENCES

[1] R. Abeyaratne and J. Knowles, *A continuum model of a thermoelastic solid capable of undergoing phase transitions*, J. Mech. Phys. Solids, 41 (1993), pp. 541–571.

[2] W. Atkinson and N. Cabrera, *Motion of a Frenkel-Kontorova dislocation in a one-dimensional crystal*, Phys. Rev. A, 138 (1965), pp. 763–766.

[3] O. M. Braun and Y. S. Kivshar, *Nonlinear dynamics of the Frenkel-Kontorova model*, Phys. Rep., 306 (1998), pp. 1–108.

[4] O. M. Braun, Y. S. Kivshar, and I. I. Zelenskaya, *Kinks in the Frenkel-Kontorova model with long-range interparticle interactions*, Phys. Rev. B, 41 (1990), pp. 7118–7138.

[5] A. Carpio and L. L. Bonilla, *Oscillatory wave fronts in chains of coupled nonlinear oscillators*, Phys. Rev. E, (2003), p. 056621.

[6] V. Celli and N. Flytzanis, *Motion of a screw dislocation in a crystal*, J. Appl. Phys., 41 (1970), pp. 4443–4447.

[7] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, Heidelberg, 2000.

[8] J. Ericksen, *Equilibrium of bars*, J. Elasticity, 5 (1975), pp. 191–202.

[9] Y. Gaididei, N. Flytzanis, A. Neupera, and F. G. Mertensa, *Effect of non-local interactions on soliton dynamics in anharmonic chains: Scale competition*, Phys. D, 107 (1997), pp. 83–111.

[10] T. Y. Hou and P. Lax, *Dispersive approximations in fluid dynamics*, Comm. Pure Appl. Math., 44 (1991), pp. 1–40.

[11] O. Kresse and L. Truskinovsky, *Mobility of lattice defects: Discrete and continuum approaches*, J. Mech. Phys. Solids, 51 (2003), pp. 1305–1332.

[12] O. Kresse and L. Truskinovsky, *Lattice friction for crystalline defects: From dislocations to cracks*, J. Mech. Phys. Solids, 52 (2004), pp. 2521–2543.

[13] P. G. LeFloch, *Hyperbolic Systems of Conservation Laws*, ETH Lecture Note Series, Birkhäuser, Boston, 2002.

[14] M. Marder and S. Gross, *Origin of crack tip instabilities*, J. Mech. Phys. Solids, 43 (1995), pp. 1–48.

[15] R. D. Mindlin, *Second gradient of strain and surface tension in linear elasticity*, Internat. J. Solids Structures, 1 (1965), pp. 417–438.

[16] S.-C. Ngan and L. Truskinovsky, *Thermal trapping and kinetics of martensitic phase boundaries*, J. Mech. Phys. Solids, 47 (1999), pp. 141–172.

[17] S.-C. Ngan and L. Truskinovsky, *Thermo-elastic aspects of dynamic nucleation*, J. Mech. Phys. Solids, 50 (2002), pp. 1193–1229.

[18] M. Peyrard, *Simple theories of complex lattices*, Phys. D, 123 (1998), pp. 403–424.

[19] M. Peyrard, S. Pnevmatikos, and N. Flytzanis, *Discreteness effects on non-topological kink soliton dynamics in nonlinear lattices*, Phys. D, 19 (1986), pp. 268–281.
[20] D. Serre, *Systems of Conservation Laws*, vols. 1, 2, Cambridge University Press, Cambridge, UK, 1999.
[21] M. Slemrod, *Admissibility criteria for propagating phase boundaries in a van der Waals fluid*, Arch. Ration. Mech. Anal., 81 (1983), pp. 301–315.
[22] L. I. Slepyan, *Dynamics of a crack in a lattice*, Soviet Phys. Dokl., 26 (1981), pp. 538–540.
[23] L. I. Slepyan, *The relation between the solutions of mixed dynamical problems for a continuous elastic medium and a lattice*, Soviet Phys. Dokl., 27 (1982), pp. 771–772.
[24] L. I. Slepyan, *Models and Phenomena in Fracture Mechanics*, Springer-Verlag, New York, 2002.
[25] L. I. Slepyan, A. Cherkaev, and E.Cherkaev, *Transition waves in bistable structures.* II. *Analytical solution: wave speed and energy dissipation*, J. Mech. Phys. Solids, 53 (2005), pp. 407–436.
[26] L. I. Slepyan and L. V. Troyankina, *Fracture wave in a chain structure*, J. Appl. Mech. Tech. Phys., 25 (1984), pp. 921–927.
[27] M. Toda, *Theory of nonlinear lattices*, Springer-Verlag, Berlin, 1989.
[28] N. Triantafyllidis and S. Bardenhagen, *The influence of scale size on the stability of periodic solids and the role of associated higher order gradient continuum models*, J. Mech. Phys. Solids, 44 (1996), pp. 1891–1928.
[29] L. Truskinovsky, *Equilibrium interphase boundaries*, Soviet Phys. Dokl., 27 (1982), pp. 306–331.
[30] L. Truskinovsky, *Dynamics of nonequilibrium phase boundaries in a heat conducting elastic medium*, J. Appl. Math. Mech., 51 (1987), pp. 777–784.
[31] L. Truskinovsky, *Kinks versus shocks*, in IMA Series in Mathematics and Its Applications, E. D. R. Fosdick and M. Slemrod, eds., Vol. Math. Appl. 52, Springer-Verlag, 1993, pp. 185–229.
[32] L. Truskinovsky and A. Vainchtein, *Peierls-Nabarro landscape for martensitic phase transitions*, Phys. Rev. B, 67 (2003), p. 172103.
[33] L. Truskinovsky and A. Vainchtein, *The origin of nucleation peak in transformational plasticity*, J. Mech. Phys. Solids, 52 (2004), pp. 1421–1446.
[34] L. Truskinovsky and A. Vainchtein, in preparation.
[35] L. Truskinovsky and A. Vainchtein, *Explicit kinetic relation from "first principles,"* in Advances in Mechanics and Mathematics, vol. 11, P. Steinmann and G. Maugin, eds., Springer, New York, 2005.
[36] L. Truskinovsky and A. Vainchtein, *Quasicontinuum modeling of short-wave instabilities in crystal lattices*, Philosophical Magazine, to appear.
[37] L. Truskinovsky and A. Vainchtein, *Quasicontinuum models of dynamic phase transitions*, Continuum Mechanics and Thermodynamics, submitted.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 554–572

SEMICONDUCTOR SIMULATIONS USING A COUPLED QUANTUM DRIFT-DIFFUSION SCHRÖDINGER–POISSON MODEL*

ASMA EL AYYADI† AND ANSGAR JÜNGEL†

Abstract. A coupled quantum drift-diffusion Schrödinger–Poisson model for stationary resonant tunneling simulations in one space dimension is proposed. In the ballistic quantum zone with the resonant quantum barriers, the Schrödinger equation is solved. Near the contacts, where collisional effects are assumed to be important, the quantum drift-diffusion model is employed. The quantum drift-diffusion model was derived by a quantum moment method from a collisional Wigner equation by Degond et al. [*J. Statist. Phys.*, 118 (2005), pp. 625–665]. The derivation yields an $O(\hbar^4)$ approximation of the equilibrium Wigner function which is used as the "alimentation function" in the mixed-state formula for the electron and current densities at the interface. The coupling of the two models is realized by assuming the continuity of the electron and current densities at the interface points. Current-voltage characteristics of a one-dimensional tunneling diode are numerically computed. The results are compared to those from the three models: quantum drift-diffusion equations, the Schrödinger–Poisson system, and the coupled drift-diffusion Schrödinger–Poisson equations.

Key words. Schrödinger–Poisson system, quantum drift-diffusion model, quantum microscopic-macroscopic coupling, finite differences, resonant tunneling diode, hysteresis

AMS subject classifications. 65M06, 82D37, 81V99

DOI. 10.1137/040610805

1. Introduction.

Quantum effects in modern semiconductor devices are becoming of increasing importance in VLSI design. Devices which are based on quantum effects, like resonant tunneling diodes, can be used in logic applications [19] and are expected to improve the system performance of multi-GHz circuits in wireless communication systems [27]. Resonant tunneling diodes can be modeled by the Wigner equation [23] or the mixed-state Schrödinger equation [26]. However, the numerical computation of these equations is very expensive, even in one space dimension. Therefore, macroscopic quantum equations like quantum hydrodynamic or quantum drift-diffusion models have been devised [1, 15], whose numerical solution is much cheaper than solving microscopic models [18, 20, 22]. On the other hand, quantum diffusion models do not always give sufficiently physical accurate solutions [7].

In order to meet both demands (physical accuracy and numerical efficiency), *coupled microscopic-macroscopic models* can be employed. In these models, a microscopic quantum description is used in regions with dominant quantum effects, and a macroscopic (fluid-type) model is employed in subregions in which collisional effects are expected to be dominant. In the case of a resonant tunneling diode, it was proposed in [10] to use the stationary mixed-state Schrödinger equation in the (ballistic) channel region and the stationary drift-diffusion equations in the diffusion zones near the contacts. This approach has two advantages. First, the spatial domain in which the Schrödinger equation is solved can be reduced, thus also reducing the computational effort. Second, as the diffusion zone is assumed to be collision dominated, a diffusion

*Received by the editors June 30, 2004; accepted for publication (in revised form) July 25, 2005; published electronically December 30, 2005. The authors were partially supported by a European Union project, grant HPRN-CT-2002-00282, the DAAD-Procope program, and Deutsche Forschungsgemeinschaft (DFG) grant JU359/3. The second author was supported by DFG grant JU359/5.
http://www.siam.org/journals/siap/66-2/61080.html
†Fachbereich Mathematik und Informatik, Universität Mainz, Staudingerweg 9, 55099 Mainz, Germany (elayyadi@mathematik.uni-mainz.de, juengel@mathematik.uni-mainz.de).

approximation of the Wigner equation leads naturally to a coupling strategy between the quantum and classical equations.

Similar coupling approaches have been proposed in the literature. A coupled kinetic-quantum model was introduced in [4]. More precisely, a Boltzmann equation is solved in the classical zone and the stationary Schrödinger equation is computed in the quantum zone. At the interface between the classical and quantum zones, the boundary conditions for the Boltzmann equation depend on the reflection and transmission coefficients of the Schrödinger solution. The distribution function solving the Boltzmann equation is used as an "alimentation function" in the definition of the electron density in the quantum region. A time-dependent classical-quantum coupling strategy was studied in [5]. Employing the drift-diffusion model with interface conditions from a diffusion approximation leads to the mentioned approach of [10]. The coupled drift-diffusion Schrödinger model was recently extended to include collisions via a Pauli master equation [2]. A model in which a classical transport is assumed in the direction parallel to the electron gas and a quantum description in the transversal direction is analyzed in [6]. In [14] the quantum drift-diffusion model is used in the parallel direction instead of a classical transport description. For other coupling models, see, e.g., [3, 8].

In this paper we propose a slightly different approach compared to [10]. Instead of choosing a classical collision operator in the Boltzmann equation (from which the drift-diffusion model is derived) we start from the Wigner equation with a Bhatganar-Gross-Krook (BGK) collision operator. In [11] it was shown that a diffusion approximation of the Wigner–BGK model leads to the so-called quantum drift-diffusion model (also called the density-gradient model [1]),

$$\frac{\partial n}{\partial t} - \frac{1}{e}\operatorname{div} J = 0, \quad J = \mu_n \left(U_{\mathrm{th}} \nabla n - n \nabla (V + Q[n])\right),$$

where the variables are the electron density $n(x,t)$, the current density $J(x,t)$, and the electrostatic potential $V(x,t)$; the physical constants are the elementary charge e, the electron mobility $\mu_n = e\tau/m$, the effective electron mass m, the momentum relaxation time τ, and the thermal voltage U_{th}. The expression

$$Q[n] = \frac{\hbar^2}{6em} \frac{\Delta\sqrt{n}}{\sqrt{n}}$$

denotes the quantum Bohm potential, where $\hbar$ is the reduced Planck constant. The fourth-order parabolic quantum drift-diffusion model was analyzed and numerically solved in [22].

More precisely, we use the quantum drift-diffusion model in the diffusion region and the mixed-state Schrödinger equation in the ballistic zone. We restrict ourselves to the spatial one-dimensional stationary case in order to avoid complicated topological conditions on the device geometry. The advantage of our approach is that no (artificial) separation of the quantum and classical zones is necessary since a quantum description is employed in the whole device. The coupled model is solved self-consistently with the Poisson equation.

The coupling of the models is realized through connection conditions relating the macroscopic variables, namely, the electron density and the current density, at the interface boundary points. We suppose that the particle and current densities are continuous across the interface. The current density computed from the Schrödinger equation depends on the statistics (or alimentation function) used in the mixed-state

formula. At the interface we assume that the statistics of the incoming particles equal the $O(\hbar^4)$ approximation of the so-called quantum Maxwellian, which is related to the quantum drift-diffusion model. This yields nonlinear boundary conditions for the macroscopic electron density and its derivatives.

The coupled model is numerically implemented using a finite-difference discretization and tested against a test case for a one-dimensional resonant tunneling diode taken from [26]. The numerical results show negative differential resistance in the current-voltage characteristic at room temperature, whereas the quantum drift-diffusion model in the whole domain is not able to reproduce these effects at room temperature (with the physical effective electron mass). Furthermore, hysteresis in the current-voltage curve can be observed in computations from our coupled model but not from the quantum drift-diffusion model. Compared to the numerical solution of the Schrödinger–Poisson (SP) system in the whole domain, the numerical effort of the coupled model is significantly reduced with comparable numerical solutions.

The paper is organized as follows. In the next section, the Schrödinger equation with open boundary conditions is presented and a sketch of the derivation of the quantum drift-diffusion model following [11] is given. Furthermore, the coupling of the two models is explained. Section 3 is concerned with the numerical discretization of the equations and the iteration procedure. Finally, some numerical results for a one-dimensional resonant tunneling diode are presented in section 4.

2. Presentation of the models. The semiconductor is assumed to occupy the interval $\Omega = (0, L)$ in which the ballistic quantum zone $\Omega_s = (x_1, x_2)$ is sandwiched between two quantum diffusion regions $\Omega_q = (0, x_1) \cup (x_2, L)$ and $0 < x_1 < x_2 < L$.

2.1. The Schrödinger model. We consider the Schrödinger equation in the interval (a, b), where $a = 0$, $b = L$ or $a = x_1$, $b = x_2$. In the first case we solve the Schrödinger equation in the whole semiconductor domain; in the latter case we solve only in the ballistic quantum zone.

Let the electrostatic potential $V(x)$ be given and let $\widetilde{V} = V + V_{\rm ext}$ be the sum of electrostatic and external potential ($V_{\rm ext}$ models, for instance, the double barriers). We solve the Schrödinger equation

$$-\frac{\hbar}{2}\frac{d}{dx}\left(\frac{1}{m}\frac{d\psi_p}{dx}\right) - e\widetilde{V}(x)\psi_p = E_p\psi_p, \quad x \in (a,b),\ p \in \mathbb{R}, \tag{2.1}$$

where $\hbar = h/2\pi$ is the reduced Planck constant, m the (generally position-dependent) effective mass of the electrons, e the elementary charge, and E_p is the total energy of the corresponding scattering state ψ_p, given by

$$E_p = \begin{cases} p^2/2m - e\widetilde{V}(a) & : \ p > 0, \\ p^2/2m - e\widetilde{V}(b) & : \ p < 0. \end{cases}$$

Here, $p = \hbar k$ is the crystal momentum and k the wave vector. We use the Lent–Kirkner boundary conditions for (2.1) [16, 24],

$$\hbar\psi_p'(a) + ip\psi_p(a) = 2ip, \quad \hbar\psi_p'(b) = ip_+(p)\psi_b(b) \quad \text{if } p > 0, \tag{2.2}$$

$$\hbar\psi_p'(b) - ip\psi_p(b) = -2ip, \quad \hbar\psi_p'(a) = -ip_-(p)\psi_p(a) \quad \text{if } p < 0, \tag{2.3}$$

where

$$p_\pm(p) = \sqrt{p^2 \pm 2em(\widetilde{V}(b) - \widetilde{V}(a))}. \tag{2.4}$$

These boundary conditions can be derived by solving the above Schrödinger equation in $\mathbb{R}$, extending the potential by the definitions $\widetilde{V}(x) = \widetilde{V}(a)$ for $x < a$ and $\widetilde{V}(x) = \widetilde{V}(b)$ for $x > b$. Then the solutions are plane waves in the intervals $(-\infty, a)$ and (b, ∞), i.e., for $p > 0$ [10],

$$\begin{aligned} \psi_p(x) &= e^{ip(x-a)/\hbar} + r(p)e^{-ip(x-a)/\hbar} \qquad (x < a), \\ \psi_p(x) &= t(p)e^{ip_+(p)(x-b)/\hbar} \qquad (x > b), \end{aligned} \tag{2.5}$$

and for $p < 0$,

$$\begin{aligned} \psi_p(x) &= e^{-ip(x-b)/\hbar} + r(p)e^{ip(x-b)/\hbar} \qquad (x > b), \\ \psi_p(x) &= t(p)e^{-ip_-(p)(x-a)/\hbar} \qquad (x < a). \end{aligned} \tag{2.6}$$

The incoming wave is thus assumed to have amplitude one. The reflection and transmission amplitudes $r(p)$ and $t(p)$ are uniquely determined from the solution (see [10, p. 226]). By elimination of the unknowns $r(p)$ and $t(p)$, the boundary conditions (2.2)–(2.3) are obtained.

From the amplitudes $r(p)$ and $t(p)$ the reflection and transmission coefficients can be computed:

$$R(p) = |r(p)|^2, \quad T(p) = \frac{\mathrm{Re}(p_\pm(p))}{|p|}|t(p)|^2 \quad \text{if } \pm p > 0,$$

where Re denotes the real part of a complex number. It holds $R(p) + T(p) = 1$ for all $p \in \mathbb{R}$ and $T(p) = T(-p_+(p))$ for all $p > 0$ (reciprocity property).

We must introduce some *macroscopic* quantities. The electron density $n_s(x)$ is defined by

$$n_s(x) = \int_{\mathbb{R}} g(p)|\psi_p|^2 dp, \tag{2.7}$$

where $g(p)$ is the statistics of the left reservoir if $p > 0$ and of the right reservoir if $p < 0$ (also called alimentation function), and the current density is given by

$$J_s(x) = \frac{e\hbar}{m}\int_{\mathbb{R}} g(p)\mathrm{Im}(\overline{\psi_p(x)}\psi_p'(x))dp, \tag{2.8}$$

where Im denotes the imaginary part of a complex number. In one space dimension, the expression for the current density can be reformulated. Indeed, using (2.2)–(2.3) and (2.5)–(2.6), we obtain

$$\begin{aligned} \hbar\,\mathrm{Im}(\overline{\psi_p(a)}\psi_p'(a)) &= \mathrm{Im}(ip(1 - |r(p)|)^2) = pT(p) \quad \text{for } p > 0, \\ \hbar\,\mathrm{Im}(\overline{\psi_p(a)}\psi_p'(a)) &= -\mathrm{Im}(ip_-(p)|t(p)|^2) = pT(p) \quad \text{for } p < -p_0, \end{aligned}$$

and $\hbar\,\mathrm{Im}(\overline{\psi_p(a)}\psi_p'(a)) = 0$ if $-p_0 < p \le 0$, where $p_0 = \mathrm{Re}(2em(\widetilde{V}(b) - \widetilde{V}(a)))^{1/2}$. Therefore, since $J_s(x)$ is constant,

$$\begin{aligned} J_s(x) = J_s(a) &= \frac{e}{m}\int_0^\infty g(p)T(p)pdp + \frac{e}{m}\int_{-\infty}^{-p_0} g(p)T(p)pdp \\ &= \frac{e}{m}\int_0^\infty g(p)T(p)pdp + \frac{e}{m}\int_\infty^0 g(-p_+)T(-p_+)pdp \\ &= \frac{e}{m}\int_0^\infty (g(p) - g(-p_+))T(p)pdp, \end{aligned} \tag{2.9}$$

where we have used the substitution $p \mapsto -p_+(p)$ and the reciprocity property of $T(p)$.

The choice of the alimentation function $g(p)$ depends on the choice of a and b. If $a = 0$ and $b = L$, it is taken to be the Fermi–Dirac distribution:

$$g(p) = \frac{mk_BT_0}{2\pi^2\hbar^3} \ln\left\{1 + \exp\left[\frac{1}{k_BT_0}\left(-\frac{p^2}{2m} + E_F\right)\right]\right\},$$

where k_B is the Boltzmann constant, T_0 the lattice temperature, and E_F the Fermi energy computed from the charge-neutrality condition at the left or right reservoir boundary. This formula holds if the system is macroscopically large in its transversal dimensions (see [17, Ch. 9] or [25, Ch. 1.5.2.1]). In the case $a = x_1$ and $b = x_2$ we choose $g(p)$ as an approximation of the so-called quantum Maxwellian (see (2.16) and (2.11)).

2.2. The quantum drift-diffusion model. In order to explain the coupling with the quantum drift-diffusion equations, we need to review its derivation from a Wigner–BGK model as performed by Degond, Méhats, Ringhofer [11, 12]. We start from the collisional Wigner equation in one space dimension,

$$w_t + \frac{p}{m}w_x + \frac{e}{m}\theta[\widetilde{V}] = Q(w), \quad x, p \in \mathbb{R},\ t > 0, \tag{2.10}$$

where w_t, w_x denote the partial derivatives of w with respect to t and x, respectively, and $\theta[\widetilde{V}]$ is a pseudodifferential operator given by

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

(We do not indicate here the time-dependency of $\widetilde{V}$.) The collision operator is assumed to be of BGK type, i.e.,

$$Q(w) = \frac{1}{\tau}(M[w] - w),$$

where τ is the relaxation time and $M[w]$ is the so-called quantum Maxwellian defined as the minimizer of the quantum entropy, subject to the constraint of given particle density [13]. To make this precise, we introduce first the so-called relative quantum entropy.

Let W^{-1} be the inverse Wigner transform (or Weyl quantization):

$$(W^{-1}[w])\phi(x) = \frac{1}{2\pi\hbar}\int_{\mathbb{R}^2} w\left(\frac{x+y}{2}, p, t\right)\phi(y)e^{ip(x-y)/\hbar}dpdy \quad \text{for suitable } \phi(x).$$

The relative quantum entropy for the density matrix $\rho = W^{-1}[w]$ is defined as follows:

$$S(\rho) = \frac{1}{2\pi\hbar}\int_{\mathbb{R}^2} w\left(\mathrm{Ln}(w) - 1 + \frac{H}{k_BT_0}\right)dxdp,$$

where $H = |p|^2/2m - e\widetilde{V}(x)$ is the classical Hamiltonian, $\mathrm{Ln}(w) := W[\ln(W^{-1}[w])]$ is called the quantum logarithm, and $\ln(f)$ is the usual operator logarithm. We wish to find, for given $n(x)$, the minimizer of

$$S(\rho^*) = \min\left\{S(\rho) : \frac{1}{2\pi\hbar}\int_{\mathbb{R}} W[\rho]dp = n(x) \quad \forall x\right\}.$$

The solution (if it exists) is $\rho_a = W^{-1}[w_a]$, where $w_a = \mathrm{Exp}(a(x) - H/k_BT_0)$ and $\mathrm{Exp}(f) := W[\exp(W^{-1}[f])]$ is called the quantum exponential. The function $a(x)$ is such that $\int w_a(x,p)dp/2\pi\hbar = n(x)$. We call w_a a quantum Maxwellian. In other words, for given $w(x,p)$, we define $M[w]$ as the *quantum Maxwellian*

$$M[w] = \mathrm{Exp}\left(b(x) - \frac{|p|^2}{2mk_BT_0}\right)$$

such that $\int (M[w] - w)dp = 0$ and $b(x) = a(x) - e\widetilde{V}(x)/k_BT_0$. We assume that the integral constraint fixes the function $b(x)$ in a unique way [11].

The quantum drift-diffusion model is derived from (2.10) in the diffusion limit. For this, we introduce the scaling $t \to t/\delta$ and $Q(w) \to Q(w)/\delta$, which yields

$$\delta^2 w_t^\delta + \delta\left(\frac{p}{m}w_x^\delta + \frac{e}{m}\theta[\widetilde{V}]\right) = Q(w^\delta).$$

As $\delta \to 0$, the formal limit $w_0 = \lim_{\delta\to 0} w^\delta$ satisfies $Q(w_0) = 0$, hence $w_0 = M[w_0] = \mathrm{Exp}(b_0(x) - |p|^2/2mk_BT_0)$ for some function $b_0(x)$, and $n(x) = \int w_0(x,p)dp/2\pi\hbar$. In [11] it was shown by a Chapman–Enskog expansion method that n satisfies the equation

$$n_t - \frac{1}{e}J_x = 0, \quad J = \frac{\tau e k_BT_0}{m}nb_{0,x} - \frac{\tau e^2}{m}n\widetilde{V}_x,$$

and n and b_0 are related by

$$n = \int_{\mathbb{R}} \mathrm{Exp}\left(b_0(x) - \frac{|p|^2}{2mk_BT_0}\right)\frac{dp}{2\pi\hbar}$$

(see [12, Lemma 6.5]). Furthermore, we can expand $w_0 = \mathrm{Exp}(b_0 - |p|^2/2mk_BT_0)$ and thus n and J in terms of $\hbar$.

LEMMA 2.1. *The following (formal) expansion holds for all $x, p \in \mathbb{R}$ up to order $O(\hbar^4)$:*

$$w_0(x,p) = A_0 e^{-p^2/2mk_BT_0} n\left[1 + \frac{\hbar^2}{12mk_BT_0}\left(1 - \frac{p^2}{mk_BT_0}\right)\left(\frac{(\sqrt{n})_{xx}}{\sqrt{n}} - \frac{(\sqrt{n})_x^2}{n}\right)\right], \tag{2.11}$$

where $A_0 = \sqrt{2\pi\hbar^2/mk_BT_0}$.

Notice that $(\sqrt{n})_{xx}/\sqrt{n} - (\sqrt{n})_x^2/n = (\log n)_{xx}$ but the formulation in (2.11) is more convenient later.

Proof. We use Lemma 5.6 of [11] to obtain

$$\begin{aligned} w_0(x,p) = {} & \exp\left(b_0 - \frac{p^2}{2mk_BT_0}\right) - \frac{\hbar^2}{8mk_BT_0}\exp\left(b_0 - \frac{p^2}{2mk_BT_0}\right) \\ & \times \left[\left(-1 + \frac{p^2}{3mk_BT_0}\right)b_{0,xx} - \frac{1}{3}b_{0,x}^2\right] + O(\hbar^4). \end{aligned} \tag{2.12}$$

This gives

$$n(x) = \int_{\mathbb{R}} w_0(x,p)\frac{dp}{2\pi\hbar} = n_0(x) + \frac{\hbar^2}{24mk_BT_0}n_0(x)(2b_{0,xx} + b_{0,x}^2) + O(\hbar^4),$$

where $n_0 := \exp(b_0)/A_0$. Consequently, $n = n_0 + O(\hbar^2)$ and we can solve the above equation for n_0:

$$n_0 = n - \frac{\hbar^2}{24mk_BT_0} n(2b_{0,xx} + b_{0,x}^2) + O(\hbar^4).$$

Insertion of the above formula into (2.12) yields, after some computations,

$$w_0 = A_0 e^{-p^2/2mk_BT_0} n \left[1 + \frac{\hbar^2}{24mk_BT_0}\left(1 - \frac{p^2}{mk_BT_0}\right) b_{0,xx}\right] + O(\hbar^4).$$

Since

$$b_{0,xx} = 2\frac{(\sqrt{n})_{xx}}{\sqrt{n}} - \frac{n_x^2}{2n^2} + O(\hbar^2) = 2\frac{(\sqrt{n})_{xx}}{\sqrt{n}} - 2\frac{(\sqrt{n})_x^2}{n} + O(\hbar^2)$$

(see [11, sec. 5.3]), we conclude the assertion. □

In [11] it has been shown that we can expand $n = n_q + O(\hbar^4)$, $J = J_q + O(\hbar^4)$, and n_q, J_q in such a way that n_q, J_q satisfy the so-called quantum drift-diffusion equations

$$n_{q,t} - \frac{1}{e} J_{q,x} = 0, \quad J_q = \frac{\tau e k_B T_0}{m} n_{q,x} - \frac{\tau e^2}{m} n_q (\widetilde{V} + Q[n_q])_x, \tag{2.13}$$

where

$$Q[n_q] = \frac{\hbar^2}{6em} \frac{(\sqrt{n_q})_{xx}}{\sqrt{n_q}}$$

is the so-called Bohm potential.

We can rewrite (2.13) by introducing the function $\sigma = \sqrt{n_q}$ and the quantum quasi-Fermi potential

$$F = U_{\rm th} \ln n_q - \widetilde{V} - Q[n_q]$$

(with the thermal voltage $U_{\rm th} = k_BT_0/e$). Then the stationary version of (2.13) becomes

$$(\sigma^2 F_x)_x = 0, \quad F = U_{\rm th} \ln \sigma^2 - \widetilde{V} - \frac{\hbar^2}{6em} \frac{\sigma_{xx}}{\sigma}. \tag{2.14}$$

The current density equals $J_q = (\tau e^2/m)\sigma^2 F_x = e\mu_n \sigma^2 F_x$.

In order to specify boundary conditions for (2.14), we need to distinguish the two cases for the choice of a and b. Let first $a = 0$ and $b = L$. Then, following [22], we assume that the total space charge vanishes and that no quantum effects occur at the boundary (in the sense $(\sqrt{n_q})_{xx} = 0$). Thus

$$n_q(0) = n_D(0), \quad n_q(L) = n_D(L), \quad F(0) = 0, \quad F(L) = -V_a, \tag{2.15}$$

where V_a denotes the applied voltage. The case $a = x_1$ and $b = x_2$ is studied in the next subsection.

2.3. The coupled model. Let $a = x_1$ and $b = x_2$. We solve the Schrödinger equation (2.1) in (a, b) with boundary conditions (2.2)–(2.3) and the quantum drift-diffusion model (2.14) in the intervals $(0, x_1)$ and (x_2, L).

To compute the electron and current densities n_s and J_s, respectively (see (2.7) and (2.8)), we need to specify the alimentation function $g(p)$. We choose $g(p)$ as the $O(\hbar^4)$ approximation (2.11) of the quantum Maxwellian (see Lemma 2.1):

$$g(p) = w_0(a,p) \quad \text{if } p > 0, \quad g(p) = w_0(b,p) \quad \text{if } p < 0. \tag{2.16}$$

Although in general w_0 does not need to be a nonnegative function, we observed in the numerical simulations that $w_0(a,p)$ and $w_0(b,p)$ are always positive. Another idea could be to choose the classical Maxwellian instead of the (approximation of the) quantum Maxwellian w_0 in (2.16). However, this choice did not lead to a converging algorithm. (See section 3 for details on the discretization and the iterative procedure.) A possible explanation could be that the use of the classical Maxwellian is not consistent with the use of the quantum drift-diffusion model.

The coupling of both models is realized through connection conditions relating the macroscopic unknowns (the electron density and the current density) at the two interface points x_1 and x_2. We assume that at the interface, the particle density and the current density are continuous, i.e.,

$$n_q(x_1) = n_s(x_1), \quad n_q(x_2) = n_s(x_2), \quad J_q(x_1) = J_s(x_1), \quad J_q(x_2) = J_s(x_2).$$

Thus, the quantum drift-diffusion model is solved in $(0, x_1)$ with the four boundary conditions

$$n_q(0) = n_D(0), \quad F(0) = 0, \quad n_q(x_1) = n_s(x_1), \quad J_q(x_1) = J_s(x_1). \tag{2.17}$$

On the boundary of the interval (x_2, L) we impose the conditions

$$n_q(x_2) = n_s(x_2), \quad J_q(x_2) = J_s(x_2), \quad n_q(L) = n_D(L), \quad F(L) = -V_a. \tag{2.18}$$

The interface conditions for the current densities can be written in a form which is more convenient for the numerical computations. For this, we remark that we have from (2.16) and (2.9)

$$J_s(x_1) = \int_0^\infty w_0(a,p)T(p)pdp - \int_0^\infty w_0(b,-p_+(p))T(p)pdp$$

and J_s is constant in $[x_1, x_2]$. Insertion of the above formula into $J_q(x_j) = J_s(x_j)$ $(j = 1, 2)$ gives, after some elementary computations, up to order $O(\hbar^4)$,

$$\begin{aligned} J_q(x_2) = J_q(x_1) &= \sqrt{2\pi\theta}A_0\left(n_s(x_1) - e^{-\delta V/2\theta}n_s(x_2)\right)I_1 \\ &+ \sqrt{2\pi\theta}\frac{\hbar^2 A_0}{12}n_s(x_1)\left(\frac{\sigma_{xx}}{\sigma} - \frac{\sigma_x^2}{\sigma^2}\right)_{x=x_1}(\theta I_1 - I_2) \\ &+ \sqrt{2\pi\theta}\frac{\hbar^2 A_0}{12}n_s(x_2)\left(\frac{\sigma_{xx}}{\sigma} - \frac{\sigma_x^2}{\sigma^2}\right)_{x=x_2}e^{-\delta V/2\theta}((\theta - \delta V)I_1 - I_2), \end{aligned}$$

where $\sigma = \sqrt{n_q}$, $\theta = mk_BT_0$, $\delta V = 2em(\widetilde{V}(x_2)\widetilde{-}V(x_1))$, and

$$I_1 = \frac{1}{\sqrt{2\pi\theta}}\int_0^\infty pT(p)e^{-p^2/2\theta}dp, \quad I_2 = \frac{1}{\sqrt{2\pi\theta}}\int_0^\infty p^3T(p)e^{-p^2/2\theta}dp.$$

For the numerical computations we replace the Bohm potential term σ_{xx}/σ by the expression (2.14) in order to avoid the computation of the second derivatives of σ. This gives two nonlinear boundary conditions for the quantum quasi-Fermi potential F:

$$n_s(x_j)F_x(x_j) = f(F(x_1), F(x_2), \sigma_x(x_1), \sigma_x(x_2)), \quad j = 1, 2,$$

where

$$\begin{aligned} f(F(x_1), F(x_2), \sigma_x(x_1), \sigma_x(x_2)) &= \sqrt{2\pi\theta}A_0 I_1 \left(n_s(x_1) - e^{-\delta V/2\theta} n_s(x_2)\right) \\ &+ \frac{em}{2}\sqrt{2\pi\theta}A_0(\theta I_1 - I_2)n_s(x_1)\left(F(x_1) - U_{\rm th}\ln n_s(x_1) + \widetilde{V}(x_1) + \frac{\hbar^2}{6em}\frac{\sigma_x^2(x_1)}{n_s(x_1)}\right) \\ &+ \frac{em}{2}\sqrt{2\pi\theta}A_0 e^{-\delta V/2\theta}((\theta - \delta V)I_1 - I_2)n_s(x_2) \\ &\times \left(F(x_2) - U_{\rm th}\ln n_s(x_2) + \widetilde{V}(x_2) + \frac{\hbar^2}{6em}\frac{\sigma_x^2(x_2)}{n_s(x_2)}\right). \end{aligned}$$

Finally, the electrostatic potential is self-consistently coupled through the Poisson equation

$$\frac{d}{dx}\left(\varepsilon_s\frac{dV}{dx}\right) = e(n - n_D(x)), \quad x \in (0, L),$$

where ε_s is the semiconductor permittivity and the particle density $n(x)$ is given by

$$n(x) = \begin{cases} n_q(x) & : x \in (0, x_1) \cup (x_2, L), \\ n_s(x) & : x \in (x_1, x_2). \end{cases} \tag{2.19}$$

3. Numerical discretization. We discretize the equations by introducing a uniform mesh $\xi_k = k\triangle x$, $\triangle x > 0$, $k = 0, \ldots, K$, and $L = K\triangle x$.

The Schrödinger equation is solved by central finite differences as in [10]. For the convenience of the reader we recall the discretization scheme. We assume that the effective mass is constant in $[x_1, x_2]$ since we wish to compare our results with those from the literature, e.g., [7, 26]. Moreover, a space-dependent effective electron mass leads to quite complicated quantum drift-diffusion models whose numerical solution is delicate [28]. The Schrödinger equation (2.1) with boundary conditions (2.2) for $p > 0$ (the case $p < 0$ can be treated analogously) can be equivalently rewritten as

$$y'' = -\frac{2m}{\hbar^2}(E_p + e\widetilde{V})y \quad \text{in } (x_1, x_2), \quad y(x_2) = 1, \quad y'(x_2) = \frac{i}{\hbar}p_+(p), \tag{3.1}$$

where $p_+(p)$ is defined in (2.4) and y and ψ_p are related by

$$\psi_p(x) = \frac{2ipy(x)}{\hbar y'(x_1) + ipy(x_1)}.$$

With the approximations $y_k \approx y(\xi_k)$ and $\widetilde{V}_k \approx \widetilde{V}(\xi_k) = V_{\rm ext}(\xi_k) + V(\xi_k)$ the discrete problem is

$$\frac{1}{(\triangle x)^2}(y_{k+1} - 2y_k + y_{k-1}) = -\frac{2m}{\hbar^2}(E_p + e\widetilde{V}_k)y_k.$$

This problem is solved as in [10] by Stoermer's method, i.e., writing

$$z_k = \frac{y_{k+1} - y_k}{\triangle x} \quad (k = 0, \dots, K-1), \quad z_K = y'(x_2) - \frac{m}{\hbar^2}(E_p + e\widetilde{V}(x_2))y(x_2)$$

and noticing that z_K is known in view of (3.1), the iteration reads

$$\begin{aligned} z_k &= z_{k+1} - \frac{m}{\hbar^2}\triangle x(E_p + e\widetilde{V}_{k-1})y_{k-1}, \\ y_k &= y_{k+1} - \triangle x z_k, \\ y_0' &= y'(x_1) = z_1 - \frac{m}{\hbar^2}\triangle x(E_p + e\widetilde{V}_0)y_0, \end{aligned}$$

which allows us to calculate z_k and y_k recursively. The algorithm is vectorized and implemented in MATLAB.

The quantum drift-diffusion model (2.14) is approximated by central finite differences as in [22]. The proposed scheme has been proved to be positivity preserving, i.e., the discrete electron density is positive (see [21, 22] for details). Let σ_k and F_k be approximations of $\sigma(\xi_k)$ and $F(\xi_k)$, respectively. Then the discrete problem corresponding to (2.14) is

$$\frac{1}{(\triangle x)^2}\left(\sigma_{k+1/2}^2 F_{k+1} - (\sigma_{k+1/2} + \sigma_{k-1/2})F_k + \sigma_{k-1/2}^2 F_{k-1}\right) = 0, \tag{3.2}$$

$$F_k = U_{\rm th} \ln \sigma_k^2 - \widetilde{V}_k - \frac{\hbar^2}{6em(\triangle x)^2}\frac{\sigma_{k+1} - 2\sigma_k + \sigma_{k-1}}{\sigma_k}, \tag{3.3}$$

where $\sigma_{k\pm1/2} = (\sigma_k + \sigma_{k\pm1})/2$.

Finally, the discrete Poisson equation for $V_k \approx V(\xi_k)$ reads as follows:

$$\frac{\varepsilon_s}{(\triangle x)^2}(V_{k+1} - 2V_k + V_{k-1}) = e(n_k - n_D(\xi_k)) \quad (k = 1, \dots, K-1), \quad V_0 = 0, \quad V_K = U,$$

where we have assumed a constant semiconductor permittivity ε_s, and the electron density is either given by (2.7) (discretized by a standard quadrature formulae) in the interval $[x_1, x_2]$ or by σ_k^2 otherwise.

We describe now the iterative procedures for the various models. We use a fixed-point strategy to solve the SP system in $(0, L)$ (i.e., $a = 0$ and $b = L$). More precisely, we choose the electrostatic potential $V^{(0)}$ of the thermal equilibrium state as an initial guess. This potential is the (discrete) solution of the Poisson equation in which the electron density is replaced by the Thomas–Fermi approximation:

$$\varepsilon_s \frac{d^2V}{dx^2} = e\left(N_c F_{1/2}\left(\frac{\mu - V}{k_B T_0}\right) - n_D(x)\right) \quad \text{in } (0, L), \quad V(0) = V(L) = 0.$$

Here, $N_c = 2(mk_BT_0/2\pi\hbar^2)^{3/2}$ is the effective density of states and $F_{1/2}$ is the Fermi integral of order 1/2 [17, Ch. 9]. The chemical potential μ is a constant in thermal equilibrium and computed from the nonlinear equation $n_D(0) = n(0) = F_{1/2}(\mu/k_BT_0)$ (where we assumed charge neutrality). With this initial potential we solve the discrete Schrödinger problem to obtain the (discrete) scattering states which allow us to compute the discrete electron density from (2.7) and the discrete current density from (2.8). Finally, the update of the electrostatic potential can be computed from the Poisson equation written in the Gummel formulation (see (3.8) in [10]),

$$\varepsilon_s \frac{d^2V^{(j+1)}}{dx^2} + e(V^{(j+1)} - V^{(j)})n^{(j)} = e(n^{(j)} - n_D(x)).$$

The solution of the Schrödinger eigenvalue problem is the most costly part of the iteration. Indeed, we use a uniform grid of 10,000 values for p with grid size $\triangle p = 0.0005\sqrt{mk_BT_0}$. An adaptive mesh size strategy has been proposed for the SP system in the whole domain in [7], but we observed that the adaptive algorithm did not converge for the coupled model.

Another idea to reduce the computing time is to choose different mesh sizes $\triangle x_q$ in the collisional zone and $\triangle x_s$ in the ballistic zone. However, it turned out in our numerical experiments that the computing time is minimized when using the same mesh size in both zones, i.e., $\triangle x = \triangle x_s = \triangle x_q$. We have used $\triangle x = 0.25\,\mathrm{nm}$ (540 grid points).

The quantum drift-diffusion model in the whole interval $(0, L)$ is solved by Newton's method. The initial guess is chosen to be the potential in thermal equilibrium with $V_a = 0$ (see (2.15)). In thermal equilibrium, the quantum quasi-Fermi potential F is constant and, in view of the boundary conditions (2.15), the constant is zero. Then the thermal equilibrium potential is computed by a fixed-point scheme, i.e., we solve first the discrete equation

$$\frac{\hbar^2}{6em(\triangle x)^2}(\sigma_{k+1} - 2\sigma_k + \sigma_{k-1}) = \sigma_k(U_{\mathrm{th}} \ln \sigma_k^2 - \widetilde{V}_k), \quad \sigma_0 = n_D(0), \quad \sigma_K = n_D(L),$$

employing Newton's method and then the linear Poisson equation with homogeneous boundary conditions.

In the case $a = x_1$ and $b = x_2$ we use again a fixed-point-type iteration. More precisely, let an initial guess for the potential be given (namely, the thermal equilibrium potential of the quantum drift-diffusion model). Then compute the scattering states from the discrete Schrödinger equation. The electron and current densities are calculated according to (2.7) and (2.8), where the approximation (2.16) is employed. The quantum drift-diffusion system is solved by Newton's method according to (3.2)–(3.3) using the boundary conditions (2.17) and (2.18), respectively. Finally, an update for the electrostatic potential is obtained through the solution of the Poisson equation using the definition (2.19).

For all models we use the continuation method in the applied voltage, i.e., with the solution for the applied voltage U as an initial guess, we solve the problem applying the potential $V_a = U + \triangle U$ and use this solution again as initial guess for the next applied voltage. For the computations of the next section, we have choosen $\triangle U = 0.005\,\mathrm{V}$.

4. Numerical results. In this section we simulate a simple one-dimensional resonant tunneling diode. We choose the same geometry and data as in [26] (essentially taken from [23]). The tunneling diode consists of highly doped GaAs regions near the contacts and a lightly doped middle region of 35-nm length (see Figure 4.1). The middle region contains a quantum well of 5-nm length sandwiched between two 5-nm AlGaAs barriers. The double barrier heterostructure is placed between two 10-nm GaAs spacer layer with a doping of $5 \cdot 10^{15}\ \mathrm{cm}^{-3}$. These spacers are enclosed by two layers of 50-nm length and with doping $10^{18}\,\mathrm{cm}^{-3}$. The total length is thus 135 nm. The double barrier height is 0.3 eV. The physical effect of the barriers is a shift in the quasi-Fermi potential level, which we model by an additional step function V_{ext} added to the electrostatic potential. The physical constants are chosen as in [26] and are summarized in Table 4.1.

First we present the current-voltage characteristics of the above tunneling diode for four different model equations: the SP model in the whole interval, the coupled drift-diffusion SP (DD-SP) model of [10], the coupled quantum DD-SP (QDD-SP)

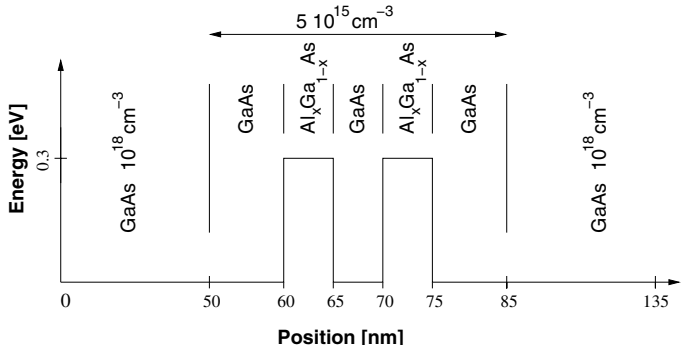


FIG. 4.1. *Geometry of the resonant tunneling diode and external potential V_{ext} modeling the double barriers.*

TABLE 4.1
Physical parameters and their numerical values.

Parameter	Physical meaning	Numerical value
m	electron mass	$0.067 \cdot 9.11 \cdot 10^{-31}$ kg
T	lattice temperature	300 K
ε_s	semiconductor permittivity	$11.44 \cdot 8.85 \cdot 10^{-12}$ As/Vm
τ	relaxation time	10^{-12} s
k_B	Boltzmann constant	$1.36 \cdot 10^{-23}$ J/K
$\hbar$	reduced Boltzmann constant	$1.055 \cdot 10^{-34}$ Js

model presented in this paper, and the quantum drift-diffusion (QDD) model in the whole interval. See [2, 10] for a description of the DD-SP model and its numerical discretization. Figure 4.2 displays the current-voltage curves for the first three models. In all these models, a region of negative differential resistance (NDR), in which the current is decreasing, can be observed. The valley current appears at approximately the same voltage, but the voltage at which the peak current is observed is slightly different in the models. Moreover, the peak-to-valley ratio in the QDD-SP model is smaller than in the SP model. This comes probably from the fact that there are no collisions modeled in the SP system. In [2], a decrease of the peak-to-valley ratio was also observed in simulations from the collisional DD-SP model (compared to the ballistic DD-SP model). The electron density and the electrostatic potential at applied voltage $V_a = 0.25$ V are presented in Figures 4.3 and 4.4, respectively. The results obtained here compare well with those of [2] and [26], where the coupled DD-SP model and the SP model, respectively, were solved.

The current-voltage curve computed from the QDD model does not show any NDR region (Figure 4.5). In fact, the QDD model is too diffusive, thus destroying the quantum resonance behavior (at least at large temperatures). It is known that a nonmonotone behavior of the current-voltage characteristic from the QDD model can be obtained by fitting the effective electron mass. Notice that the values for the current density are overestimated compared to the other models. Since we are using a constant relaxation time model (see section 2.2), the scattering effects are comparable for both temperatures and the current density for $T = 300$ K is larger than that for $T = 77$ K in view of the larger thermal energy.

For comparison, the current-voltage characteristics for the SP, QDD-SP, and DD-SP models at $T = 77$ K are displayed in Figure 4.6. The current-voltage curves of the two coupled models differ significantly from the curve computed from the SP

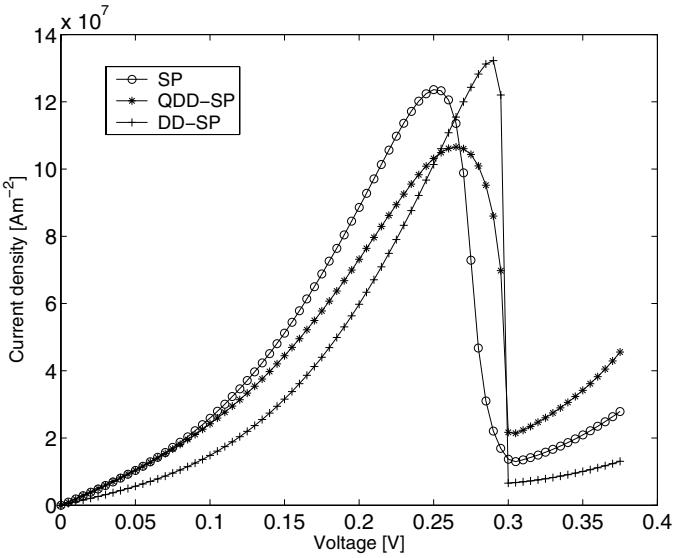


FIG. 4.2. *Current-voltage characteristics for a resonant tunneling diode using the SP, QDD-SP, and DD-SP models.*

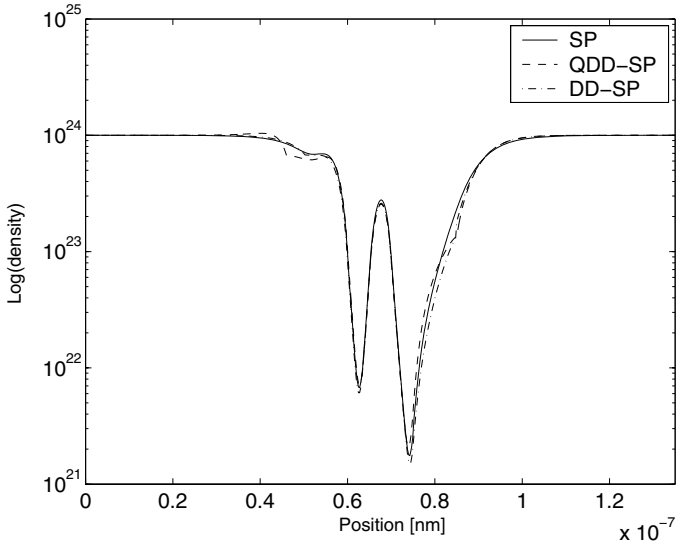


FIG. 4.3. *Electron densities at applied voltage $V_a = 0.25$ V from the SP, QDD-SP, and DD-SP models.*

model. This can be understood by the fact that at this low temperature, collisional effects are expected to be less important such that the use of diffusion models is questionable. Mathematically, the difference of the curves rather comes from the Bohm potential term than from the interface conditions, since a similar difference can be observed comparing the QDD and DD models without coupling to the SP system. The peak current from the QDD-SP model coincides with the peak current from the SP model, whereas the DD-SP model overestimates the peak current. Therefore, for

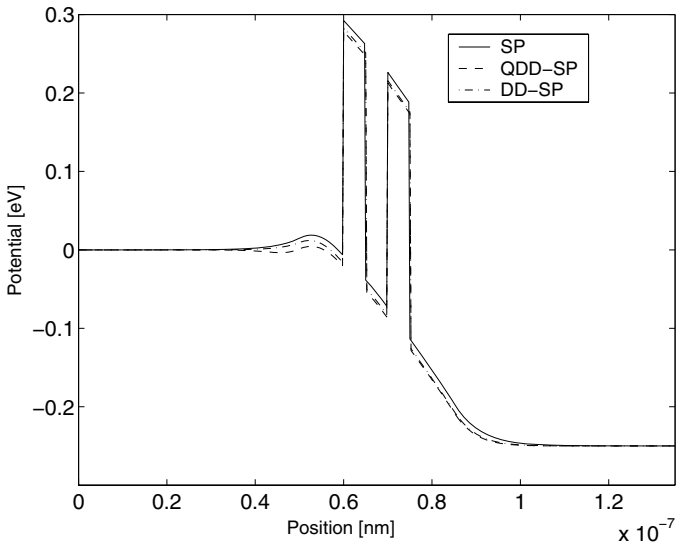


FIG. 4.4. *Potential profiles at applied voltage $V_a = 0.25$ V from the SP, QDD-SP, and DD-SP models.*

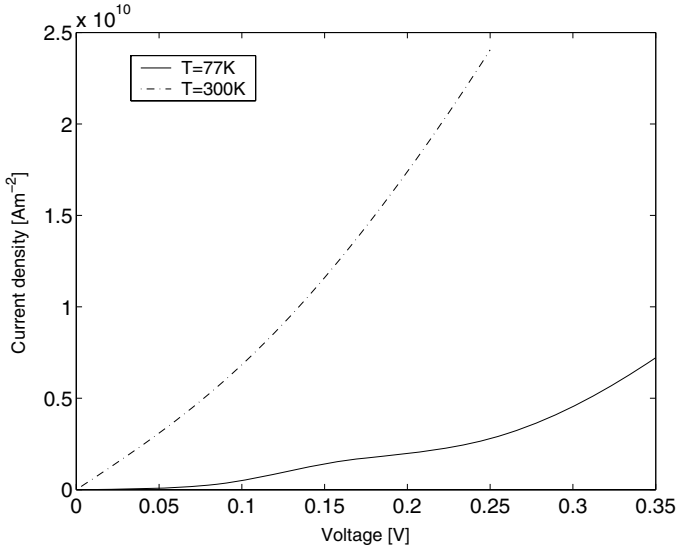


FIG. 4.5. *Current-voltage characteristics for a resonant tunneling diode using the QDD model for two different lattice temperatures.*

low temperature, the QDD-SP model seems to provide more accurate results than the DD-SP model.

In [10] it was observed that the current-voltage values depend quite sensitively on the position of the left interface point $a = x_1$, but the influence of the position of the right interface $b = x_2$ is very small. This observation holds true also for the QDD-SP model (Figures 4.7 and 4.8). When the left interface is too close to the double barrier, the potential in the quantum region cannot reproduce the correct quantum

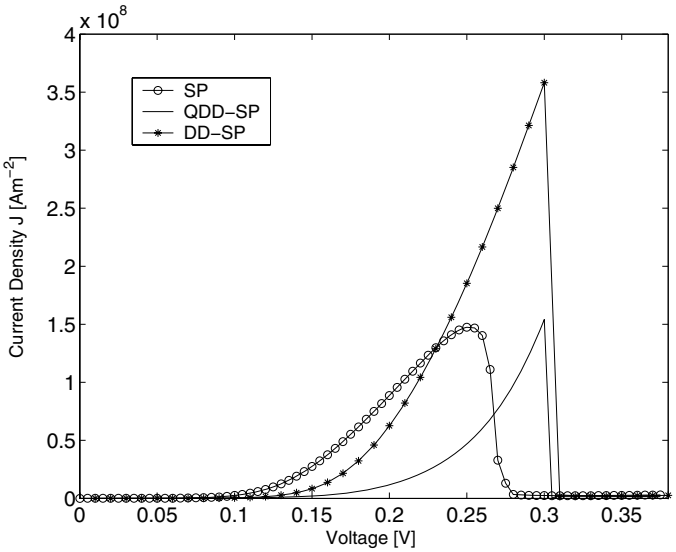


FIG. 4.6. *Current-voltage characteristics for a resonant tunneling diode using the SP, QDD-SP, and DD-SP models at temperature $T = 77\,K$.*

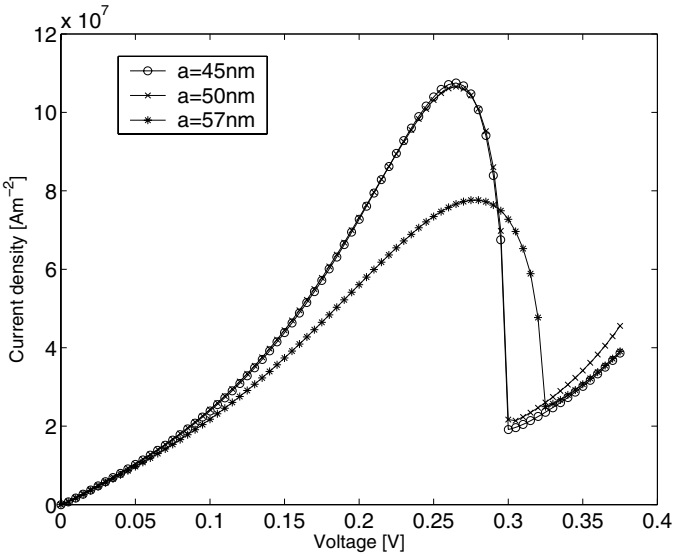


FIG. 4.7. *Influence of changes of the interface $a = x_1$ on the current-voltage characteristics using the QDD-SP model.*

resonances. It is argued in [10] that the insensitivity of the choice of the right interface position comes from the fact that the electrons crossing the double barriers have high energy and can be described equally well by a classical or a quantum model.

We have also investigated the effect of the relaxation time τ on the current-voltage curve. Figure 4.9 shows that the results are insensitive of the choice of τ. This holds true also in the ballistic DD-SP model of [10] (see Figure 4.10). In the collisional DD-SP model of [2], however, the characteristic is very sensitive with respect to τ.

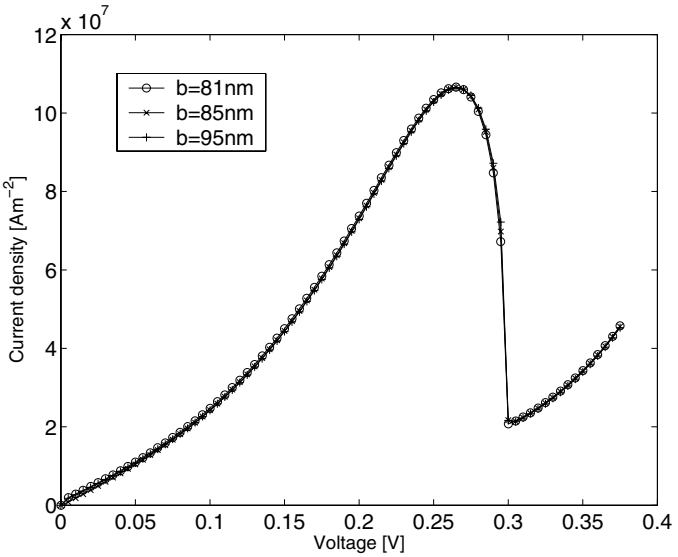


FIG. 4.8. *Influence of changes of the interface $b = x_2$ on the current-voltage characteristics using the QDD-SP model.*

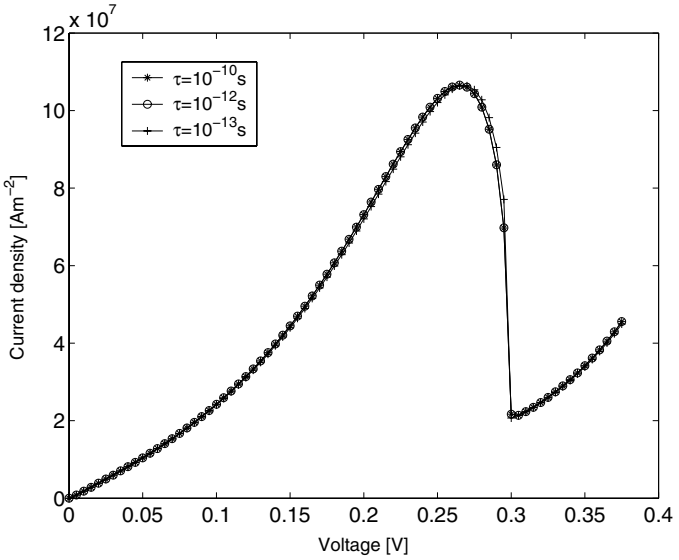


FIG. 4.9. *Influence of the relaxation time on the current-voltage characteristic using the QDD-SP model.*

This seems to come from the collision events in the quantum region modeled by the Pauli master equation. Notice that in the coupled DD-SP and QDD-SP models, no collisions are taken into account in the microscopic quantum region.

It is well known that the current-voltage curve of a tunneling diode exhibits hysteresis, probably resulting from storage effects of the charges in the quantum well [23]. Hysteresis can be found in simulations from the Wigner–Poisson model [23] or from the quantum hydrodynamic equations [9]. It cannot be observed in simulations

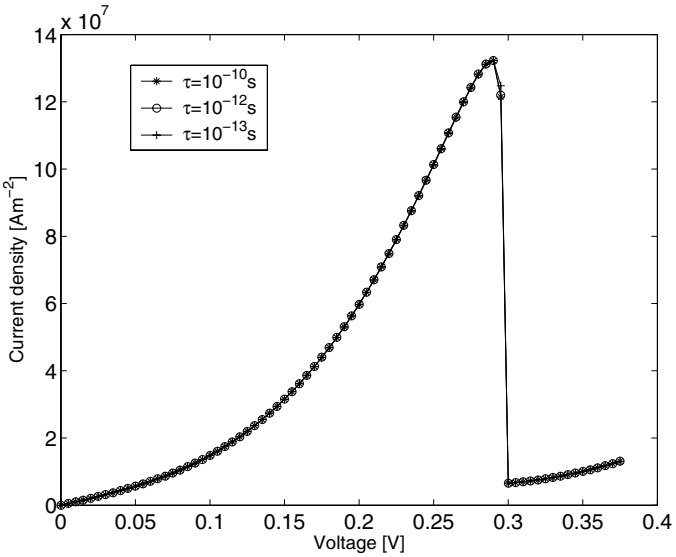


FIG. 4.10. *Influence of the relaxation time on the current-voltage characteristic using the DD-SP model.*

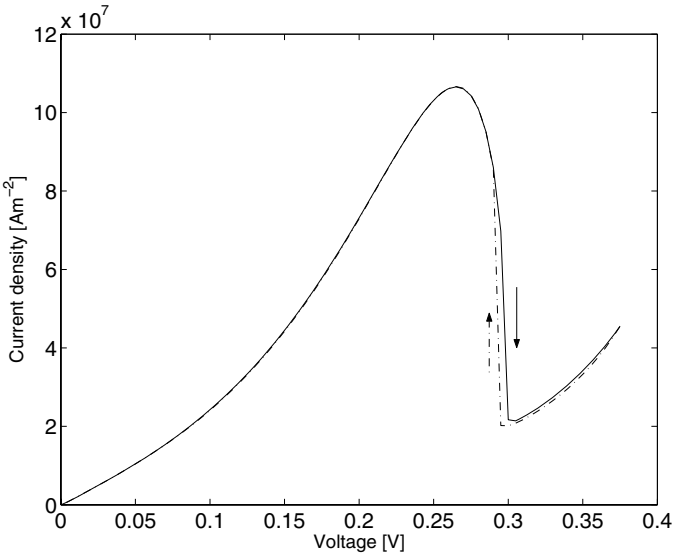


FIG. 4.11. *Hysteresis in the current-voltage characteristic using the QDD-SP model.*

from the QDD model. However, employing the coupled QDD-SP model, the current-voltage characteristic shows hysteresis (Figure 4.11). We notice that also with the DD-SP model, hysteresis effects can be found (Figure 4.12).

Finally, Table 4.2 displays the CPU times (for a 2.4-GHz Pentium 4 processor) needed to compute the current-voltage characteristics in various voltage ranges for the SP, QDD-SP, and DD-SP models. All algorithms are vectorized in the same way such that the CPU times are comparable. The CPU time needed to calculate the current-voltage curve with the QDD model is of the order of a few seconds only;

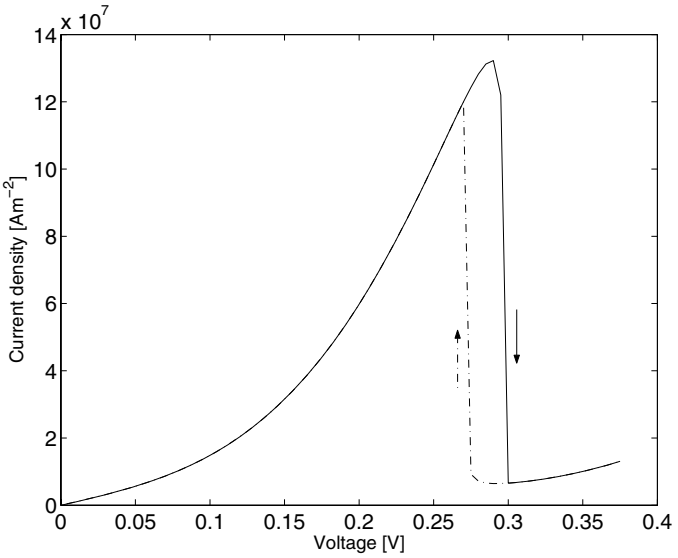


FIG. 4.12. *Hysteresis in the current-voltage characteristic using the DD-SP model.*

TABLE 4.2

CPU times needed to compute the current-voltage characteristic in the indicated voltage ranges, using different models.

model	[0 0.25eV]	[0.25eV 0.3eV]	[0.3eV 0.375eV]	[0 0.375eV]
SP	5776 s	1255 s	1371 s	8402 s
QDD-SP	2305 s	570 s	629 s	3504 s
DD-SP	695 s	187 s	168 s	1050 s

however, the numerical results are not satisfactory. The QDD-SP model needs only about half the CPU time compared to the SP model. This shows that the coupled model allows to reduce significantly the computing time compared to the full SP model. The DD-SP model is even faster; the reduction factor is about 8 compared to the SP model and about 3 compared to the QDD-SP model. The latter model is faster since the current density in the drift-diffusion region can be computed by an analytic expression [2, formula (16)], whereas the current density of the QDD-SP model is a result of the solution of the QDD model. Although the CPU time for the DD-SP model is smaller than for the QDD-SP model, the latter model has the advantage that there is a quantum description in the whole semiconductor device, avoiding any artificial separation of classical and quantum zones. From a more practical point of view, the DD-SP model may be prefered due to the smaller CPU time.

Acknowledgment. The authors thank Dr. Méhats (Toulouse) for very fruitful discussions.

REFERENCES

[1] M. ANCONA AND G. IAFRATE, *Quantum correction to the equation of state of an electron gas in a semiconductor*, Phys. Rev. B, 39 (1989), pp. 9536–9540.

[2] M. BARO, N. BEN ABDALLAH, P. DEGOND, AND A. EL AYYADI, *A 1D coupled Schrödinger drift-diffusion model including collisions*, J. Comput. Phys., 203 (2005), pp. 129–153.

[3] M. Baro, H.-C. Kaiser, H. Neidhardt, and J. Rehberg, *A quantum transmitting Schrödinger-Poisson system*. Rev. Math. Phys., 16 (2004), pp. 281–330.
[4] N. Ben Abdallah, *A hybrid kinetic-quantum model for stationary electron transport in a resonant tunneling diode*, J. Statist. Phys., 90 (1998), pp. 627–662.
[5] N. Ben Abdallah, P. Degond, and I. Gamba, *Coupling one-dimensional time-dependent classical and quantum transport models*, J. Math. Phys., 43 (2002), pp. 1–24.
[6] N. Ben Abdallah, F. Méhats, and N. Vauchelet, *Analysis of a drift-diffusion-Schrödinger-Poisson model*. C. R. Acad. Sci. Paris Ser. I, 335 (2002), pp. 1007–1012.
[7] N. Ben Abdallah, O. Pinaud, C. Gardner, and C. Ringhofer, *A comparison of resonant tunneling based on Schrödinger's equation and quantum hydrodynamics*, VLSI Design, 15 (2002), pp. 695–700.
[8] B. Biegel, *Simulation of Ultra-Small Electronic Devices: The Classical-Quantum Transition Regime*, Technical Report 97-028, NASA, 1997.
[9] Z. Chen, B. Cockburn, C. Gardner, and J. Jerome, *Quantum hydrodynamic simulation of hysteresis in the resonant tunneling diode*, J. Comput. Phys., 117 (1995), pp. 274–280.
[10] P. Degond and A. El Ayyadi, *A coupled drift-diffusion model for quantum semiconductor device simulations*, J. Comput. Phys., 181 (2002), pp. 222–259.
[11] P. Degond, F. Méhats, and C. Ringhofer, *Quantum energy-transport and drift-diffusion models*, J. Statist. Phys., 118 (2005), pp. 625–665.
[12] P. Degond, F. Méhats, and C. Ringhofer, *Quantum hydrodynamic models derived from the entropy principle*, Contemp. Math., 371 (2005), pp. 107–131.
[13] P. Degond and C. Ringhofer, *Quantum moment hydrodynamics and the entropy principle*, J. Statist. Phys., 112 (2003), pp. 587–628.
[14] C. de Falco, E. Gatti, A. Lacaita, and R. Sacco, *Quantum-corrected drift-diffusion model for transport in semiconductor devices*, J. Comput. Phys., 204 (2005), pp. 533–561.
[15] D. Ferry and J.-R. Zhou, *Form of the quantum potential for use in hydrodynamic equations for semiconductor device modeling*, Phys. Rev. B, 48 (1993) pp. 7944–7950.
[16] W. Frensley, *Boundary conditions for open quantum systems driven far from equilibrium*, Rev. Modern Phys., 62 (1990), pp. 745–791.
[17] W. Frensley and N. Einspruch, (eds.), *Heterostructures and quantum devices*, in VLSI Electronics: Microstructure Science, Academic Press, San Diego, 1994.
[18] C. Gardner, *The quantum hydrodynamic model for semiconductor devices*, SIAM J. Appl. Math., 54 (1994), pp. 409–427.
[19] K. Gullapalli, D. Miller, and D. Neilirk, *Simulation of quantum transport in memory-switching double-barrier quantum-well diodes*, Phys. Rev. B, 49 (1994), pp. 2622–2628.
[20] J. Höntschel, W. Klix, and R. Stenzel, *Investigation of Quantum Transport Phenomena in Resonant Tunneling Structures by Simulations with a Novel Quantum Hydrodynamic Transport Model*, IoP Conf. Ser. 174 (2002), pp. 255–258.
[21] A. Jüngel, *Quasi-Hydrodynamic Semiconductor Equations*, Birkhäuser, Basel, 2001.
[22] A. Jüngel and R. Pinnau, *A positivity-preserving numerical scheme for a nonlinear fourth-order parabolic equation*, SIAM J. Numer. Anal., 39 (2001), pp. 385–406.
[23] N. Kluksdahl, A. Kriman, D. Ferry, and C. Ringhofer, *Self-consistent study of the resonant-tunneling diode*, Phys. Rev. B, 39 (1989), pp. 7720–7735.
[24] C. Lent and D. Kirkner, *The quantum transmitting boundary method*, J. Appl. Phys., 67 (1990), pp. 6353–6359.
[25] M. Lundstrom, *Fundamentals of Carrier Transport*, 2nd ed., Cambridge University Press, Cambridge, UK, 2000.
[26] O. Pinaud, *Transient simulations of a resonant tuneling diode*, J. Appl. Phys., 92 (2002), pp,. 1987–1994.
[27] T. Waho, *Resonant tunneling diode and its application to multi-GHz analog-to-digital converters*, Proceedings of Quantum Nanoelectronics for Meme-Media-Based Information Technologies, Sapporo, Japan, 2003, pp. 53–58.
[28] A. Wettstein, *Quantum Effects in MOS Devices*, Ph.D. thesis, ETH, Zürich, 2000.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 573–590

CRITERIA FOR THE CONVERGENCE, OSCILLATION, AND BISTABILITY OF PULSE CIRCULATION IN A RING OF EXCITABLE MEDIA*

H. SEDAGHAT†, C. M. KENT‡, AND M. A. WOOD§

Abstract. A discrete model based on a nonlinear difference equation (equivalent to a coupled map lattice of high dimension) is used to study the dynamics of a circulating pulse in a ring of excitable media, such as cardiac cells. Based on the global and local properties of monotonic restitution and dispersion curves, criteria are obtained for the asymptotic stability of the unique steady state (pulse circulating at constant frequency) as well as for nonconvergent oscillatory behavior of all nonequilibrium trajectories (pulse circulating at variable frequency). We also demonstrate that in certain cases the system is bistable, where an asymptotically stable equilibrium coexists with stable oscillatory solutions.

Key words. difference equation, nonlinear, reentry, restitution, asymptotic stability, persistent oscillations, bistability

AMS subject classifications. 39A11, 37N25, 92C99

DOI. 10.1137/040617078

1. Introduction.

The periodic contractions of muscles that result in the beating of our hearts are caused by electrochemical signals or excitations called *action potentials* that propagate through chains of cardiac cells. Normally, cardiac cells generate and conduct action potentials in response to excitation by the self-oscillatory pacemaker cells in the heart's sinoatrial and atrioventricular nodes. However, in certain circumstances a closed loop or ring of tissue is formed within the heart that unidirectionally recycles a previously generated action potential. Such a *reentrant circuit* is capable of blocking the much slower pacemaker signals by transmitting its rapid pulses outward through adjacent cell layers, thus taking over the beating of the heart and leading to potentially life-threatening arrhythmias.

Reentry of an action potential pulse in a ring of cardiac cells or other excitable media and the resulting self-sustained propagation is relatively easy to model mathematically because of the simple one-dimensional geometry. The study of such models contributes to our understanding of cardiac arrhythmias, and the results of the study find concrete applications to experimental models of reentrant electrical activity in cardiac muscle. Nevertheless, the mathematical expressions of the manner in which a reentrant pulse propagates in a loop involve complex nonlinear equations whose study requires the application of a variety of different methods from the dynamical systems theory.

In [14] a discrete model of a reentrant circuit is developed based on the *restitution* and *dispersion* properties of cardiac cells. Mathematically, the centerpiece of this

*Received by the editors October 15, 2004; accepted for publication (in revised form) August 17, 2005; published electronically December 30, 2005.
http://www.siam.org/journals/siap/66-2/61707.html
†Corresponding author. Department of Mathematics, Virginia Commonwealth University, Richmond, VA 23284-2014 (hsedagha@vcu.edu).
‡Department of Mathematics, Virginia Commonwealth University, Richmond, VA 23284-2014 (cmkent@vcu.edu).
§Department of Internal Medicine, Virginia Commonwealth University Health Systems, Richmond, VA 23298-0053 (mawood@vcu.edu).

model is a coupled-map lattice whose dimension is equal to the number of cells (or excitable units) in the loop. Close agreement was shown between certain important experimental facts and the model's simulations (using typical exponential-type maps to fit the restitution curve data). The authors discuss a number of issues, including the formation of a unidirectional block. For sustained propagation, they also study local stability and through numerical simulations establish the occurrence of discrete Hopf (or Neimark–Sacker) bifurcations with the variation of a parameter in their dispersion curve. In this way they exhibit the occurrence of almost periodic solutions with multiple incomensurate frequencies.

Two subsequent papers [5], [6] presented equivalent continuous space versions of the above model in terms of a delay differential equation and an integral delay equation, respectively. These papers present a mix of analytical and numerical results. The analytical results establish the occurrence of solutions with multiple frequencies via Hopf bifurcations in the continuous case. The length of the ring is used as the bifurcation parameter in these papers, where the nonconvergent solutions arise when this length is sufficiently reduced. These issues are discussed in greater detail in [6] than in [5]. Notably, the numerical simulations of the delay-differential and the integral delay equation used different versions of the mapping in [14] for their discretizations. These versions are nearly equivalent and one of them is discussed in this paper.

The concepts of restitution and dispersion are well known and have been widely studied in various contexts, both theoretical and experimental; see, e.g., [2], [3], [5], [6], [7], [10], [13], [14], [15], [16], [24], [25], [26]. In this paper we reexamine the dynamical system in [14], limiting our focus to dynamics of sustained reentry. Our purpose is twofold: First, for sustained propagation our results extend those obtained in [14] to include global behavior in a bounded (but not infinitesimal) invariant region of the phase space. In particular, we obtain conditions for (a) the convergence of all orbits within the invariant region to a unique stable equilibrium so that the reentrant pulse circulates at constant frequency; (b) the persistent oscillation (bounded but nonconverging) of all nonequilibrium orbits, resulting in pulse circulation at variable frequency; and (c) the occurrence of bistable behavior with coexisting stable oscillatory and convergent steady states so that it becomes possible to shift from one of these states to another through premature stimulations. Standard methods from the mathematical literature on discrete dynamics (e.g., [1], [8], [12], [17], [18], [19], [20], [23]) are used to advantage in the proofs of Theorems 1 and 2.

Our second goal in this paper is to further generalize various results in previous studies by using *generic* restitution and dispersion curves that satisfy a few minimal conditions. This shows indirectly that the aforementioned qualitatively different types of behavior are general manifestations of the dynamical system under consideration and are not peculiar to a specific class of elementary mathematical functions. This substantial extension of the previously published material is made possible by reliance on rigorous analytical methods that lead to a deeper understanding of the mathematical nature of the model. We limit use of numerical simulations largely to examples to illustrate the main results.

2. The model. We consider a loop of cardiac tissue (or, more generally, of excitable cells) of fixed length L that consists of a fixed number m of cells or, more generally, aggregate units or nodes in the sense of [14]. If the real number $\Delta L_i > 0$ denotes the ith cardiac unit spacing or internodal separation, then $L = \sum_{i=1}^{m} \Delta L_i$. If we denote the average of ΔL_i, $i = 1, 2, \ldots, m$, by ΔL and define

$$\delta_i = \frac{\Delta L_i}{\Delta L}$$

for each i, then

$$\sum_{i=1}^{m} \delta_i = \sum_{i=1}^{m} \frac{\Delta L_i}{\Delta L} = m \quad \text{and} \quad L = m\Delta L. \tag{1}$$

In most published works in the literature the spacings ΔL_i are assumed to deviate negligibly from the average ΔL. In such cases one assumes that $\delta_i = 1$ for all i and considers a *homogeneous loop* in which all nodes are uniformly spaced. This is an important special case which captures all the significant features with a minimum of technical details. Finally, for numerical simulations we arbitrarily take $m = 500$, although much smaller values of m give qualitatively similar outcomes [14].

2.1. Restitution curves.

The action potential duration restitution. The *action potential duration* (APD) is the length of time (usually measured in milliseconds (ms)) that a node or cell is active after excitation by a pulse. After passage of the APD, if no new excitation takes place, the cell enters a recovery period called the *diastolic interval* (DI), also measured in ms. The DI ends only by the arrival of a new excitation and it is only during the DI that a cell can fire or become active again if excited.

Let $\mathrm{APD}_{i,n}$ and $\mathrm{DI}_{i,n}$ denote the APD and DI, respectively, for the cell i in beat n. The most basic temporal relationship that is possible between the APD and the DI may be stated as

$$\mathrm{APD}_{i,n} = A(DI_{i,n-1}), \qquad i = 1, 2, \ldots, m, \tag{2}$$

where A is an increasing single variable function whose graph is called the *APD-restitution curve*.

We use the basic form (2) in this paper but point out that in various papers (e.g., [4], [9], [11], [13], [15], [26]) it has been suggested that in general the function A may contain additional delays or past APD dependence to account for memory effects.

Dispersion and the conduction time (CT) restitution. Each cardiac cell is capable of conducting the action potential pulse through it in a finite amount of time. The speed with which the pulse propagates through a cell is the *conduction velocity* or CV, often measured in cm/sec. If $\mathrm{CV}_{i,n}$ is the conduction velocity through cell i in beat n, then we may express our second restitution hypothesis as follows:

$$CV_{i,n} = V(DI_{i,n-1}), \qquad i = 1, 2, \ldots, m, \tag{3}$$

where V is a nondecreasing single variable function. Its graph is called the *dispersion curve*. With ΔL sufficiently small, we may assume that V does not change from one end of a unit or cell to its other end, so (3) defines a unitwise or cellular *conduction time restitution* as follows:

$$CT_{i,n} = T(i, DI_{i,n-1}) = \frac{\Delta L_i}{V(DI_{i,n-1})} = \frac{\delta_i \Delta L}{V(DI_{i,n-1})}.$$

If we define the function of one variable,

$$C(t) = \frac{\Delta L}{V(t)},$$

then $\mathrm{CT}_{i,n} = \delta_i C(DI_{i,n-1})$ for $i = 1, 2, \ldots, m$. Note that the function C is nonincreasing; we refer to its graph as the *CT-restitution curve*. Except for the factor ΔL the function C is the same as the recovery curve in [14].

2.2. The propagation equation. During sustained reentry the pulse or the excitation front moves from node to node around the loop. The length of each cycle (or beat) n is divided into two distinct periods, the APD followed by DI. The length of each cycle is also the sum of CT for all nodes or cells in the loop. Therefore, we have

$$APD_{i,n} + DI_{i,n} = \sum_{j=1}^{i-1} CT_{j,n+1} + \sum_{j=i}^{m} CT_{j,n}, \qquad i = 1, 2, \dots, m. \tag{4}$$

Since in this setting $i = 1$ is the reference point on the loop where a cycle begins and ends, the split in the sums reflects the fact that for $i > 1$ conduction in cells 1 through $i - 1$ takes place during beat $n + 1$. Using the restitution relations, (4) may be written as the following system:

$$DI_{i,n} = \sum_{j=1}^{i-1} T(j, DI_{j,n}) + \sum_{j=i}^{m} T(j, DI_{j,n-1}) - A(DI_{i,n-1}), \qquad i = 1, \dots, m. \tag{5}$$

Equation (5) is the coupled-map lattice discussed in [14]. It is also a partial difference equation in the spatial variable i and the temporal variable n. See [24] for an adaptation of this argument to a nonloop structure.

Rather than working directly with (5), we first transform it to an ordinary mth-order difference equation by taking advantage of the periodic nature of the loop. Define the combined space-time variable

$$x_{mn+i} = DI_{i,n}$$

and note that

$$A(DI_{i,n-1}) = A(x_{m(n-1)+i})$$

and

$$\sum_{j=1}^{i-1} T(j, DI_{j,n}) + \sum_{j=i}^{m} T(j, DI_{j,n-1}) = \sum_{j=1}^{i-1} T(j, x_{mn+j}) + \sum_{j=i}^{m} T(j, x_{m(n-1)+j}) = \sum_{j=mn-m+i}^{mn+i-1} T(j, x_j).$$

Next, we substitute $k = mn + i$ to get the following:

$$x_k = \sum_{j=k-m}^{k-1} \delta_j C(x_j) - A(x_{k-m}). \tag{6}$$

This equation is the main object of interest in this paper. It is equivalent to (5) provided that we extend the coefficient δ periodically, i.e., $\delta_{i+mn} = \delta_i$ for all positive integers n and each $i = 1, 2, \dots, m$ (a valid assumption since the loop consists of a finite number m of units).

In the homogeneous case where $\delta_i = 1$ for all i, (6) becomes an autonomous difference equation for which a greater number of results exist in the literature. This autonomous version of (6) is the same as that obtained in [6] on a discretization of

their independently developed integral equation (also see [7]). However, no analysis of the discrete case is given in these references.

Each solution of (6) is an infinite sequence $\{x_k\}_{k=1}^{\infty}$ of DI quantities that is generated by a given set of initial values which may be thought of as an initial state vector,

$$\mathbf{DI}_0 = [DI_{1,0}, DI_{2,0}, \dots, DI_{m,0}] = [x_{-m+1}, x_{-m+2}, \dots, x_0].$$

Also, *each subsequent state vector*

$$\mathbf{DI}_n = [x_{m(n-1)+1}, x_{m(n-1)+2}, \dots, x_{mn}], \quad n \geq 0,$$

constitutes the dynamic state of the loop in one cycle. Each such vector determines the DI values for all the cells in the loop within a given cycle, and from the DI values and the restitution and dispersion functions, other quantities such as APD and CV can be computed. Further, plotting the components of $\mathbf{DI}_n$ as functions of the spatial coordinate i, one obtains the spatial profile of the DI values in each cycle (see Figure 2).

Exponential restitution and dispersion curves. Experimental data from pacing experiments as well as data generated by the numerical simulations of the ionic PDE models are typically fitted with exponential maps for numerical studies of pulse propagation. Monotonic maps that are variations of the following two functions appear frequently in the literature on cardiac electrophysiology:

$$A(t) = a - be^{-\sigma t} + pe^{-\gamma(t-\tau)^2}, \quad C(t) = \frac{\Delta L}{c}[1 + de^{-\omega t}], \tag{7}$$

where the parameters a, b, c, d, τ, σ, γ, $\omega > 0$ and p is a real number. We use these representations in our examples and figures below. The parameter c is in fact the limiting (or maximum) value of conduction velocity; i.e., $c = \lim_{t\to\infty} V(t)$. The second exponential term has been added here to the definition of A as a simple device for modifying A locally near the value τ. For small values of $|p|$ the mapping A is strictly increasing and in this paper we shall be concerned with this range of p values only. In particular, if $p = 0$, then A takes the form used in [14]; where the above definition of A is used in the sequel (except for the one on bistability) we assume that $p = 0$.

Using definitions (7) in (6) and rearranging a few terms gives the following:

$$x_k = \frac{d\Delta L}{c} \sum_{j=k-m}^{k-1} \delta_j e^{-\omega x_j} + be^{-\sigma x_{k-m}} - pe^{-\gamma(x_{k-m}-\tau)^2} + \frac{L}{c} - a. \tag{8}$$

Equation (8) defines a complex dynamical system that displays a wide range of behaviors. Nevertheless, it is mathematically a rather special case of (6) in which certain situations do not occur that are possible for (6); e.g., nonconcavity or nondifferentiability. In this paper we generally use the following parameter values in (8), unless otherwise stated:

$$a = 24,\ b = 12,\ \sigma = 0.5,\ \ p = 0,\ c = 6,\ d = 1,\ \omega = 1,\ \Delta L = 0.3. \tag{9}$$

These values are largely arbitrary but not far-fetched; they are within scientifically acceptable ranges and here they are used mainly for numerical verifications of our results in illustrative examples. For the sake of interpretation, $L = m\Delta L = 150$ may be considered to be in millimeters and time units for DI, APD, etc. in various diagrams will be 10 milliseconds each, unless otherwise specified.

3. Dynamics of sustained reentry. A positive solution $\{x_k\}_{k=1}^{\infty}$ of (6), being an infinite sequence, represents *sustained reentry*. As noted above, a partitioning of this sequence into m-dimensional vectors gives the history of the loop's dynamic states in the phase space. In this section we study the qualitative properties of the solutions of (6).

3.1. The existence of a unique equilibrium. We begin with a set of basic assumptions concerning the restitution functions. The functions A, C in (7) in particular satisfy all these hypotheses.

(A1) There is $r_A \geq 0$ such that the APD restitution function A is continuous and increasing on the interval $[r_A, \infty)$ with $A(r_A) \geq 0$.
(A2) There is $r_C \geq 0$ such that the CT restitution function C is continuous and nonincreasing on the interval $[r_C, \infty)$ with $\inf_{x \geq r_C} C(x) \geq 0$.
(A3) There is $r \geq \max\{r_A, r_C\}$ such that

$$mC(r) > A(r) + r.$$

The first two assumptions express basic physiological facts that are commonly attributed to the functions A and C. The third assumption guarantees (via Lemma 1) the existence of a structurally stable equilibrium or steady state. Physiologically, this means that conduction around the ring must be sufficiently slow to achieve the desired effect. Note also that (A1)–(A3) allow for the possibility that the restitution functions may not be monotonic or otherwise well-behaved near the origin.

Next, we define the *auxiliary function* $F = mC - A$ and note that by (A1)–(A3), F is continuous and decreasing on the interval $[r, \infty)$ and satisfies

$$F(r) > r. \tag{10}$$

Let x^* be a steady state solution or *equilibrium* of (6), i.e., a solution of the equation

$$x = \sum_{j=k-m}^{k-1} \delta_j C(x) - A(x) = mC(x) - A(x) = F(x). \tag{11}$$

In particular, x^* is also a fixed point of the auxiliary map F so it is the same in the autonomous case $\delta_i = 1$.

LEMMA 1. *Assume that* (A1)–(A3) *hold.*
(a) *Equation* (6) *has a unique positive equilibrium* $x^* \in (r, F(r))$.
(b) $(x^*, F(r)] = F([r, x^*))$.
(c) $[r, F(r)] = [r, x^*] \cup F([r, x^*))$ *disjointly.*

Proof. (a) Let $f(x) = F(x) - x$ so that by (11) x^* is a zero of f. Then by (10) $f(r) > 0$. Further, since F is decreasing for $x \geq r$, applying F to (10) gives

$$f(F(r)) = F(F(r)) - F(r) < 0.$$

The existence of $x^* \in (r, F(r))$ is now established by applying the intermediate value theorem to the continuous f. The uniqueness of x^* is a clear consequence of the strictly decreasing nature of f.

(b) Let $y \in (x^*, F(r)]$. Then $x = F^{-1}(y) > F^{-1}(F(r)) = r$ and $x < F^{-1}(x^*) = x^*$. Thus $(x^*, F(r)] \subset F([r, x^*))$. The converse is clear.

(c) The disjoint union follows immediately from part (b). ∎

The interval $[r, F(r)]$ represents a *relevant interval* for this model since what happens outside this interval is not relevant to our discussion of sustained reentry. The relevant interval is not generally invariant, although it may contain an invariant interval.

Remark. We are interested only in nonnegative DI values; therefore, if a zero of F occurs in the relevant interval, i.e., if there is $p \in [r, F(r)]$ such that $F(p) = 0$, then $F(x) < 0$ for $x > p$ so the relevant interval reduces to $[r, p]$. We note that since $F(x^*) = x^* > 0$, it must be that $x^* \in (r, p)$.

3.2. The invariant interval. The existence of a nontrivial invariant interval in particular guarantees an open set of bounded solutions for (6) and thus assures the robust occurrence of sustained reentry. An additional hypothesis is required.

(A4) There is $s \in [r, x^*)$ such that $s \le F(F(s)) = F^2(s)$.

LEMMA 2. *Assuming* (A1)–(A4)*, the interval* $I = [s, F(s)]$ *is nontrivial, contains* x^**, and is invariant for* (6)*; i.e., if the initial values* $x_0, \dots, x_{-m+1}$ *are in* I*, then* $x_k \in I$ *for all* $k \ge 1$*. Also* $I \subset [r, F(r)]$*.*

Proof. First, observe that since $s < x^*$ and F is decreasing, then $F(s) > x^* > s$. Hence, the interval I contains x^* and is nontrivial, i.e., it has a nonempty interior. Next, assume that $x_0, \dots, x_{-m+1} \in I$. Then for $j = -m+1, \dots, 0$, $x_j \ge s$, so $A(x_{-m+1}) \ge A(s)$ and $C(x_j) \le C(s)$. Therefore,

$$x_1 = \sum_{j=-m+1}^{0} \delta_j C(x_j) - A(x_{-m+1}) \le mC(s) - A(s) = F(s).$$

Similarly, for $j = -m+1, \dots, 0$, $x_j \le F(s)$ so $A(x_{-m+1}) \le A(F(s))$ and $C(x_j) \ge C(F(s))$. Therefore,

$$\begin{aligned} x_1 &= \sum_{j=-m+1}^{0} \delta_j C(x_j) - A(x_{-m+1}) \\ &\ge mC(F(s)) - A(F(s)) \\ &= F(F(s)) \\ &\ge s. \end{aligned}$$

It follows that $x_1 \in I$. Now, assume inductively that for $k \ge 1$, we have established that $x_{k-1}, \dots, x_{k-m} \in I$. Then repeating the above argument gives $x_k \in I$ and shows I to be invariant. ∎

3.3. Convergence to a stable steady state. In this section we look at a special case of (A4) that implies the asymptotic stability of the equilibrium (convergence of all trajectories in I to a stable steady state). This special case of condition (A4) is stated as follows:

(A4S) There is $s \in [r, x^*)$ such that $F^2(x) > x$ for all $x \in (s, x^*)$.

Figure 1 depicts a case where (A4S) holds with $r = s = 0$. Note in particular that if $A'(x^*) < 1$, then it is easy to see that (A4S) holds, at least when $C'(x^*) = 0$ (although C need not be constant). However, (A4S) is a weaker condition in that the differentiability of A is not required and that if differentiable, then the derivative A' need not be uniformly bounded by 1 in a left-neighborhood of x^* (i.e., small irregularities in the APD curve do not affect the qualitative behavior of the circulating pulse).

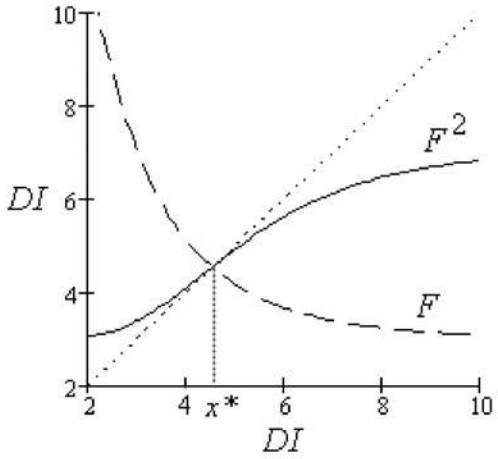


FIG. 1. *Conditions for the stability of the fixed point x^* of the mapping F, which is required in Theorem* 1. *When the graph of $F^2 = F \circ F$ crosses the diagonal going from above it to below it as shown, x^* is asymptotically stable.*

Before Theorem 1 is presented, it is necessary to consider the dynamics of the auxiliary map F under (A4S). It is known [23, sec. 2.1], that condition (A4S) is in fact necessary *and* sufficient for x^* to attract all orbits of F that start in $(s, F(s))$. But since F is a relatively simple mapping, we can be more specific here.

LEMMA 3. *Assume that* (A1)–(A3) *plus* (A4S) *hold. Then for every* $x_0 \in (s, x^*)$

$$s < x_0 < F^2(x_0) < \cdots < x^* < \cdots < F^3(x_0) < F(x_0) < F(s) \tag{12}$$

and

$$\lim_{n\to\infty} F^{2k}(x_0) = \lim_{n\to\infty} F^{2k+1}(x_0) = x^*. \tag{13}$$

Proof. Since F is decreasing, if $x_0 \in (s, x^*)$, then $F(x_0) > F(x^*) = x^*$ and $F(x_0) < F(s)$. Thus

$$x^* < F(x_0) < F(s). \tag{14}$$

Applying F to (14) in the above fashion gives

$$F^2(s) < F^2(x_0) < x^*.$$

Now (12) follows by simple induction. Statements (13) follow from (12) because F has no fixed points in $(s, F(s))$ other than x^* to which the odd and even iterates of F can converge. □

THEOREM 1. *Let* (A1)–(A3) *and* (A4S) *hold. Then the equilibrium x^* is stable and every solution of* (6) *with initial values in $(s, F(s))$ is attracted to x^*.*

Proof. First we establish the attracting nature of x^*. Let $x_0, \ldots, x_{-m+1}$ be in the interval $(s, F(s))$, and define

$$\mu_1 = \min\{x^*, x_0, \ldots, x_{-m+1}\}, \quad \mu_2 = \max\{x^*, x_0, \ldots, x_{-m+1}\}.$$

Since F is continuous, we have $F(x) \to F(s)$ as $x \to s$. Thus we can find $q \in (s, \mu_1)$ sufficiently close to s that $F(q) \in (\mu_2, F(s))$. Next, observe that since $x_0, \dots, x_{-m+1} > q$,

$$x_1 = \sum_{j=-m+1}^{0} \delta_j C(x_j) - A(x_{-m+1}) < mC(q) - A(q) = F(q).$$

Similarly, $x_0, \dots, x_{-m+1} < F(q)$ implies

$$x_1 = \sum_{j=-m+1}^{0} \delta_j C(x_j) - A(x_{-m+1}) > mC(F(q)) - A(F(q)) = F^2(q).$$

If (A4S) holds, then $F^2(q) > q$ so that $x_1 \in (F^2(q), F(q)) \subset (q, F(q))$. Repeating a similar calculation for $x_2, \dots, x_m$ we conclude that

$$x_k \in (F^2(q), F(q)) \subset (q, F(q)), \quad k = 1, \dots, m. \tag{15}$$

Next, we move on to the next cycle and look at x_{m+1}. Since by (15) $x_1, \dots, x_m > F^2(q)$,

$$x_{m+1} = \sum_{j=1}^{m} \delta_j C(x_j) - A(x_1) < mC(F^2(q)) - A(F^2(q)) = F^3(q);$$

further, $x_1, \dots, x_m < F(q)$ gives

$$x_{m+1} = \sum_{j=1}^{m} \delta_j C(x_j) - A(x_1) > mC(F(q)) - A(F(q)) = F^2(q).$$

Since $F^3(q) < F(q)$, this argument can be repeated for $x_{m+2}, \dots, x_{2m}$ to yield

$$x_k \in (F^2(q), F^3(q)) \subset (F^2(q), F(q)), \quad k = m+1, \dots, 2m.$$

Continuing this argument inductively leads to the conclusion that

$$\begin{aligned} x_k &\in (F^{2n}(q), F^{2n-1}(q)), \quad k = m(2n-2)+1, \dots, m(2n-1), \\ x_k &\in (F^{2n}(q), F^{2n+1}(q)), \quad k = m(2n-1)+1, \dots, 2mn. \end{aligned} \tag{16}$$

From these relations and Lemma 3 it is clear that x_k converges to x^* as $k \to \infty$. It remains to show that x^* is stable (dynamically in the sense of Liapunov). Let $\varepsilon > 0$ and use the continuity of F to pick $\delta \in (0, \varepsilon)$ small enough that $F(x^* - \delta) < x^* + \varepsilon$. If $x_0, \dots, x_{-m+1} \in (x^* - \delta, x^* + \delta)$, then it follows from Lemma 3 and (16) that

$$x_k \in (x^* - \delta, F(x^* - \delta)) \subset (x^* - \varepsilon, x^* + \varepsilon), \quad k \geq 1.$$

Hence x^* is stable. □

Remark 1. If x^* is attracting (e.g., conditions of Theorem 1 hold), then the cycle length $mC(x^*)$ may be easily computed as the fixed period T^* (analogous to the *basic cycle length* (BCL)) for the oscillation of the reentrant pulse. Note that

$$T^* = mC(x^*) = \frac{m\Delta L}{V(x^*)} = \frac{L}{V^*},$$

where V^* is steady state conduction velocity in the loop. Similarly, the APD is calculated from the restitution relation as $A^* = A(x^*)$. It may be emphasized that since $F(x^*) = x^*$ we have the cycle period $T^* = A^* + x^* > A^*$ as required.

The frequency $1/T^*$ can be quite high in this sort of reentrant regime. We calculate this frequency in a hypothetical case using (8) subject to the parameter values (9) but with L increased to 162 (from 150 by increasing m to 540). In this case, we obtain the situation depicted in Figure 1 so by Theorem 1 the fixed point $x^* \approx 4.55$ (estimated numerically and interpreted as a steady state DI of 45.5 ms) is globally asymptotically stable (with respect to I).

The fixed cycle length or period of the reentrant pulse is approximately

$$mC(4.55) = \frac{L}{c}(1 + de^{-4.55\omega}) = \frac{162}{6}(1 + e^{-4.55}) = 27.3,$$

which we interpret as 273 milliseconds, corresponding to a frequency of 60000/273 or a rather fast 220 cycles (or beats) per minute.

Remark 2. (A1)–(A3) plus (A4S) are sufficient for the asymptotic stability of the equilibrium but in general they are not necessary. A special case where these hypotheses are both necessary and sufficient for asymptotic stability is when C is constant (see Theorem 2.1.2 in [23]).

3.4. Persistent oscillations. In a series of experiments on animal cardiac tissue by Frame and Simson [10] it was found that the reentrant pulse does not always cycle around a loop with a fixed period. In some cases, the cycle length tended to oscillate without approaching a specific value. In [14] these oscillations were attributed to the appearance of quasiperiodic solutions for (5) due to local bifurcations (Neimark–Sacker or discrete Hopf). In this section we discuss sufficient conditions for oscillations to occur in all nontrivial solutions of (6) within the invariant I^m. Throughout this section we assume that the restitution functions A and C are continuously differentiable and that $\delta_i = 1$ for all i, i.e., the loop is homogeneous.

For the autonomous version of (6), the characteristic polynomial of the linearization at x^* is given as

$$P(\lambda) = \lambda^m + \sum_{i=1}^{m-1} \beta\lambda^i + \beta + \alpha, \quad \alpha = A'(x^*) > 0,\ \beta = -C'(x^*) \geq 0.$$

See, e.g., [18] or [23]. Note that the roots of P are the eigenvalues of the linearization of (6) at x^*. We now list some special properties of P.

LEMMA 4.

(a) *P has no nonnegative (real) roots.*

(b) *If some root of P lies on the unit circle in the complex plane, then $\alpha + \beta = 1$.*

(c) *If $\alpha + \beta = 1$, then for each root λ of P, $1/\lambda$ is also a root.*

(d) *If $\beta = 0$ (even if C is not constant), then either all roots of P are inside the unit disk in the complex plane if $\alpha < 1$ or they are all outside if $\alpha > 1$.*

Proof. (a) This is clear from the facts that $\beta \geq 0$ and $\alpha + \beta > 0$. (b) First we show that (i) implies (iii). Let $\alpha + \beta = 1$. By part (b), roots $\lambda_j = \rho_j \exp(i\theta_j)$ of P must have modulus $\rho_j = 1$ for all $j = 1, \ldots, m$ and thus they are on the unit circle. In particular, the only possible real root of P is -1 which occurs when m is odd.

Next, (iii) trivially implies (ii), so it remains to show that (ii) implies (i). Let λ_1 be a root that is on the unit circle. Then $\lambda_1 = \exp(i\theta_1)$ so the conjugate $\exp(-i\theta_1) =$

$1/\lambda_1$ is also a root. Since $P(\lambda_1) = 0$ it follows that

$$\beta \sum_{i=1}^{m-1} \lambda_1^i = -\lambda_1^m - \alpha - \beta. \tag{17}$$

Also, $P(1/\lambda_1) = 0$ so

$$\begin{aligned} 0 &= \frac{1}{\lambda_1^m} + \beta \sum_{i=1}^{m-1} \frac{1}{\lambda_1^i} + \beta + \alpha \\ &= \frac{1}{\lambda_1^m} \left[1 + \beta \sum_{i=1}^{m-1} \lambda_1^i \right] + \alpha + \beta \\ &= \frac{1 - \lambda_1^m - \alpha - \beta}{\lambda_1^m} + \alpha + \beta \\ &= \left(\frac{1}{\lambda_1^m} - 1 \right) (1 - \alpha - \beta). \end{aligned} \tag{18}$$

Note that $\lambda_1^m \neq 1$, since otherwise the sum of the mth roots of unity in (17) would add up to -1, leaving $-\beta$ on the left but giving $-1 - \alpha - \beta$ on the right. Therefore, by (18) it is the case that $1 - \alpha - \beta = 0$ as required.

(c) If $\alpha + \beta = 1$, then

$$P\left(\frac{1}{\lambda}\right) = \frac{1}{\lambda^m} + \beta \sum_{i=1}^{m-1} \frac{1}{\lambda^i} + 1 = \frac{1}{\lambda^m} \left[\lambda^m + \beta \sum_{i=1}^{m-1} \lambda^i + 1 \right] = \frac{P(\lambda)}{\lambda^m}.$$

(d) This is straightforward since P reduces to the simple equation $\lambda^m + \alpha$ if $\beta = 0$. □

Next, a global oscillation result [23, p. 166] is needed which we quote here as a lemma. We say that a sequence $\{x_n\}_{n=1}^{\infty}$ *oscillates persistently* if it is bounded and has at least two distinct limit points. In particular, persistently oscillating solutions cannot converge to a point.

LEMMA 5. *Consider the general difference equation*

$$x_n = f(x_{n-1}, \ldots, x_{n-m}), \tag{19}$$

where $f : D \to \mathbb{R}$ for a set $D \subset \mathbb{R}^m$. Assume that (19) *has a unique fixed point x^* and that all the eigenvalues of the linearization of* (19) *at x^* (i.e., the roots of its characteristic polynomial) lie outside the unit disk in the complex plane. Further, assume that*

$$f(x^*, \ldots, x^*, x) \neq x^* \quad \text{if} \quad x \neq x^*. \tag{20}$$

Then all nontrivial solutions of (19) *that are bounded oscillate persistently.*

We are now ready for the main result of this subsection.

THEOREM 2. *Assume that* (A1)–(A4) *and one of the following inequalities hold:*

$$A'(x^*) + (m-2)C'(x^*) > 1 \tag{21}$$

or

$$A'(x^*) > 0, \quad C'(x^*) \leq -1. \tag{22}$$

Then x^* is unstable and all nontrivial solutions of (6) *oscillate persistently in the invariant interval I.*

Proof. In the notation of Lemma 4, inequality (21) may be written as

$$\alpha - (m-2)\beta > 1. \tag{23}$$

We first show that if (23) holds, then all roots of the characteristic polynomial P lie outside the unit disk in the complex plane. To this end, define the polynomial

$$Q(\lambda) = \frac{\lambda^m}{\alpha+\beta} P\left(\frac{1}{\lambda}\right) = \lambda^m + \frac{1}{\alpha+\beta}\sum_{j=1}^{m-1} \beta\lambda^j + \frac{1}{\alpha+\beta}$$

and observe that λ_0 is a root of Q if and only if $1/\lambda_0$ is a root of P. Thus if every root of Q is inside the unit disk, then every root of P will be outside as desired. A well-known sufficient condition (e.g., [23, pp. 209, 210]) for all the roots of Q to be inside the unit disk is that the sum of all the coefficients (except for that of the highest power λ^m) be less than unity, i.e.,

$$1 > \sum_{j=1}^{m-1} \frac{\beta}{\alpha+\beta} + \frac{1}{\alpha+\beta} = \frac{(m-1)\beta+1}{\alpha+\beta}. \tag{24}$$

It is easy to see that (24) is equivalent to (23).

Next, assume that (22) holds. Then in the notation of Lemma 4, $\beta \geq 1$ with $\alpha > 0$; in particular, $\alpha + \beta > 1$. If all roots of P are *not* outside the unit disk, then some root

$$\lambda_0 = r_0 e^{i\theta_0} = r_0(\cos\theta_0 + i\sin\theta_0)$$

of P has modulus $r_0 \leq 1$. By Lemma 4(c) we may assume that $r_0 < 1$. Setting $P(\lambda) = 0$ and writing its middle terms in a compact way gives the equation

$$\lambda^m + \beta\frac{\lambda^m - 1}{\lambda - 1} + \alpha = 0. \tag{25}$$

Since by part (a) $\lambda \neq 1$ for any root, the zeros of (25) are precisely those of $P(\lambda) = 0$ and also the same as the nonunit zeros of

$$\lambda^m(\lambda - 1) + \beta(\lambda^m - 1) + \alpha(\lambda - 1) = 0. \tag{26}$$

Inserting λ_0 in (26) and rearranging terms, we get

$$\lambda_0^{m+1} + (\beta - 1)\lambda_0^m + \alpha\lambda_0 = \alpha + \beta. \tag{27}$$

Setting the real parts on the two sides of (27) equal we obtain

$$r_0^{m+1}\cos(m+1)\theta_0 + (\beta-1)r_0^m\cos m\theta_0 + \alpha r_0\cos\theta_0 = \alpha + \beta. \tag{28}$$

For $\beta \geq 1$ and $\alpha > 0$ the left side of (28) is bounded above by the quantity

$$r_0^{m+1} + (\beta-1)r_0^m + \alpha r_0 < 1 + \beta - 1 + \alpha = \alpha + \beta.$$

But this contradicts (28), which was assumed to hold with $r_0 \in (0,1)$ for some θ_0. It follows that λ_0 cannot exist and thus all roots of P must be outside the unit circle.

Having shown that a unique, unstable equilibrium exists in a positively invariant region, the proof can now be concluded by Lemma 5 as soon as we show that (20) holds. For (6), the function f is given by

$$f(u_1, \dots, u_m) = \sum_{j=1}^{m} C(u_j) - A(u_m),$$

so (20) is equivalent to

$$(m-1)C(x^*) + C(x) - A(x) \neq x^* \quad \text{if} \quad x \neq x^*. \tag{29}$$

Since $x^* = F(x^*) = mC(x^*) - A(x^*)$, (29) is equivalent to

$$C(x) - A(x) \neq C(x^*) - A(x^*) \quad \text{if} \quad x \neq x^*.$$

This last inequality is true since the function $C(x) - A(x)$ is strictly decreasing on the invariant interval I. □

Remark 3. Writing (21) in the equivalent form

$$A'(x^*) > 1 + (m-2)|C'(x^*)|$$

we see that for any given value of $A'(x^*)$ larger than 1, persistent oscillations occur by Theorem 2 if $|C'(x^*)|$ is sufficiently small, i.e., if C is sufficiently flat in a neighborhood of x^*. On the other hand, if $A'(x^*) < 1$, then the steepness condition $|C'(x^*)| \geq 1$ is sufficient to guarantee the occurrence of persistent oscillations. Thus in this sense (21) and (22) are complementary conditions.

As an application of Theorem 2, consider (8) subject to the parameter values (9) except with $\omega = 2$ now. Then we may estimate $x^* \approx 3.32$ (DI of about 33.2 ms) and compute

$$A'(x^*) \approx 1.141, \quad 1 + (m-2)|C'(x^*)| \approx 1.065.$$

Thus (21) is satisfied and solutions oscillate by Theorem 2. Figure 2 shows both the spatial and temporal DI profile in six consecutive cycles of a sample trajectory (the thick curve) after the transient effects have dissipated; in addition to its oscillatory nature, this solution is quasi-periodic, a fact that does not follow from Theorem 2 but may be inferred from local analysis.

Alternatively, if we have a CT restitution curve that flattens more quickly (e.g., $\omega = 3$), then we would get the solution shown by the thin curves in Figure 2. In this case, Theorem 2 again confirms that this solution is oscillatory because

$$x^* \approx 3.29, \quad A'(x^*) \approx 1.158, \quad 1 + (m-2)|C'(x^*)| \approx 1.004.$$

We note in passing that although A is independent of ω, if C is not constant, then changes in ω affect x^* and thus the value of the slope $A'(x^*)$.

Remark 4. In [14] it is conjectured that x^* is locally asymptotically stable (or unstable) if the quantity $A'(x^*) - C'(x^*) - 1$ is negative (respectively, positive). Although neither (21) nor (22) implies the positivity of this quantity, numerical simulations and certain results such as Lemma 4 suggest that this conjecture is probably true. In the notation of this paper, we restate this open problem as follows: *All roots of the characteristic polynomial P are inside (respectively, outside) the unit disk if and only if $\alpha + \beta < 1$ (respectively, $\alpha + \beta > 1$).*

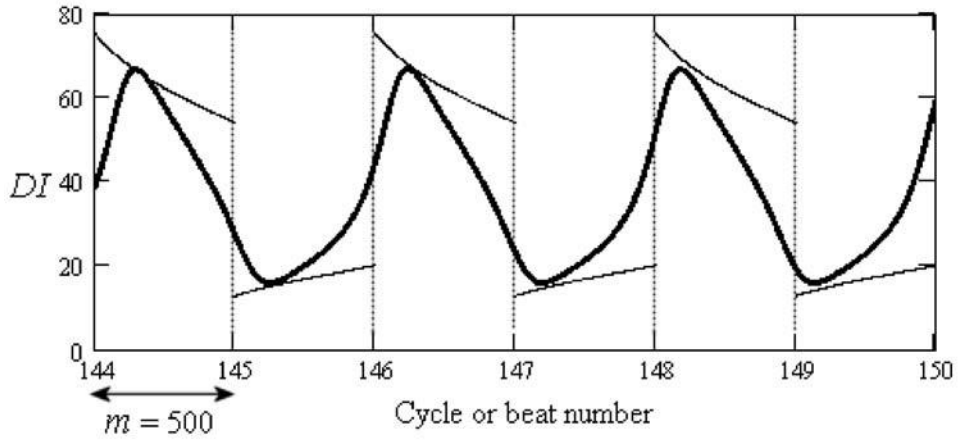


FIG. 2. *Two different types of oscillatory solution: The thick curve shows quasi-periodic oscillations when* $\omega = 2$ *and the thin curves show periodic oscillations when* $\omega = 3$. *Each point on either curve indicates the DI value of a particular unit in a given cycle. Both curves shown are generated from the same initial configuration vector* $DI_0 = [c, \ldots, c]$, *where* c *is a positive constant that may represent, e.g., the dynamic state of the loop in terms of cellular DI values just before the initiation of reentry. In each of the vertial strips each curve indicates the configuration of DI values over the entire ring (*500 *units) in one cycle or beat.*

3.5. Bistability. When cardiac tissue receives a premature stimulus such as an electrical shock of the type imparted by a defibrillator or through electrodes in a lab preparation, this can change its rhythm. In the case of a one-dimensional reentrant circuit, a particular type of behavior such as stably convergent may change to stably persistent oscillatory or conversely. Such changes can occur if (6) is bistable (or, more generally, multistable; note that a premature stimulus in cycle n changes the components of $\mathbf{DI}_n$, thus shifting the orbit in phase space). In this section we use Theorems 1 and 2 to show that a type of bistability that is caused by slight "dents" in the APD curve may exist in a general setting. An advantage of this approach is that the emergence of bistability in the higher-order equation (6) can be observed in a bifurcation diagram of the one-dimensional auxiliary map F even when C is not constant; see Remark 5.

The main idea. Consider a case where the solutions of (6) are persistently oscillatory, e.g., inequality (21) is satisfied at the equilibrium under hypotheses (A1)–(A4). Suppose that now the APD curve is made locally flat (i.e., its slope small) in the vicinity of the unstable equilibrium so that condition (A4S) holds in an interval I_0 containing x^* (A is unaltered or altered negligibly outside I_0). The endpoints of I_0 then represent an unstable 2-cycle for the auxiliary map F. Further, assume that I_0 is small enough to not contain some previously oscillatory solutions (in the sense that the oscillatory orbit in the phase space does not enter the hypercube I_0^m). Then by Theorem 1 solutions that are generated by initial values inside I_0 will converge to a new stable equilibrium without substantially affecting the previously oscillatory solutions that were outside I_0. The resulting system is thus bistable.

We emphasize that the existence of I_0 is sufficient but not necessary for the bistability of (6), unless of course the CT restitution curve C is constant. Let us now illustrate the preceding ideas using (8) subject to the parameter values (9) except that we now set $p = 0.3$ (with $\gamma = 1$, $\tau = 2.5$) to modify the APD curve, and also pick a suitable value for ω to enhance the bistability effect. The graph of $F^2(t) - t$

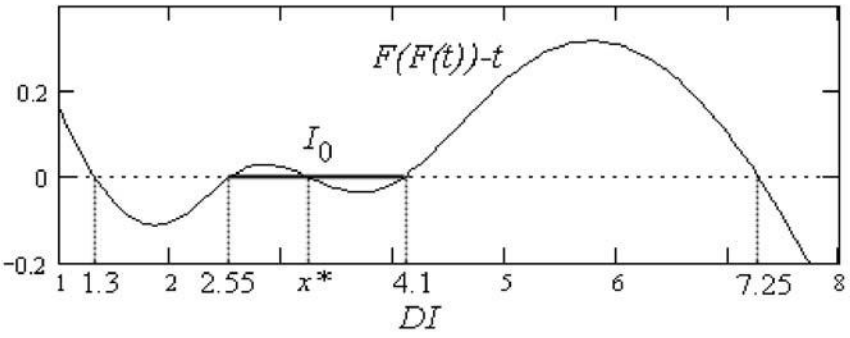


FIG. 3. *The two invariant intervals* $I = [1.3, 7.25]$ *and* $I_0 = [2.55, 4.1]$ *with the latter highlighted as the interval of convergence.*

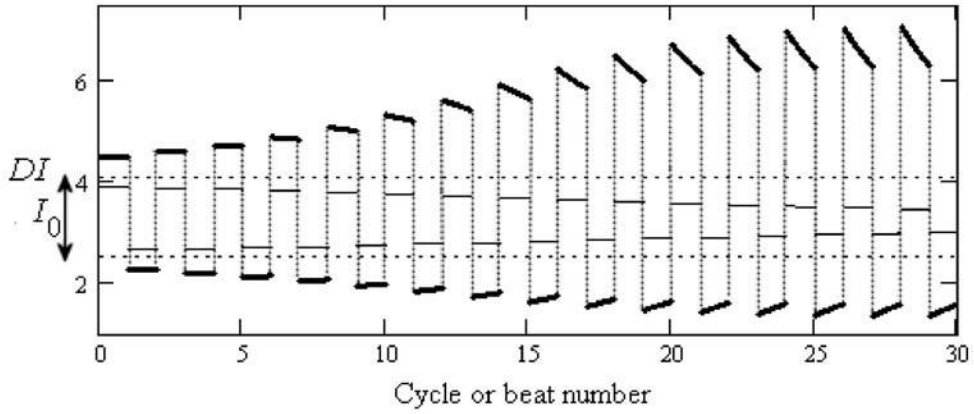


FIG. 4. *Qualitatively different outcomes that result from slightly different initial values. With initial values* $DI_0 = [3.9, \ldots, 3.9]$ *we obtain the convergent solution shown as the thin, square-shaped oscillatory curve, since* 3.9 *is in* I_0. *However, with* $DI_0 = [4.5, \ldots, 4.5]$ *we obtain the divergent, thick oscillatory curve. Note that* 4.5 *is in* I *but not in* I_0. *The horizontal strip between the dotted lines represents* I_0.

in Figure 3 shows the primary invariant interval I discussed previously, as well as the smaller one I_0 for a particular value of ω.

Figure 4 shows the different outcomes that result depending on whether we choose the initial values in I_0 or outside it. Evidently, the divergent oscillatory curve does not enter I_0 in this case. If $\mathbf{DI}_0$ is not constant and its components are only partially in I_0, then the corresponding solution may or may not converge.

More complex types of oscillatory behavior (not shown) where the trajectory may pass through I_0 repeatedly also occur in the above context with different parameter values. In such cases, the occurrence of convergent solutions may again be explained by Theorem 1 if $\mathbf{DI}_0 \in I_0^m$ (even with inhomogeneities). However, explaining the stable occurrence of the nonconvergent orbits requires further extensions of Theorem 2, or perhaps different methods that are global; local arguments and linearization in m dimensions do not apply since the equilibrium is asymptotically stable within I_0^m.

Remark 5. (emergence of bistability). Variations of shapes or changes in the relationship between the CT and APD curves may cause the emergence of bistability

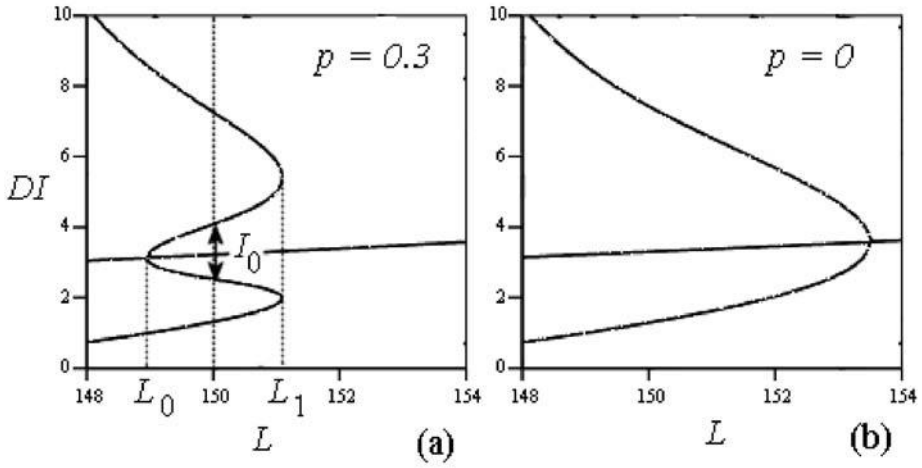


FIG. 5. *As the length L of the ring decreases, bistable behavior emerges in* (a) *but not in* (b). *The occurrences of tangent (or fold) bifurcations in* (a) *create two different period-2 orbits, the inner ones being unstable and enclosing a copy of the interval* I_0 *of Fig.* 3 *between them.*

by bringing x^* close to the region of local flatness for the APD. For illustration, consider (8) again where we change the parameter L (the length of the loop, a common bifurcation parameter). Note that $x^* = \phi^{-1}(L)$ is a strictly increasing function of L, where

$$\phi(x) = \frac{144 + 6x - 72e^{-0.5x} + 1.8e^{-(x-2.5)^2}}{1 + e^{-4x}}, \quad x > 0,$$

is easily obtained by solving (11) for L subject to the functions A, C and their parameter values for Figure 3. Thus we can reduce x^* by decreasing L and examine the effect of local APD flatness on bistability as x^* passes through the region of relative flatness for the APD curve. To visually demonstrate this effect of L on bistability we plot the two variable curve

$$F(L, F(L,x)) = x, \text{ where } F(L,x) = \frac{L}{6}[1 + e^{-4x}] - 24 + 12e^{-0.5x} - pe^{-(x-2.5)^2},$$

for both zero and nonzero (fixed) values of p. Such a bifurcation diagram shows the motion of the equilibrium for (8) as well as the occurrence and development of all period-2 points of the auxiliary map F as L varies. Figure 5 shows a comparison between two cases.

In both of the cases (a) and (b) shown in Figure 5, the nearly horizontal thick curve indicates the path of x^* as L changes. The other thick curve in Figure 5(a) where $p \neq 0$ (local flatness exists in APD) clearly shows the emergence of bistability at $L_1 \approx 151.1$ where tangent (or fold) bifurcations of the auxiliary map F create two period-2 orbits (for F). The two outer curves indicate the stable 2-cycles of F and the inner curves show the unstable 2-cycles which converge to the equilibrium x^* at $L_0 \approx 149$ through a reverse period-doubling (or flip) bifurcation. The interval between the two inner curves exists nontrivially for each $L \in (L_0, L_1)$ and gives the interval $I_0 = I_0(L)$ of bistability (by Theorem 1, x^* is asymptotically stable for all $L \in (L_0, L_1)$.) The line segment with arrows in Figure 5(a) indicates the interval I_0

at the particular value $L = 150$; this is just the interval I_0 that appears in Figures 3 and 4. By contrast, in Figure 5(b) where there is no region of relative flatness for the APD curve, the interval I_0 does not exist. For some related remarks concerning the occurrence of bistable behavior, see Vinet [25]. Vinet also uses a parameter p by which the APD function can be altered. However, the bifurcation diagrams in [25] are based on variation of p rather than L.

4. Concluding remarks. As indicated, many models of reentry in a loop of cardiac tissue or, more generally, of excitable media have been studied in the literature both in continuous and discrete form. In this paper, we refined and extended previous work using ideas and methods that yield global results. But more work remains to be done, especially at the theoretical level. For instance, it is necessary to extend Theorems 1 and 2 to cover a broader range of possibilities that include the relevant ranges of parameters in specific equations like (8), which is likely to be obtained through curve fitting of experimental data.

The interaction between Theorems 1 and 2 is a simple but not sufficiently satisfactory way to explain the occurrence of bistability. For instance, certain oscillatory solutions (periodic or almost periodic ones) repeatedly enter and exit the interval I_0 without converging, although solutions that start inside I_0 do converge (by Theorem 1). This situation requires a more detailed explanation than what is provided here.

The occurrence of almost periodic solutions has a standard description in terms of the Hopf–Neimark–Sacker bifurcation. However, it is less clear why these solutions break up into periodic solutions in some cases as the parameter ω increases from 2 to 3. A more detailed study of both periodic and almost periodic solutions of (6) is certainly desirable.

Beyond the mathematical issues mentioned above and related matters, it is also necessary to study extensions of the model that use more general types of restitution, e.g., to account for beat-to-beat memory, latency, etc. In particular, the addition of latency may lead to the occurrence of complex behavior for small values of DI.

Indeed, one sees that the persistent oscillation exhibited by the solutions of (6) is generally not complicated. (The reasons for the absence of complexity with *monotonic* APD are easy to discern when C is constant.) Can more complex types of persistent oscillations occur, e.g., with higher periods? Can they be possibly chaotic? In certain papers evidence is given that the APD restitution curve A may be nonmonotonic, e.g., be unimodal [24] or contain dents [22] at small DI values. In these works it is shown that chaotic behavior and complex bifurcations may occur whether it is for reentry in a loop [22] or for a regularly paced, nonclosed strip of tissue [24]. These observations may also be confirmed through numerical simulations of (8) with large enough p (not presented here). Such studies indicate the existence of a spatiotemporal form of chaotic behavior that needs further mathematical study and clarification.

REFERENCES

[1] R. P. Agarwal, *Difference Equations and Inequalities: Theory, Methods and Applications*, 2nd ed., Marcel-Dekker, New York, 2000.

[2] I. Banville and R. A. Gray, *Effects of action potential duration and conduction velocity restitution on alternans and the stability of arrhythmias*, J. Cardiovasc. Electrophysiol., 13 (2002), pp. 1141–1149.

[3] D. R. Chialvo and J. Jalife, *On the nonlinear equilibrium of the heart: Locking behavior and chaos in Purkinje fibers*, in Cardiac Electrophysiology: From Cell to Bedside, D. P. Zipes and J. Jalife, eds., Saunders, Philadelphia, 1990.

[4] P. Collet and J.-P. Eckmann, *Iterated Maps on the Interval as Dynamical Systems*, Birkhauser, Boston, 1980.
[5] M. Courtemanche, L. Glass, and J. P. Keener, *Instabilities of a propagating pulse in a ring of excitable media*, Phys. Rev. Lett., 70 (1993), pp. 2182–2185.
[6] M. Courtemanche, J. P. Keener, and L. Glass, *A delay equation representation of pulse circulation on a ring in excitable media*, SIAM J. Appl. Math., 56 (1996), pp. 119–142.
[7] M. Courtemanche and A. Vinet, *Reentry in excitable media*, in Nonlinear Dynamics in Physiology and Medicine, A. Beuter, L. Glass, M. C. Mackey, and M. S. Titcombe, eds., Springer, New York, 2003.
[8] S. N. Elaydi, *An Introduction to Difference Equations*, 2nd ed., Springer, New York, 1999.
[9] J. J. Fox, E. Bodenschatz, and R. F. Gilmour, *Period-doubling instability and memory in cardiac tissue*, Phys. Rev. Lett., 89 (2002), pp. 8101–8104.
[10] L. H. Frame and M. B. Simson, *Oscillations of conduction, action potential duration and refractoriness*, Circulation, 78 (1988), pp. 2182–2185.
[11] R. F. Gilmour, N. F. Otani, and M. A. Watanabe, *Memory and complex dynamics in cardiac Purkinje fibers*, Amer. J. Physiol., 272 (1997), pp. H1826–H1832.
[12] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurctions of Vector Fields*, Springer, New York, 1983.
[13] G. M. Hall, S. Bahar, and D. J. Gautheir, *Prevalence of rate-dependent behaviors in cardiac muscle*, Phys. Rev. Lett., 82 (1999), pp. 2995–2998.
[14] H. Ito and L. Glass, *Theory of reentrant excitation in a ring of cardiac tissue,* Phys. D, 56 (1992), pp. 84–106.
[15] S. S. Kalb, H. M. Dobrovolny, E. G. Tolkacheva, S. F. Idriss, W. Krassowska, and D. J. Gauthier, *The restitution portrait: A new method for investigating rate-dependent restitution*, J. Cardiovasc. Electrophysiol., 15 (2004), pp. 698–709.
[16] J. P. Keener, *Arrhythmias by dimension*, in *An Introduction to Mathematical Modeling in Physiology, Cell Biology and Immunology*, J. Sneyd, ed., pp. 57–81, American Mathematical Society, Providence, RI, 2002.
[17] W. G. Kelley and A. C. Peterson, *Difference Equations: An Introduction with Applications*, 2nd ed., Harcourt/Academic Press, San Diego, 2001.
[18] V. L. Kocic and G. Ladas, *Global Behavior of Nonlinear Difference Equations of Higher Order with Applications*, Kluwer, Dordrecht, The Netherlands, 1993.
[19] V. Lakshmikantham and D. Trigiante, *Theory of Difference Equations: Numerical Methods and Applications*, 2nd ed., Marcel-Dekker, New York, 2002.
[20] R. Mickens, *Difference Equations: Theory and Applications*, 2nd ed., CRC Press, Boca Raton, FL, 1991.
[21] N. F. Otani and R. F. Gilmour, *Memory models for the electrical properties of local cardiac systems*, J. Theor. Biol., 187 (1997), pp. 409–436.
[22] Z. Qu, J. N. Weiss, and A. Garfinkel, *Spatiotemporal chaos in a simulated ring of cardiac cells*, Phys. Rev. Lett., 78 (1997), pp. 1387–1390.
[23] H. Sedaghat, *Nonlinear Difference Equations: Theory with Applications to Social Science Models*, Kluwer, Dordrecht, The Netherlands, 2003.
[24] M. D. Stubna, R. H. Rand, and R. F. Gilmour, *Analysis of a nonlinear partial difference equation, and its application to cardiac dynamics*, J. Differ. Equations Appl., 8 (2002), pp. 1147–1169.
[25] A. Vinet, *Quasiperiodic circus movement in a loop model of cardiac tissue: Multistability and low dimensional equivalence*, Ann. Biomed. Engrg., 28 (2000), pp. 704–720.
[26] M. A. Watanabe and M. L. Koller, *Mathematical analysis of dynamics of cardiac memory and accomodation: Theory and experiment,* Amer. J. Physiol., 282 (2002), pp. H1534–H1547.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 591–615

THE CONVEX BACK-SCATTERING SUPPORT*

HOUSSEM HADDAR†, STEVEN KUSIAK‡, AND JOHN SYLVESTER§

Abstract. A monochromatic, i.e., fixed-frequency, back-scattering kernel measured at all angles does not uniquely determine the index of refraction in an inhomogeneous medium, nor can it guarantee any upper bound on the support of the inhomogeneity. We show that it is possible to associate with any such kernel its *convex back-scattering support*, a convex set which must be a subset of the convex hull of the support of any inhomogeneity with that back-scattering kernel. For the Born approximation, we further demonstrate that there is an inhomogeneity supported in any neighborhood of the convex back-scattering support which has exactly that back-scattering kernel. Last, we discuss a practical implementation of these results and include a numerical example.

Key words. Helmholtz equation, inverse scattering, back-scattering, acoustic scattering, electromagnetic scattering, far field

AMS subject classifications. 81U40, 74J25, 65N21

DOI. 10.1137/040616231

1. Introduction.

The goal of inverse scattering is to use acoustic or electromagnetic waves to deduce properties of a scatterer from remote observations. Exactly which properties and how well we can deduce them depend on exactly what scattering data we measure. In this paper the measured data are the back-scattered far field at all angles and a single frequency. A back-scattering experiment requires only a single sensor, which acts as both source and receiver. The sensor radiates at a single temporal frequency and measures the amplitude and phase of the resulting time-harmonic field. We can move the sensor to an arbitrary location on a sphere of large radius that surrounds the scatterer and repeat the experiment. The complex field (amplitude and phase) measured at each point on the sphere is the back-scattering data.

We will use the Helmholtz equation as our model for the propagation of time-harmonic waves,

$$(\Delta + k^2 n^2(x))u = 0, \quad x \in \mathbb{R}^d, \quad d \geq 2. \tag{1.1}$$

Here, $n(x) = \frac{c_0}{c(x)}$ is the index of refraction, which is the ratio of the wave speed in the vacuum to that in the medium, while $k = 2\pi/\lambda$ denotes the wave number. It will be convenient to define the scattering potential, $q(x) := k^2(1 - n^2)$, and to rewrite

*Received by the editors October 3, 2004; accepted for publication (in revised form) July 20, 2005; published electronically December 30, 2005.
http://www.siam.org/journals/siap/66-2/61623.html

†Laboratoire POems, UMR 2706 CNRS/ENSTA/INRIA, INRIA, Domaine de Voluceau-Rocquencourt, BP 105, 78153 Le Chesnay cedex, France (houssem.haddar@inria.fr).

‡MIT Lincoln Laboratory, Lexington, MA 02420-9180 (kusiak@ll.mit.edu). This work was completed while this author was a CNRS postdoctoral fellow at Laboratoire POems, UMR 2706 CNRS/ENSTA/INRIA, ENSTA, 32 Boulevard Victor, 75739 Paris cedex 15, France.

§Department of Mathematics, University of Washington, Seattle, WA 98195 (sylvest@math.washington.edu). This author's research was supported by NSF grant DMS-0099838 and ONR grant N00014-93-1-0295.

(1.1) as

$$(\Delta + k^2)u = qu, \tag{1.2}$$

$$u = e^{ik\Theta\cdot x} + u_{sc}, \tag{1.3}$$

$$u_{sc} \sim \frac{e^{ikr}}{r^{\frac{d-1}{2}}} s_q(\Phi, \Theta), \quad |x| = r \to \infty. \tag{1.4}$$

Equation (1.3) expresses the fact that although the sensor, located at $\Theta \in S^{d-1}$, acts as a point source and emits a spherical wave, the sphere is far enough from the scatterer that the incident wave appears to be a plane wave in a neighborhood of the scatterer. Equation (1.4) expresses the asymptotics of the *outgoing* scattered wave.[1] The field measured by an additional dislocated sensor positioned at $\Phi \in S^{d-1}$ is denoted by $s_q(\Phi, \Theta)$. This quantity is called the *scattering kernel*. We define the back-scattering kernel as $s(\Theta) := s_q(\Theta, -\Theta)$. The back-scattering kernel represents the complex time-harmonic scattered field measured at the source position. We will discuss both the back-scattering kernel and what we will call the *Born-back-scattering* kernel. We denote the latter quantity by $b(\Theta) := b_q(\Theta, -\Theta)$ and remark that it is the analogue of $s(\Theta)$ in the Born, or single scattering, approximation. In both cases, a main feature is that the fixed frequency back-scattering data does not uniquely determine the scattering potential $q(x)$. Indeed, the Born-back-scattering kernel is equal to a constant (in Θ) multiple of the Fourier transform of q, restricted to the sphere of radius $2k$, i.e.,

$$b(\Theta) = \frac{k^{d-2}}{2i} \widehat{q}(2k\Theta). \tag{1.5}$$

We learn from (1.5) that as long as $g(x)$ is a smooth compactly supported function, then $\tilde{q} = (\Delta + (2k)^2)g$ has zero Born-back-scattering kernel, so that

$$b_{q+\tilde{q}}(\Theta) = b_q(\Theta). \tag{1.6}$$

The theorems below will relate the back-scattering data to the support of q. A glance at (1.6) makes it clear that it is impossible to produce an upper bound for the $\operatorname{supp} q$. We will, however, compute a lower bound. We will associate (and compute numerically) with $b(\Theta)$, or $s(\Theta)$, its convex back-scattering support, a convex set which must be a subset of the convex hull of the support of any q which produces that back-scattering data. In the Born approximation we will also find a potential supported in any neighborhood of this set that reproduces the data. In this case, the convex back-scattering support is the unique smallest convex set that supports a potential that can produce this data.

We state all our theorems below, and we include only very brief descriptions of the necessary notation here and defer both their detailed discussion and proofs to the following sections. In what follows, the symbol $\sigma_n(2kR)$ denotes the L^2 norm of the d-dimensional Bessel function of order n, and argument $2k|x|$, restricted to the ball of radius R. The number $N = N(n, d) \approx n^{d-2}$ denotes the dimension of the space of spherical harmonics of degree n. For a function $b \in L^2(S^{d-1})$, the functions $b_n^{(c)}(\Theta)$ represent the terms in the *condensed spherical harmonic expansion* (Fourier series expansion in two dimensions) of the function $e^{2i\Theta\cdot c} b(\Theta)$. Additionally, $B_c(R)$ denotes the closed ball of radius R centered at the point $c \in \mathbb{R}^d$.

[1] We will give a more precise mathematical description in the next section.

We now state our main results.

THEOREM 1.1 (linear, born-back-scattering). *Let $q \in L^2(\mathbb{R}^d)$ be compactly supported with Born-back-scattering kernel $b(\Theta)$, i.e.,*

$$b(\Theta) := b_q(\Theta, -\Theta) = \frac{k^{d-2}}{2i}\widehat{q}(2k\Theta), \tag{1.7}$$

and let

$$e^{2ik\Theta\cdot c}b(\Theta) = \sum_{n=0}^{\infty} b_n^{(c)}(\Theta)$$

be its expansion in spherical harmonics centered at c. If the $\operatorname{supp} q \subset B_c(R)$, *then*

$$\sum_{n=0}^{\infty} \frac{||b_n^{(c)}||^2_{L^2(S^{d-1})}}{\sigma_n^2(2kR)} < \infty. \tag{1.8}$$

Conversely, if $b \in L^2(S^{d-1})$ satisfies (1.8), *then there exists a $q \in L^2(\mathbb{R}^d)$ with* $\operatorname{supp} q \subset B_c(R)$ *and Born-back-scattering kernel $b(\Theta)$.*

Moreover, if $b \in L^2(S^{d-1})$ and $\mathcal{B}$ denotes the collection of all balls, $B_c(R)$, for which (1.8) *is satisfied, then for any $\epsilon > 0$, there exists $q \in C^\infty(\mathbb{R}^d)$ with*[2]

$$\operatorname{supp} q \subset \mathcal{N}_\epsilon\left(\bigcap_{B_c(R)\in\mathcal{B}} B_c(R)\right) \tag{1.9}$$

and Born-back-scattering kernel equal to $b(\Theta)$.

THEOREM 1.2 (nonlinear, full back-scattering). *Let $q \in L^p(\mathbb{R}^d)$, where $p > \max(2, \frac{d}{2})$, be compactly supported with back-scattering kernel*

$$s(\Theta) := s_q(\Theta, -\Theta),$$

and let

$$e^{2ik\Theta\cdot c}s(\Theta) = \sum_{n=0}^{\infty} s_n^{(c)}(\Theta)$$

be its expansion in spherical harmonics centered at c. If the $\operatorname{supp} q \subset B_c(R)$, *then, for $n > \max(4R, 2R^2)$,*

$$||s_n^{(c)}||^2_{L^2(S^{d-1})} \le C(q)\ N\sigma_n^2(2kR), \tag{1.10}$$

where $C(q)$ is given explicitly in (4.13).

If we observe that for any $\epsilon > 0$, (1.10) implies (1.8) with R replaced by $R + \epsilon$, we obtain the following corollary.

COROLLARY 1.3. *If the* $\operatorname{supp} q \subset B_c(R)$, *then there is a (complex-valued) $\tilde{q} \in C^\infty(\mathbb{R}^d)$ supported in an ϵ-neighborhood of $B_c(R)$ with Born-back-scattering kernel exactly equal to the full back-scattering kernel of q, i.e.,*

$$b_{\tilde{q}}(\Theta) = s_q(\Theta).$$

[2] $\mathcal{N}_\epsilon(\Omega)$ denotes an open ϵ-neighborhood of Ω, the set of points whose distance from Ω is less than ϵ.

As a consequence of these theorems, we define the convex scattering support to be as follows.

DEFINITION 1.4 (convex scattering support).

$$\mathrm{cS}_k\mathrm{supp}\, b = \bigcap_{B_c(R)\in\mathcal{B}} B_c(R),$$
$$\mathrm{cS}_k\mathrm{supp}\, s = \bigcap_{B_c(R)\in\mathcal{S}} B_c(R),$$

where $\mathcal{B}$ denotes the collection of balls such that b satisfies (1.8) *and $\mathcal{S}$ is the collection of balls such that s satisfies* (1.10).

We record the main property of the convex scattering support below.

THEOREM 1.5. *In both the Born approximation and in the full back-scattering cases, the convex scattering support must be a subset of the convex hull of the support of any potential q with that back-scattering kernel.*

Proof. Note that any point that does not belong to the convex hull of the support of q can be separated from that set by a ball. The Hahn–Banach theorem tells us there is a separating hyperplane, and because this convex set is bounded (q has compact support), a large enough ball will approximate the hyperplane as close as necessary on any compact set and thus accomplish the separation. ∎

In the Born approximation, the linear dependence on q allows a stronger statement.

THEOREM 1.6. *For any $b \in L^2(S^{d-1})$, the convex scattering support of the Born back-scattering kernel b is*

1. *the smallest convex set such that there is a q, supported in every neighborhood of that set, with Born back-scattering kernel b;*
2. *the largest convex set that is contained in the convex hull of the support of every q with Born back-scattering kernel equal to b.*

A simple consequence of Theorem 1.6 is that the convex scattering support of a nonzero Born back-scattering kernel is nonempty. A consequence of Corollary 1.3 is that this same conclusion holds for the full back-scattering kernel.

We introduced the convex scattering support of a far field of a source in [8] and [9]. We extended the notion to include fields scattered by a variation in wave speed or an obstacle, in response to a single incident wave, and described one method to compute it. Different methods that compute roughly the same set have also appeared in [10], [12], [13], [4], [6], and [5].

Our main reason for extending the notion of convex scattering support to back-scattering is that the data are more realistically acquired, and therefore, the methods can be more readily adapted to be of practical use. To measure the far field of a single incident wave requires at least two sensors, one at a fixed position that radiates and another that will move around the scatterer to measure the field. Calibration requires accurate knowledge of the angle between the two devices, while back-scattering simply requires moving a single sensor.

It may appear that a condition like (1.10) is difficult to apply to noisy data. Exactly the opposite is true. The coefficients in a Fourier series or spherical harmonic expansion are easily computed. As we will point out in the final section, the function $\sigma_n(2kR)$, viewed as a function of n at a fixed value of $2kR$, is uniformly large for $n < 2kR$ and rapidly becomes uniformly small when $n > 2kR$. Thus, the norm of the terms in the spherical harmonic expansion for $e^{i2k\Theta\cdot c}s(\Theta)$ undergo a rapid transition

to, and become approximately, zero as n passes through the radius of the smallest ball that contains the convex scattering support. This transition is readily observable and remains so in the presence of appreciable noise.

2. The Herglotz, far field, and scattering operators. For any $\delta > \frac{1}{2}$, the unique $L^2_{-\delta}$ solution[3] of the free (homogeneous) Helmholtz equation, parametrized by α, satisfies

$$(\Delta + k^2)v = 0, \tag{2.1}$$

$$v \sim \frac{e^{ikr}}{r^{\frac{d-1}{2}}}\alpha(-\Theta) + \frac{e^{-ikr}}{r^{\frac{d-1}{2}}}\alpha(\Theta), \quad |x| = r \to \infty, \tag{2.2}$$

and may be expressed in terms of the Herglotz operator

$$v(r\Phi) = (\mathcal{H}\alpha)(r\Phi) := k^{\frac{d-1}{2}} \int_{S^{d-1}} e^{ikr\Theta\cdot\Phi}\alpha(\Theta)d\Theta.$$

The adjoint of the Herglotz operator, which maps $L^2_\delta(\mathbb{R}^d)$ into $L^2(S^{d-1})$, is therefore the Fourier transform (times a power of k), followed by restriction to the sphere of radius k, and may be written as

$$(\mathcal{H}^* f)(\Theta) = k^{\frac{d-1}{2}} \int_{\mathbb{R}^d} e^{-ikr\Theta\cdot\Phi} f(r\Phi) r^{d-1} dr d\Phi = k^{\frac{d-1}{2}} \widehat{f}(k\Theta). \tag{2.3}$$

Here we have simply written $d\Phi$ and $d\Theta$ rather than $dS(\Phi)$ and $dS(\Theta)$ to denote the surface measure on the unit sphere S^{d-1}.

We note that for any right-hand-side $f \in L^2_\delta$, with $\delta > 1/2$, the source problem

$$(\Delta + k^2)u = f \tag{2.4}$$

has a unique *outgoing* solution. Such an *outgoing* solution has asymptotics similar to those in (2.2); however, the second term vanishes. Specifically,

$$u \sim \frac{e^{ikr}}{r^{\frac{d-1}{2}}}\beta(\Theta) + \frac{e^{-ikr}}{r^{\frac{d-1}{2}}} \times 0, \quad |x| = r \to \infty. \tag{2.5}$$

The function $\beta \in L^2(S^{d-1})$ is called the far field of the outgoing solution u. We will use the notation

$$u = Gf$$

to denote the solution operator which solves (2.4). Additionally, we define the far field operator $\mathcal{F}$ as the map between the source f and the far field β, so that

$$\beta = \mathcal{F}f.$$

We will make use of the fact that, except for a factor of $2ik$, the far field operator is the adjoint of the Herglotz operator, as follows.

[3] $||u||_{L^2_\delta(\mathbb{R}^d)} = ||(1+|x|^2)^{\frac{\delta}{2}}u||_{L^2(\mathbb{R}^d)}$. These spaces were first used in the context of the Helmholtz equation in [1] as a means for studying long-range potentials. The existence and uniqueness statements we quote here can be found in [8].

PROPOSITION 2.1.

$$\mathcal{F}f = \frac{1}{2ik}\mathcal{H}^* f = \frac{k^{\frac{d-3}{2}}}{2i}\widehat{f}(k\Theta).$$

Proof. The second equality is a consequence of the first and (2.3). To establish the first, we apply Green's formula to u and v as previously defined in (2.4) and (2.1), i.e.,

$$\begin{aligned}(\mathcal{H}\alpha, f) &= (v, f)\\ &= \int_{\mathbb{R}^d} \overline{v} f\\ &= \int_{\mathbb{R}^d} \overline{v}(\Delta + k^2)u - \overline{(\Delta + k^2)v}u\\ &= \lim_{R\to\infty} \int_{S_R^{d-1}} \left(\overline{v}\frac{\partial u}{\partial \nu} - \frac{\overline{\partial v}}{\partial \nu}u\right) d\Theta,\end{aligned}$$

which becomes, on inserting the asymptotics from (2.2) and (2.5),

$$\begin{aligned}&= 2ik\int_{S^{d-1}} \overline{\alpha}\beta d\Theta\\ &= (\alpha, 2ik\mathcal{F}f). \qquad \square\end{aligned}$$

To define the scattering operator, we return to (1.2),

$$(\Delta + k^2)u = qu,$$

and seek u as an *outgoing* perturbation of the solution of the free Helmholtz equation with the incident field $\mathcal{H}\alpha$ so that

$$u = \mathcal{H}\alpha + u_{sc},$$

where u_{sc} is an outgoing solution. This means that u_{sc} is the unique outgoing solution of

$$(\Delta + k^2)u_{sc} = q\mathcal{H}\alpha + qu_{sc}. \tag{2.6}$$

Such an outgoing field has the asymptotics

$$u_{sc} \sim \frac{e^{ikr}}{r^{\frac{d-1}{2}}}\beta_q(\Theta) + \frac{e^{-ikr}}{r^{\frac{d-1}{2}}} \times 0.$$

This observation allows us to define the relative scattering operator

$$\mathcal{S}\alpha = \beta_q.$$

The Born approximation replaces (2.6) with

$$(\Delta + k^2)u_{born} = q\mathcal{H}\alpha. \tag{2.7}$$

The unique outgoing solution has the asymptotics

$$u_{born} \sim \frac{e^{ikr}}{r^{\frac{d-1}{2}}}\beta_{born}(\Theta) + \frac{e^{-ikr}}{r^{\frac{d-1}{2}}} \times 0$$

and so we define the (relative) Born scattering operator

$$\mathcal{B}\alpha = \beta_{born}.$$

The Born scattering operator is the Fréchet derivative of the relative scattering operator with respect to q, evaluated at $q \equiv 0$. The factorizations below (similar to those in [7]) will enable us to derive useful properties of the scattering operator from analogous properties of the Herglotz operator.

PROPOSITION 2.2. *The full relative scattering operator admits the factorization*

$$\mathcal{S} = \frac{1}{2ik}\mathcal{H}^* q(I - Gq)^{-1}\mathcal{H}, \tag{2.8}$$

$$= \frac{1}{2ik}\mathcal{H}^*(I - qG)^{-1}q\mathcal{H}, \tag{2.9}$$

while the Born relative scattering operator may be decomposed as

$$\mathcal{B} = \frac{1}{2ik}\mathcal{H}^* q\mathcal{H}. \tag{2.10}$$

The kernel of the Born relative scattering operator is

$$b(\Theta, \Phi) = \frac{k^{d-2}}{2i}\widehat{q}(k(\Theta - \Phi)). \tag{2.11}$$

Proof. The proof below relies on the invertibility of $(I - qG)$, the proof of which we defer to Lemma 4.1. We begin with (2.6) and apply G to both sides,

$$\begin{aligned} u_{sc} &= G(q\mathcal{H}\alpha + qu_{sc}), \\ (I - Gq)u_{sc} &= Gq\mathcal{H}\alpha, \\ u_{sc} &= (I - Gq)^{-1}Gq\mathcal{H}\alpha, \\ &= G(I - qG)^{-1}q\mathcal{H}\alpha \end{aligned}$$

so that the far field of u_{sc} is

$$\begin{aligned} \mathcal{S}\alpha &= \mathcal{F}(I - qG)^{-1}q\mathcal{H}\alpha \\ &= \frac{1}{2ik}\mathcal{H}^*(I - qG)^{-1}q\mathcal{H}\alpha, \end{aligned}$$

establishing (2.8). The analogous calculation applied to (2.7) instead of (2.6) establishes (2.10). Once we know that $(I - qG)$ is invertible, the identity

$$(I - qG)^{-1}q = q(I - Gq)^{-1}$$

follows from

$$q(I - Gq) = (I - qG)q,$$

which transforms (2.9) into (2.8). Finally, writing the integral representation of (2.10) as

$$\begin{aligned} \mathcal{B}\alpha &= \frac{k^{d-2}}{2i}\int_{\mathbb{R}^d} e^{-irk\Psi\cdot\Phi}q(r\Psi)\left[\int_{S^{d-1}} e^{irk\Theta\cdot\Psi}\alpha(\Theta)d\Theta\right] r^{d-1}drd\Psi \\ &= \frac{k^{d-2}}{2i}\int_{S^{d-1}}\left[\int_{\mathbb{R}^d} e^{-irk(\Phi-\Theta)\cdot\Psi}q(r\Psi)r^{d-1}drd\Psi\right]\alpha(\Theta)d\Theta \end{aligned}$$

yields (2.11). ∎

3. The Herglotz operator and the spherical harmonics. In this and the next sections, the notation will be slightly less cluttered if we restrict to the case $k = 1$. Because of the representations of the scattering operator in (2.8) and (2.9), the properties of the Herglotz operator will figure prominently into our analysis of the scattering operator. The singular value decomposition of the Herglotz operator in terms of the spherical harmonics and (spherical) Bessel functions will provide a basic tool for all our subsequent calculations. We begin with the expansion of an incident plane wave in spherical harmonics, which for $\Theta, \Phi \in S^{d-1}$ and $0 \leq r < \infty$ is

$$e^{ir\Theta\cdot\Phi} = \sum_{n=0}^{\infty} i^n j_n(r) p_n(\Theta \cdot \Phi).$$

The most useful way to define the functions p_n in our context is as the kernel of the orthogonal projection, $\mathcal{P}_n$, from $L^2(S^{d-1})$ onto the subspace of degree n spherical harmonics, i.e.,

$$\left(\mathcal{P}_n \alpha\right)(\Theta) = \int_{S^{d-1}} p_n(\Theta \cdot \Phi)\alpha(\Phi) d\Phi.$$

The functions $p_n(\Theta \cdot \Phi)$ play a dual role: they act as both kernels of projection operators and are themselves spherical harmonics. When we wish to emphasize their second role we will write

$$p_n^{\Phi}(\Theta) := p_n(\Theta \cdot \Phi).$$

Up to a constant, the function p_n^{Φ} is the unique spherical harmonic that is invariant under rotations about the Φ axis [11]. The constants involved here will be important to us, hence we compute

$$\begin{aligned} \mathcal{P}_n p_n^{\Phi} &= p_n^{\Phi}, \\ \int_{S^{d-1}} p_n^{\Psi}(\Theta) p_n^{\Phi}(\Theta) d\Theta &= p_n^{\Phi}(\Psi), \end{aligned}$$

from which we learn that

$$||p_n^{\Psi}||_{L^2}^2 = p_n^{\Psi}(\Psi)$$

is independent of Ψ and

$$= ||p_n^{\Psi}||_{L^\infty}.$$

If we take the trace of the operator $\mathcal{P}_n$, we see that

$$\begin{aligned} \mathrm{tr}\mathcal{P}_n &= \int_{S^{d-1}} p_n(\Theta \cdot \Theta) d\Theta, \\ N &= p_n(\Theta \cdot \Theta)\ \omega, \end{aligned}$$

where ω denotes the volume of the d-dimensional sphere and $N = N(n, d)$ the dimension of the space of spherical harmonics of degree n. Hence, we conclude that

$$||p_n^{\Phi}||_{L^2}^2 = ||p_n^{\Phi}||_{L^\infty} = p_n^{\Phi}(\Phi) = \frac{N}{\omega}.$$

In terms of the standard Legendre functions, P_n, we have the relation

$$p_n(\Theta \cdot \Phi) = \frac{N}{\omega} P_n(\Theta \cdot \Phi).$$

The Bessel functions $j_n(r)$ are most easily defined in terms of the Herglotz operator acting on the p_n. Recalling that $p_n^\Phi = p_n(\Theta \cdot \Phi)$ acts as both a spherical harmonic and the kernel of a projection, we see that

$$\begin{aligned} \mathcal{H} p_n^\Phi &= \int_{S^{d-1}} e^{ir\Theta\cdot\Psi} p_n(\Theta \cdot \Phi) d\Theta \\ &= \mathcal{P}_n e^{ir\Theta\cdot\Psi} \end{aligned} \tag{3.1}$$

so that $\mathcal{H} p_n^\Phi$, for each fixed r, must again be a spherical harmonic of degree n. We can check that the right-hand side of (3.1) is invariant under rotations about the Φ-axis and thus must be equal to a constant multiple of p_n^Φ itself. Specifically, this constant is i^n times the spherical Bessel function, i.e.,

$$\mathcal{H} p_n^\Phi = i^n j_n(r) p_n^\Phi. \tag{3.2}$$

Because the p_n's act also as projection kernels, we see that any spherical harmonic of degree n may replace p_n^Φ in (3.2). The Herglotz operator is not compact. However, if we compose it with the restriction to the ball of radius R, the composition is compact. We denote the resulting operator by $\mathcal{H}_R$ and describe its singular value decomosition.

$$\begin{aligned} \mathcal{H}_R &= \sum_{n=0}^{\infty} \sigma_n(R) \frac{j_n^R(r)}{||j_n^R||} \mathcal{P}_n \\ &= \sum_{n=0}^{\infty} \sigma_n(R) \mathcal{Q}_n. \end{aligned} \tag{3.3}$$

We have used the notation j_n^R to denote the Bessel function multiplied by the characteristic function of $B_0(R)$, the ball of radius R centered at 0, and defined the singular values

$$\sigma_n^2(R) := ||j_n||^2_{L^2(B_0(R))}.$$

Each projection operator

$$\mathcal{Q}_n := \frac{j_n^R(r)}{||j_n^R||} \mathcal{P}_n$$

is an isometry from the N-dimensional space of spherical harmonics of degree n in $L^2(S^{d-1})$ to an N-dimensional subspace of $L^2(B_0(R))$. In short, we find that the $\mathcal{Q}_n$ simply project onto the spherical harmonics of degree n and then multiply the result by $\frac{j_n^R(r)}{||j_n^R||}$.

Since the singular values all have multiplicity greater than one, this looks a little different from the more familiar version of the singular value decomposition. A compact linear operator K admits the representation

$$\begin{aligned} K &= \sum_{n=0}^{\infty} \lambda_n \Psi_n \otimes \Phi_n \\ &= \sum_{n=0}^{\infty} \lambda_n \tilde{\mathcal{Q}}_n. \end{aligned} \tag{3.4}$$

In this case, Ψ_n and Φ_n are orthonormal basis vectors so that the tensor products $\tilde{\mathcal{Q}}_n = \Psi_n \otimes \Phi_n$ are isometries between one-dimensional subspaces.

The corresponding singular value decomposition for the operator ${\mathcal{H}_R}^*$ is

$$ {\mathcal{H}_R}^* = \sum_{n=0}^{\infty} \sigma_n(R) \mathcal{Q}_n^*, \tag{3.5} $$

and the ranges of the $\mathcal{Q}_n^*$ are exactly the subspaces of spherical harmonics of degree n. Note that as no $\sigma_n(R)$ is zero, ${\mathcal{H}_R}^*$ has dense range in $L^2(S^{d-1})$.

Now, we recall the following.

THEOREM 3.1 (Picard's theorem). *If $K : X \to Y$ is a compact linear operator with dense range in Y and has a singular value decomposition of the form given in* (3.4)*, then*

$$ \alpha = Kf $$

if and only if

$$ \sum_{n=0}^{\infty} \frac{||\mathcal{Q}_n \alpha||^2}{\lambda_n^2} < \infty $$

and $\alpha \in N(K^*)^{\perp}$.

The proof of Theorem 1.1 is now in hand.

Proof of Theorem 1.1. The second equality in (1.7) follows from (2.11) on setting $\Phi = -\Theta$. Now, scaling the Fourier transform gives

$$ b(\Theta) = \frac{1}{2i} \widehat{q}(2\Theta) = 2^{-d} \mathcal{H}^* q(x/2). $$

The Picard theorem applied to $\mathcal{H}_R^*$ tells us that in the case $c = 0$, if the $\operatorname{supp} q \subset B_c(R)$, then b satisfies (1.8). It also tells us the converse, that if b satisfies (1.8), there exists a $q \in L^2(\mathbb{R}^d)$ with $\operatorname{supp} q \subset B_c(R)$ and Born-back-scattering kernel $b(\Theta)$.

The Fourier shift theorem tells us that the Fourier transform—and therefore $\mathcal{H}^*$ and $\mathcal{H}_R^*$—intertwines translation by c and multiplication by $e^{ik\Theta \cdot c}$, i.e.,

$$ e^{ik\Theta \cdot c} \mathcal{H}^* q = \mathcal{H}^* T_c q := \mathcal{H}^* q(x - c), $$

which establishes the corresponding conclusions for arbitrary c.

So far we have shown that every ball for which (1.8) is satisfied supports a q with Born-back-scattering kernel b. We now wish to demonstrate that any open neighborhood of their intersection supports such a q as well. As a consequence of Lemma 3.2, given below, we find that if each of two convex sets support potentials q with corresponding Born-back-scattering kernel b, then so must any neighborhood of their intersection.

LEMMA 3.2. *Suppose the* $\operatorname{supp}\ q_1 \subset \Omega_1$*, the* $\operatorname{supp}\ q_2 \subset \Omega_2$*, and that $\mathbb{R}^d \setminus (\Omega_1 \bigcup \Omega_2)$ is connected and contains a neighborhood of ∞. If*

$$ \mathcal{H}^* q_1 = \mathcal{H}^* q_2 = b, \tag{3.6} $$

then, for any $\varepsilon > 0$, there exists an $q_3 \in C^{\infty}(\mathbb{R}^d)$ with

$$ \operatorname{supp} q_3 \subset \mathcal{N}_{\varepsilon}(\Omega_1 \cap \Omega_2) $$

and

$$\mathcal{H}^* q_3 = b.$$

Proof. A consequence of (3.6) is that the outgoing solutions of

$$\left(\Delta + (2k)^2\right) u_i = q_i, \quad i = 1, 2,$$

$u_1 = Gq_1$, and $u_2 = Gq_2$ have the same far field. According to Rellich's lemma and the unique continuation principle [2], u_1 and u_2 also agree on the $\mathbb{R}^d \setminus (\Omega_1 \bigcup \Omega_2)$. Let $\phi \in C^\infty(\mathbb{R}^d)$ satisfy

$$\phi = \begin{cases} 1, & x \in \mathbb{R}^n \backslash \mathcal{N}_\varepsilon(\Omega_1 \cap \Omega_2), \\ 0, & x \in \mathcal{N}_{\frac{\varepsilon}{2}}(\Omega_1 \cap \Omega_2); \end{cases}$$

then,

$$v = \begin{cases} \phi u_1, & x \in \mathbb{R}^d \backslash \Omega_1, \\ \phi u_2, & x \in \mathbb{R}^d \backslash \Omega_2, \\ 0, & x \in \Omega_1 \cap \Omega_2, \end{cases}$$

is a well-defined C^∞ function and $v = u_1 = u_2$ outside a compact set so that

$$q_3 = (\Delta + (2k)^2) v$$

must also have $\mathcal{H}^* q_3 = b$. ∎

Because the intersection of convex sets is convex, and the complement of the union of two convex sets must be connected, the lemma can be applied repeatedly to produce a $q \in C^\infty$ satisfying the (apparently weaker) analogue of (1.9) with $\mathcal{B}$ replaced by any finite collection of balls, $B_c(R)$, for which (1.8) holds. The following compactness argument shows that (1.9) follows from this analogue. Let R_0 be large enough that $\mathcal{N}_\epsilon(\cap_{B_c(R) \in \mathcal{B}} B_c(R)) \subset B_0(R_0)$, then $B_0(R_0) \setminus \mathcal{N}_\epsilon(\cap_{B_c(R) \in \mathcal{B}} B_c(R))$ is a compact set covered by the relatively open subsets $B_0(R_0) \setminus B_c(R)$, so a finite subcollection, $\mathcal{B}_\mathcal{M}$, of these open subsets suffices to cover that compact set, i.e.,

$$B_0(R_0) \setminus \mathcal{N}_\epsilon\left(\bigcap_{B_c(R) \in \mathcal{B}} B_c(R) \right) \subset \bigcup_{B_c(R) \in \mathcal{B}_\mathcal{M}} \Big(B_0(R_0) \setminus B_c(R) \Big). \tag{3.7}$$

Taking complements of this inclusion yields

$$\mathcal{N}_\epsilon\left(\bigcap_{B_c(R) \in \mathcal{B}} B_c(R) \right) \supset \bigcap_{B_c(R) \in \mathcal{B}_\mathcal{M}} B_c(R),$$

from which we conclude that (1.9) follows from its apparently weaker analogue. This finishes the proof of Theorem 1.1. ∎

4. Estimating the full back-scattering kernel. This section is composed of three main propositions, and a few supporting lemmas, which combine to prove Theorem 1.2. We begin with the following lemma concerning the invertibility of the operator $I - Gq$.

LEMMA 4.1. *Let $q \in L^\infty(\mathbb{R}^d)$ and be compactly supported. Then $q(I-Gq)^{-1}$ is a bounded operator from $L^2(\mathbb{R}^d)$ to itself. Moreover, let $q \in L^p(\mathbb{R}^d)$, with $p > \max(2, \frac{d}{2})$, be compactly supported. Then,*

$$q(I-Gq)^{-1} = q + Gq(I-Gq)^{-1}$$

and $Gq(I-Gq)^{-1}$ and $(I-Gq)^{-1}$ are bounded operators from $L^2(\mathbb{R}^d)$ to itself.

Proof. The lemma is a special consequence of Lemma 12 and Corollaries 13, 14, and 15 of [9]. Roughly speaking, multiplication by q loses $\frac{d}{p}$ L^2-derivatives, while G gains two. This implies that Gq is compact and that $I-Gq$ is Fredholm. If zero were an eigenvalue of $I-Gq$, then the corresponding eigenfunction would be a nonzero outgoing solution to (1.2). An application of Green's formula to this solution and its complex conjugate shows that this outgoing solution would have zero far field. Rellich's lemma and unique continuation imply that any outgoing solution with zero far field is identically zero. Thus zero is not an eigenvalue and $I-Gq$ is invertible. ∎

PROPOSITION 4.2. *If $q \in L^\infty$, the $\operatorname{supp} q \subset B_c(R)$, then scattering operator may be factored as,*

$$\mathcal{S} = e^{-i\Theta\cdot c}\mathcal{H}_R^* B \mathcal{H}_R e^{i\Theta\cdot c}, \tag{4.1}$$

where B is a bounded operator from $L^2(\mathbb{R}^d)$ to itself.

Proof. Because $\mathcal{H}$ intertwines translation by c and multiplication by $e^{ik\Theta\cdot c}$, i.e.,

$$\mathcal{H}e^{-i\Theta\cdot c} = T_c\mathcal{H},$$

it is enough to treat the case $c=0$. We begin with (2.9),

$$\mathcal{S} = \mathcal{H}^*(I-qG)^{-1}q\mathcal{H},$$

then insert a characteristic function of the ball

$$= \mathcal{H}^*(I-qG)^{-1}q\chi_R\mathcal{H}, \tag{4.2}$$

then shift to (2.8)

$$= \mathcal{H}^* q(I-Gq)^{-1}\chi_R\mathcal{H}$$

and again insert another characteristic function

$$= \mathcal{H}^*\chi_R q(I-Gq)^{-1}\chi_R\mathcal{H}, \tag{4.3}$$

which we recognize as

$$= \mathcal{H}_R^* q(I-Gq)^{-1}\mathcal{H}_R. \tag{4.4}$$

Finally, we observe that the operator in the middle is bounded according to Lemma 4.1. ∎

Remark 4.3. The only property of the scattering operator that we will use in the sequel is (4.1). The conclusions of Theorem 1.2 apply to any operator which admits such a factorization.

Remark 4.4. So as not to unnecessarily complicate the subsequent discussion, we will continue working with $q \in L^\infty$ in the rest of this section. Theorem 1.2, however,

remains true for compactly supported $q \in L^p(\mathbb{R}^d)$ with $p > \max(2, \frac{d}{2})$. In this case, the scattering operator is the sum of two operators,

$$\begin{aligned} \mathcal{S} &= \mathcal{H}_R^* q(I - Gq)^{-1}\mathcal{H}_R \\ &= \mathcal{H}_R^* q\mathcal{H}_R + \mathcal{H}_R^* Gq(I - Gq)^{-1}\mathcal{H}_R. \end{aligned} \tag{4.5}$$

The first is the Born scattering operator and the second satisfies (4.1) with a $B = Gq(I - Gq)^{-1}$ which is bounded from $L^2(\mathbb{R}^d)$ to itself.

The factorization (4.1) combines with the singular values of the Herglotz operator to give some natural estimates for the terms in what we might call a *block decomposition* of the scattering operator.

PROPOSITION 4.5. *Let $\mathcal{S}$, mapping $L^2(S^{d-1})$ to itself, admit the factorization (4.1); then $\mathcal{S}_{nm}$, defined as*

$$\mathcal{S}_{nm} = \mathcal{P}_n \mathcal{S} \mathcal{P}_m, \tag{4.6}$$

has an L^∞ kernel $s_{nm}(\Theta, \Phi)$ and satisfies

$$\begin{aligned} ||\mathcal{S}_{nm}|| &\le \sigma_n(R)\sigma_m(R)\ ||B||, \\ |s_{nm}(\Theta, \Phi)| &\le \sqrt{N}\sigma_n(R)\sqrt{M}\sigma_m(R)\ ||B||. \end{aligned} \tag{4.7}$$

Proof. We insert the factorization (4.1) into (4.6),

$$\begin{aligned} \mathcal{S}_{nm} &= \mathcal{P}_n \mathcal{H}_R^* B \mathcal{H}_R \mathcal{P}_m \\ &= (\mathcal{H}_R \mathcal{P}_n)^* B \mathcal{H}_R \mathcal{P}_m, \end{aligned}$$

and use our singular value decompositions, (3.3) and (3.5),

$$= \sigma_n \sigma_m \mathcal{Q}_n^* B \mathcal{Q}_m, \tag{4.8}$$

so that

$$||\mathcal{S}_{nm}|| \le \sigma_n \sigma_m ||B||.$$

Recalling again that the p_n^Θ are kernels of the $\mathcal{P}_n$, we see that the kernel of $\mathcal{S}_{nm}$ is given by

$$\begin{aligned} s_{nm}(\Theta, \Phi) &= \left(p_n^\Theta, \mathcal{S} p_m^\Phi\right)_{L^2(S^{d-1})} \\ &= \left(p_n^\Theta, \mathcal{S}_{nm} p_m^\Phi\right)_{L^2(S^{d-1})}, \\ |s_{nm}(\Theta, \Phi)| &\le ||p_n^\Theta||\ ||\mathcal{S}_{nm}||\ ||p_m^\Phi|| \\ &\le \sqrt{N}\sigma_n(R)\sigma_m(R)||B||\sqrt{M}. \end{aligned}$$

∎

The main conclusion of Theorem 1.2 is the estimate (1.10) of the left-hand side of the identity (4.9) below. We will apply (4.7) to show that the series on the right-hand side is summable and then to prove (1.10).

$$\mathcal{P}_l s(\Theta, -\Theta) = \sum_{n,m} \mathcal{P}_l s_{nm}(\Theta, -\Theta). \tag{4.9}$$

The next proposition tells us that many of the terms in the series are zero.

PROPOSITION 4.6.

$$\mathcal{P}_l(s_{nm}(\Theta,-\Theta)\equiv 0$$

unless the sum of any two indices is greater than or equal to the third.

Proof.

$$\begin{aligned}
s_{nm}(\Theta,-\Theta) &= \left(p_n^{-\Theta},\mathcal{S}p_m^{\Theta}\right)_{L^2(S^{d-1})}\\
&= \left(\int_{S^{d-1}} p_n(\Theta\cdot\Psi_1)p_n^{\Psi_1}d\Psi_1,\mathcal{S}\int_{S^{d-1}} p_m(\Theta\cdot\Psi_2)p_m^{\Psi_2}d\Psi_2\right)\\
&= \int_{S^{d-1}}\int_{S^{d-1}} p_n(\Theta\cdot\Psi_1)p_m(\Theta\cdot\Psi_2)\left(p_n^{\Psi_1},\mathcal{S}p_m^{\Psi_2}\right)d\Psi_1 d\Psi_2\\
\mathcal{P}_l s_{nm}(\tau) &= \int_{S^{d-1}}\int_{S^{d-1}}\left[\int_{S^{d-1}} p_l(\Theta\cdot\tau)p_n(\Theta\cdot\Psi_1)p_m(\Theta\cdot\Psi_2)d\Theta\right]\left(p_n^{\Psi_1},\mathcal{S}p_m^{\Psi_2}\right)d\Psi_1 d\Psi_2.
\end{aligned}$$

The quantities in the square brackets are closely related to the Clebsch–Gordon coefficients [14]. To see that they must be zero, fix Ψ_1 and Ψ_2, and call the quantity in the square brackets $c_{nml}(\tau)$. Since $p_l^\tau(\Theta)=p_l(\Theta\cdot\tau)$ is the kernel of the projection operator $\mathcal{P}_l$ onto degree l spherical harmonics, we see that $c_{nml}(\tau)$ are the degree l spherical harmonics in the condensed spherical harmonic expansion of the product of the two spherical harmonics, $p_n^{\Psi_1}$ and $p_m^{\Psi_2}$, i.e.,

$$p_n^{\Psi_1}(\Theta)p_m^{\Psi_2}(\Theta)=\sum_{l=0}^{\infty}c_{nml}(\Theta).$$

Recalling that every spherical harmonic extends to a homogeneous polynomial of the same degree, we see that the left-hand side extends to the unit ball as a homogeneous polynomial of degree $n+m$ if we replace Θ by $r\Theta$. Since the left-hand side goes to zero as r^{n+m} as $r\to 0$, so must the right-hand side. This is possible only if the c_{nml}, which extend as homogeneous degree l polynomials, are zero for all $l<n+m$. Finally, note that all conclusions remain valid if we permute the indices n, m, and l. ∎

Proof of Theorem 1.2. We start by applying Propositions 4.2 and 4.5 to conclude that

$$|s_{nm}(\Theta,\Phi)|\le ||q(I-Gq)^{-1}||\sqrt{N}\sigma_n(R)\sqrt{M}\sigma_m(R). \tag{4.10}$$

We set $\Phi=-\Theta$ and let $s_{nm}(\Theta)=s_{nm}(\Theta,-\Theta)$.

Now, (4.10) tells us that the terms $|s_{nm}(\Theta)|$ are summable. Hence, we may write

$$s(\Theta)=\sum_{n,m=0}^{\infty}s_{nm}(\Theta)$$

and therefore

$$\mathcal{P}_l s(\Theta)=\sum_{n,m=0}^{\infty}\mathcal{P}_l s_{nm}(\Theta),$$

which is the same as

$$= \sum_{n+m\geq l} \mathcal{P}_l s_{nm}(\Theta)$$

according to Proposition 4.6. Hence

$$||\mathcal{P}_l s||_{L^2} \leq \sum_{n+m\geq l} ||\mathcal{P}_l s_{nm}(\Theta)||_{L^2} \tag{4.11}$$

$$\leq \sum_{n+m\geq l} ||s_{nm}(\Theta)||_{L^2}.$$

Recalling that ω is the area of S^{d-1},

$$\leq \sqrt{\omega} \sum_{n+m\geq l} ||s_{nm}(\Theta)||_{L^\infty} \tag{4.12}$$

$$\leq \sqrt{\omega}||q(I-Gq)^{-1}|| \sum_{n+m\geq l} \sqrt{N}\sigma_n(R)\sqrt{M}\sigma_m(R).$$

Finally, Proposition 4.7, which we state and prove below, estimates the sum on the right-hand side in terms of $\sigma_l(2R)$.

$$\leq ||q(I-Gq)^{-1}|| \frac{\omega\sqrt{2}\, R^{\frac{d}{2}}}{(1-\frac{2R}{l})(1-\frac{R^2}{l+\frac{d}{2}})^2}\ \sqrt{L}\ \sigma_l(2R). \tag{4.13}$$

This completes the proof for $q \in L^\infty$. For $q \in L^p(\mathbb{R}^d)$ with $p > \max(2, \frac{d}{2})$, we return to (4.5) and notice that Theorem 1.2 holds for each of the two terms. It holds for the first because it is a Born back-scattering kernel and for the second because the operator in the middle is bounded. It is not hard to see that the conclusion will persist for the sum. □

The next proposition provides an estimate of the right-hand side of (4.13) and thus yields the last ingredient necessary to establish Theorem 1.2.

PROPOSITION 4.7.

$$\sum_{n+m\geq l} \sqrt{N}\ \sigma_n(R)\sqrt{M}\ \sigma_m(R) \leq \frac{\sqrt{2\omega}\ R^{\frac{d}{2}}}{(1-\frac{2R}{l})(1-\frac{R^2}{l+\frac{d}{2}})^2}\ \sqrt{L}\ \sigma_l(2R). \tag{4.14}$$

Proof. The proof requires several small lemmas. The first allows us to estimate the $\sigma_n(r)$ from above and below by ratios of Γ-functions and powers of r.

LEMMA 4.8.

$$\frac{\Gamma(\frac{d}{2})(\frac{r}{2})^n}{\Gamma(n+\frac{d}{2})}\left(1-\frac{(\frac{r}{2})^2}{n+\frac{d}{2}}\right) \leq j_n(r) \leq \frac{\Gamma(\frac{d}{2})(\frac{r}{2})^n}{\Gamma(n+\frac{d}{2})}, \tag{4.15}$$

$$\frac{\sqrt{\omega}\ 2^{\frac{d-1}{2}}\Gamma(\frac{d}{2})(\frac{r}{2})^{n+\frac{d}{2}}}{\Gamma(n+\frac{d+1}{2})}\left(1-\frac{(\frac{r}{2})^2}{n+\frac{d}{2}}\right) \leq \sigma_n(r) \leq \frac{\sqrt{\omega}\ 2^{\frac{d}{2}}\Gamma(\frac{d}{2})(\frac{r}{2})^{n+\frac{d}{2}}}{\Gamma(n+\frac{d+1}{2})}. \tag{4.16}$$

Proof. The first inequality, (4.15), is just the statement that the spherical Bessel function lies between the first and the partial sum of the first two terms of its alternating power series expansion. The second is obtained from the first by squaring, integrating over the ball of radius r, and making use of (4.18). □

The dimensions N of the spaces of spherical harmonics of degree n also appear on the left-hand side of (4.14). We will estimate them from above and below in terms of Γ-functions as well.

LEMMA 4.9.

$$N(n,d) = \begin{cases} 1, & n = 0, \\ d, & n = 1, \\ \frac{(2n+d-2)(n+d-3)!}{n!(d-2)!} & \text{otherwise}, \end{cases}$$

$$\frac{\Gamma(n+d-1)}{\Gamma(n+1)\Gamma((d-1)} \leq N \leq 2\frac{\Gamma(n+d-1)}{\Gamma(n+1)\Gamma(d-1)}. \tag{4.17}$$

Proof. The first formula follows from the observation that N satisfies the difference equation

$$N(n,d) = N(n,d-1) + N(n-1,d)$$

and the fact that (4.9) holds in the special cases $n = 0$ and $d = 2$. See [11] for a different proof. The inequality (4.17) follows from this formula and

$$n + d - 2 \leq 2n + d - 2 \leq 2(n + d - 2). \qquad \square$$

The ratio $\frac{s!}{(s-1)!} = \frac{\Gamma(s+1)}{\Gamma(s)} = s$. We need to estimate the analogous ratio when one of the arguments is a half integer. For this we state the following lemma.

LEMMA 4.10.

$$\left(s - \frac{1}{2}\right)^{\frac{1}{2}} \leq \frac{\Gamma(s+\frac{1}{2})}{\Gamma(s)} \leq s^{\frac{1}{2}}. \tag{4.18}$$

Proof. Because the gamma function is log-convex [3],

$$\Gamma\left(s + \frac{1}{2}\right) \leq \Gamma(s)^{\frac{1}{2}}\Gamma(s+1)^{\frac{1}{2}},$$
$$\frac{\Gamma(s+\frac{1}{2})}{\Gamma(s)} \leq \frac{\Gamma(s+1)^{\frac{1}{2}}}{\Gamma(s)^{\frac{1}{2}}} = s^{\frac{1}{2}}.$$

Analogously,

$$\Gamma(s) \leq \Gamma\left(s - \frac{1}{2}\right)^{\frac{1}{2}} \Gamma\left(s + \frac{1}{2}\right)^{\frac{1}{2}},$$
$$\frac{\Gamma(s)}{\Gamma(s+\frac{1}{2})} \leq \frac{\Gamma(s-\frac{1}{2})^{\frac{1}{2}}}{\Gamma(s+\frac{1}{2})^{\frac{1}{2}}} = \frac{1}{(s-\frac{1}{2})^{\frac{1}{2}}}. \qquad \square$$

We use Lemma 4.10 to establish a replacement for the binomial theorem involving Γ-functions of half integers rather than just factorials.

LEMMA 4.11.

$$\sum_{n+m=l} \frac{1}{\Gamma(n+\frac{d+1}{2})} \frac{1}{\Gamma(m+\frac{d+1}{2})} \leq \frac{2^{l+d-1}}{\Gamma(l+d)}.$$

Proof. If d is odd, the inequality is an equality which follows from the binomial expansion of $(1+1)^{l+d-1}$. If d is even, we use (4.18),

$$\begin{aligned}\sum_{n+m=l} \frac{1}{\Gamma(n+\frac{d+1}{2})}\frac{1}{\Gamma(m+\frac{d+1}{2})} &\le \sum_{n+m=l} \frac{(n+\frac{d}{2})^{\frac{1}{2}}}{\Gamma(n+\frac{d}{2}+1)}\frac{(m+\frac{d}{2})^{\frac{1}{2}}}{\Gamma(m+\frac{d}{2}+1)}\\ &\le \left(l+\frac{d}{2}\right)\sum_{n+m=l}\frac{1}{\Gamma(n+\frac{d}{2}+1)}\frac{1}{\Gamma(m+\frac{d}{2}+1)}\\ &= \left(l+\frac{d}{2}\right)\frac{2^{l+d}}{\Gamma(l+d+1)}\\ &\le \frac{2^{l+d}}{\Gamma(l+d)}. \qquad \square\end{aligned}$$

The sum in (4.14) is in fact a double summation over the indices n and m. Hence, we first estimate each single sum in Lemma 4.12 and then sum those estimates in Lemma 4.13 to estimate the double summation.

LEMMA 4.12.

$$\sum_{n+m=l} \sqrt{N}\,\sigma_n(R)\sqrt{M}\,\sigma_m(R) \le \frac{\sqrt{2\omega}\,R^{\frac{d}{2}}}{(1-\frac{R^2}{l+\frac{d}{2}})}\sqrt{L}\,\sigma_l(2R).$$

Proof.

$$\sum_{n+m=l} \sqrt{N}\,\sigma_n(R)\sqrt{M}\,\sigma_m(R) \le L\sum_{n+m=l}\sigma_n(R)\,\sigma_m(R). \tag{4.19}$$

We apply (4.15) from Lemma 4.8,

$$\le \omega 2^{d-1}\Gamma^2\left(\frac{d}{2}\right)\left(\frac{R}{2}\right)^{l+d} L\sum_{n+m=l}\frac{1}{\Gamma(n+\frac{d+1}{2})}\frac{1}{\Gamma(m+\frac{d+1}{2})},$$

then Lemma 4.11

$$\le \omega 2^{d-1}\Gamma^2\left(\frac{d}{2}\right)\left(\frac{R}{2}\right)^{l+d} L\frac{2^{l+d}}{\Gamma(l+d)},$$

followed by the left inequality in (4.16) of Lemma 4.8,

$$\le \frac{\sqrt{\omega}\sqrt{L}\sigma_l(2R)}{(1-\frac{R^2}{l+\frac{d}{2}})} \quad \left(2^{\frac{d-1}{2}}\left(\frac{R}{2}\right)^{\frac{d}{2}}\right) \quad \left(\Gamma(\frac{d}{2})\sqrt{L}\frac{\Gamma(l+\frac{d+1}{2})}{\Gamma(l+d)}\right), \tag{4.20}$$

and finally (4.17) to see that the third factor in (4.20) is less than 2,

$$\le \frac{\sqrt{\omega}\sqrt{L}\sigma_l(2R)}{(1-\frac{R^2}{l+\frac{d}{2}})} \quad \left(2^{\frac{d+1}{2}}\left(\frac{R}{2}\right)^{\frac{d}{2}}\right). \qquad \square \tag{4.21}$$

Finally, we complete these estimates with the next lemma.

LEMMA 4.13.

$$\sum_{l=l_0}^{\infty}\sqrt{L}\,\sigma_l \le \frac{\sqrt{L_0}\,\sigma_{l_0}}{(1-\frac{R}{l_0})(1-\frac{R^2}{4l_0+2d})}. \tag{4.22}$$

Proof.

$$\sum_{l=l_0}^{\infty} \sqrt{L}\, \sigma_l \leq \sqrt{L_0}\, \sigma_{l_0} \left(1 + \sum_{l=l_0+1}^{\infty} \sqrt{\frac{L}{L_0}}\, \frac{\sigma_l}{\sigma_{l_0}} \right).$$

According to Lemma 4.9,

$$\leq \sqrt{L_0}\sigma_{l_0} \left(1 + \sum_{k=1}^{\infty} \sqrt{\frac{\Gamma(l_0+k+d-1)\Gamma(l_0+1)}{\Gamma(l_0+d-1)\Gamma(l_0+k+1)}}\, \frac{\sigma_l}{\sigma}_{l_0} \right).$$

Estimating the square root gives

$$\leq \sqrt{L_0}\sigma_{l_0} \left(1 + \sum_{k=1}^{\infty} \left(1 + \frac{d-2}{l_0+1} \right)^{\frac{k}{2}}\, \frac{\sigma_l}{\sigma}_{l_0} \right). \tag{4.23}$$

Next, we apply both upper and lower bounds in (4.16) of Lemma 4.8,

$$\leq \sqrt{L_0}\sigma_{l_0} \left(1 + \sum_{k=1}^{\infty} \left(1 + \frac{d-2}{l_0+1} \right)^{\frac{k}{2}}\, \left(\frac{R}{2} \right)^k \frac{\Gamma(l_0+\frac{d+1}{2})}{\Gamma(l_0+k+\frac{d+1}{2})}\, \frac{1}{1-\frac{R^2}{4l_0+2d}} \right), \tag{4.24}$$

and finally estimate the ratio of Gamma functions, and compare to a geometric series,

$$\begin{aligned} &\leq \sqrt{L_0}\sigma_{l_0} \left(1 + \sum_{k=1}^{\infty} \left(1 + \frac{d-2}{l_0+1} \right)^{\frac{k}{2}}\, \frac{\left(\frac{R}{2}\right)^k}{(l_0+\frac{d+1}{2})^k} \frac{1}{1-\frac{R^2}{4l_0+2d}} \right) \\ &\leq \frac{\sqrt{L_0}\sigma_{l_0}}{1-\frac{R^2}{4l_0+2d}} \left(1 + \sum_{k=1}^{\infty} \left(\frac{(1+\frac{d-2}{l_0+1})^{\frac{1}{2}} r}{l_0+\frac{d+1}{2}} \right)^k \right) \\ &\leq \frac{\sqrt{L_0}\sigma_{l_0}}{1-\frac{R^2}{4l_0+2d}} \sum_{k=0}^{\infty} \left(\frac{r}{l_0}\, \frac{\sqrt{1+\frac{d-2}{l_0+1}}}{1+\frac{d+1}{2l_0}} \right)^k \\ &\leq \frac{\sqrt{L_0}\sigma_{l_0}}{1-\frac{R^2}{4l_0+2d}} \sum_{k=0}^{\infty} \left(\frac{r}{l_0} \right)^k. \end{aligned} \tag{4.25}$$

Hence, (4.22) follows on summing the geometric series. ∎

We need only observe that a combination of Lemmas 4.12 and 4.13 implies (4.14) to finish the proof of Proposition 4.7. ∎

5. Numerical applications. In this section, we illustrate a method for finding the convex back-scattering support with a simple example in two dimensions. Here, our scatterer is rectangle R_q,

$$R_q = \{(x,y) \in \mathbb{R}^2 : -1 \leq x \leq 1, -1/2 \leq y \leq 1/2\},$$

having horizontal sides at $x = \pm 1$ and vertical sides at $y = \pm 0.5$. We computed the full (not the Born) back-scattering kernel numerically, using the Helmholtz equation

$$(\Delta + k^2 n^2(x))u = 0$$

with the wavenumber $k = 5$ and index of refraction given by the formula

$$n^2 = \begin{cases} 2 & x \in R_q, \\ 1 & x \notin R_q. \end{cases}$$

5.1. The forward solution. In this section we explain how we computed the back-scattering kernel that we will use as data to numerically determine the convex back-scattering support for our example to follow.

For each incident wave, we employ a two-step process. First, we solved a discretized version of the Lippmann–Schwinger integral equation (LSIE) for the total field u on (a region slightly larger than) the support of the scatterer, using a Nyström, or collocation, method. Next, we use the product qu as a source and numerically integrate it against the two-dimensional version of the far field operator $\mathcal{F}$, computing only the value of the scattered field in the direction opposite that of the incident wave.

More specifically, we solved the LSIE numerically by discretizing the operator $I - Gq$, and the incident wave u^i, over a large, but finite, number of nodes on a prescribed rectangular region $D = D_x \times D_y$ containing the scatterer q and then solved the ensuing large system of simultaneous equations (of the form $Au = u^i$) for the unknown u on our grid of evaluation nodes. This process was then repeated over a collection of incident directions on the unit circle.

In two dimensions, we use a collection of equally spaced points $(x_{\hat{l}}, y_{\hat{p}}) \in D$, $(l = 1, 2, \ldots, M, p = 1, 2, \ldots, N)$, and then express the LSIE as

$$u(\hat{x}_l, \hat{y}_p) + \int_{D_x} \int_{D_y} K(\hat{x}_l - x, \hat{y}_p - y) u(x, y) dy dx = u^i(\hat{x}_l, \hat{y}_p),$$

where the kernel K is

$$K(\hat{x}_l - x, \hat{y}_p - y) = \frac{i}{4} H_0^{(1)} \left(k \sqrt{(\hat{x}_l - x)^2 + (\hat{y}_p - y)^2} \right) q(x, y)$$

and $H_0^{(1)}$ is the usual Hankel function of the first kind. We discretize the points of integration similarly, defining the additional points (x_l, y_p), at the same points of evaluation of the LSIE. Using the trapezoid rule on this same set of nodes allows us to write the fully discrete version of the LSIE as

$$u_{\hat{l},\hat{p}} + \frac{\Delta x \Delta y}{4} \sum_{\substack{p=1 \\ p \neq \hat{p}}}^{N-1} \sum_{\substack{l=1 \\ l \neq \hat{l}}}^{M-1} (K_{\hat{l},l;\hat{p},p} u_{l,p} + K_{\hat{l},l+1;\hat{p},p} u_{l+1,p} + K_{\hat{l},l;\hat{p},p+1} u_{l,p+1} + K_{\hat{l},l+1;\hat{p},p+1} u_{l+1,p+1}) + w_{\hat{l},\hat{p}} S_{\hat{l},\hat{p}} u_{\hat{l},\hat{p}} = u^i_{\hat{l},\hat{p}}.$$

The notation given above has been abbreviated so that the subscripts indicate the points of evaluation, namely, $u_{\hat{l},\hat{p}} = u(\hat{x}_l, \hat{y}_p)$, $u_{l,p} := u(x_l, y_p)$, and $K_{\hat{l},l;\hat{p},p} = K(\hat{x}_l - x_l, \hat{y}_p - y_p)$. Since the kernel of the integral operator is singular along the diagonal, these terms must be treated separately in the discretization scheme. The term $w_{\hat{l},\hat{p}} S_{\hat{l},\hat{p}} u_{\hat{l},\hat{p}}$ above corresponds to the appropriate trapezoid-rule weighted average of the diagonal terms in the discretization. Specifically,

$$S_{\hat{l},\hat{p}} := \begin{cases} \frac{k^2}{4} \delta^2 q_{\hat{l},\hat{p}} (1 - 2 \log \delta - 4C_k) & \text{if } (\hat{x}_l, \hat{y}_p) \in D, \\ \frac{k^2}{8} \delta^2 q_{\hat{l},\hat{p}} (1 - 2 \log \delta - 4C_k) & \text{if } (\hat{x}_l, \hat{y}_p) \in \partial D, \end{cases}$$

where

$$C_k = \frac{1}{2}\left(\log\frac{k}{2} - \gamma\right) - \frac{i\pi}{4}$$

and γ is the Euler–Mascheroni constant.

In more detail, what we have done is to assume that product qu is nearly constant over some small ball of radius δ and integrated the logarithmic singularity of the kernel on this set to define the appropriate matrix entry in the numerical integration along the diagonal. Provided we take a fine mesh of integration points, this presents a viable way to treat this weakly singular behavior. We use an equispaced grid in both x and y and chose δ to be $\Delta x/2$.

Last, we obtained the back-scattered field by simply computing the discrete sum of the form

$$s_q(\theta) = e^{i\frac{5\pi}{4}}\sqrt{\frac{1}{8k\pi}}\sum_{l=1}^{M-1}\left(A_l(\theta) + A_{l+1}(\theta)\right)\frac{\Delta x}{2},$$

where we define the iterated areas $A_l(\theta)$ and $A_{l+1}(\theta)$ as

$$A_l(\theta) = \sum_{p=1}^{N-1}\left(f(x_l, y_p, \theta) + f(x_l, y_{p+1}, \theta)\right)\frac{\Delta y}{2},$$

$$A_{l+1}(\theta) = \sum_{p=1}^{N-1}\left(f(x_{l+1}, y_p, \theta) + f(x_{l+1}, y_{p+1}, \theta)\right)\frac{\Delta y}{2}$$

and where we used the computed values of the total field u on the grid to compute

$$f(x_l, y_p, \theta) = e^{-ik(x_l\cos\theta + y_p\sin\theta)}q(x_l, y_p)u(x_l, y_p).$$

Again, we ran the above numerical scheme for a rectangular scatterer, i.e., $q = k^2\chi_{R_q}$ with χ_{R_q} the characteristic function of the rectangle

$$R_q = \{(x, y) \in \mathbb{R}^2 : -1 \le x \le 1, -1/2 \le y \le 1/2\}$$

and computed the back-scattering kernel at 100 equispaced points on the unit circle. We used a wavenumber of $k = 5$ and distributed 40 nodes along each of the x and y axes within the bounding region

$$D = \{(x, y) \in \mathbb{R}^2 : -1 \le x \le 1, -1 \le y \le 1\}.$$

Just over 1.5 wavelengths fall within D, so that 40 trapezoid-rule integration nodes (roughly 25 nodes per wavelength) should yield an accurate simulated solution.

5.2. Numerical computation of the convex back-scattering support. In this section, we illustrate how we use the back-scattering data to find only the rectangle.[4] The general outline of our method for locating the rectangle from the back-scattering kernel will be the following:

[4]In general, we can expect to find only the convex back-scattering support of the scatterer, which may be smaller than the convex hull of the scatterer. In the Born approximation, it follows from Theorem 14 of [8] that the convex back-scattering support of a rectangle is exactly the rectangle. We have not proved this for the full back-scattering data, but the numerical computations below suggest that this is the case.

1. Choose a center, $c \in \mathbb{R}^2$ (we take $c = 0$ the first time through), and expand the back-scattering kernel in a Fourier series centered at this point.
2. Find the *support*, i.e., the value of n where these coefficients *rapidly transition to zero*, in order to calculate the radius of the smallest ball centered at c that contains the (convex back-scattering support of the) rectangle. We will explain this in detail below.
3. Choose other centers and repeat the above process to produce other balls.
4. The scatterer, e.g., the rectangle, must be contained in the common intersection of all these balls.

We will illustrate this procedure with the sequence of Figures 2–5 and again with the sequence of Figures 6–9. Before proceeding to the figures, however, we need to explain the logic behind the second step in more detail. We begin by expanding the back-scattering kernel, i.e., the measured back-scattered signal, in a Fourier series centered at c. On the circle, it is more convenient to work with the azimuthal angle $\theta \in [0, 2\pi]$, which is related to the unit vector Θ through $\Theta = (\cos\theta, \sin\theta)$. The Fourier series expansion, centered at c, is

$$e^{ic\cdot\Theta} s(\Theta) = \sum_{n=-\infty}^{\infty} t_n e^{in\theta}.$$

Numerically, we compute a finite sequence $\{t_n\}$ as the discrete Fourier transform of the sequence $\{e^{ic\cdot\Theta_n} s(\Theta_n)\}$, where $\{\Theta_n\}$ are equispaced points on the unit circle. In Theorem 1.2 we expanded in condensed spherical harmonics. In two dimensions, the condensed spherical harmonics of degree n are the linear combinations of $e^{in\theta}$ and $e^{-in\theta}$, i.e., $s_0^{(c)}(\Theta) = t_0$ and

$$s_n^{(c)}(\Theta) = t_n e^{in\theta} + t_{-n} e^{-in\theta}, \quad n \geq 1, \tag{5.1}$$

so that $\|s_0^{(c)}\|_{L^2(S^1)} = |t_0|$ and

$$||s_n^{(c)}||_{L^2(S^1)} = \left(|t_{-n}|^2 + |t_n|^2\right)^{\frac{1}{2}}, \quad n \geq 1. \tag{5.2}$$

Now, if the scatterer, i.e., our rectangle, is contained in the ball of radius R, then the s_n's satisfy the estimate (1.10), which simplifies slightly in two dimensions because $N(n, 2) = 2$. Specifically,

$$||s_n^{(c)}||_{L^2(S^1)} \leq C(q) N \sigma_n(2kR) \leq 2C(q)\sigma_n(2kR).$$

Theorem 1.2 only tells us that if the rectangle is contained in the ball of radius R centered at c, then the s_n's (a shorthand notation for the $||s_n^{(c)}||_{L^2}$) are bounded by a constant times the $\sigma_n(2kR)$'s. We shall operate as if we knew the converse were true as well.[5] The wavenumber k is fixed ($k = 5$ in the example below), so we want to examine the s_n's and find the smallest value of R for which such a bound holds. Our test principle for finding R from the Fourier coefficients will be the following.

TEST PRINCIPLE 5.1. *The s_n's are effectively supported in the interval $(0, N)$ if and only if the the convex back-scattering support is contained in the ball of radius $R = |\frac{N}{2k}|$.*

We expect the s_n's to be effectively supported in the interval $(0, N)$ because the $\sigma_n(2kR)$'s have exactly this property. As a function of n, $\sigma_n(2kR)$ is positive and

[5]For the Born approximation, the converse is part of Theorem 1.1.

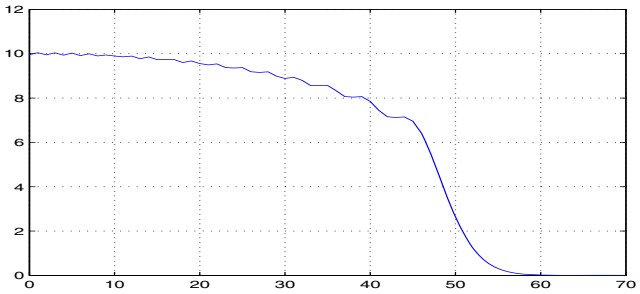


FIG. 1. *$\sigma_n(50)$ plotted as a function of n. Notice that $\sigma_n(50)$ is large (> 6) for $n < 45$ and small (< 0.5) for $n > 55$. If we knew only that this was a plot of $\sigma_n(R)$ for some R, we could have deduced an approximation for R by finding the value of n where the function rapidly transitioned to zero.*

bounded away from zero for $n < |2kR|$ and effectively zero for $n > |2kR|$, with a transition region of width proportinal to $(2kR)^{1/3}$. We see evidence of this behavior in the graph of $\sigma_n(50)$ in Figure 1 and in the asymptotic formulas

$$\sigma_n(2kR) \sim \begin{cases} \sqrt{2}\ \left((2kR)^2 - n^2\right)^{\frac{1}{4}}, & n < |2kR|, \\ \left(\frac{ekR}{n}\right)^{n+1} \sim 0, & n > |2kR|. \end{cases} \tag{5.3}$$

Equation (5.3) follows from classical asymptotics of Bessel functions when at least one of either n or kR is large. We don't yet know a proof when neither is large, but we rely on numerical computations in this case. It can be shown that the two sequences $\sigma_{2n}(2kR)$ and $\sigma_{2n+1}(2kR)$ are monotone decreasing as $n > 0$ increases.

The two sequences of graphs, Figures 2–5 and Figures 6–9, demonstrate that the transitions we witness in the σ_n's are also observed in the back-scattering data as well.

As illustrated in Figures 2–5, our test principle is based on a transition which occurs at a finite value of n, while Theorems 1.1 and 1.2 depend only on large n asymptotics. The strict conditions of these theorems may never be verified experimentally, while the test principle, or any condition that does not include a limit as $n \to \infty$, cannot be a mathematically correct theorem. One can always construct a potential q, supported in the ball of radius R, having any finite number of Fourier coefficients of $\widehat{q}(k\Theta)$ equal to an arbitrary set of numbers. This set of numbers can

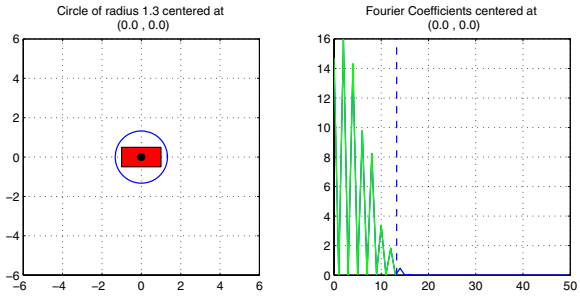


FIG. 2. *We plot the modulus of the Fourier coefficients of the back-scattering kernel on the right and locate the transition to zero at $n = 13$, indicating that value of n by the dashed vertical line. We draw the deduced circle of radius $\frac{n}{2k} = \frac{13}{2\times 5}$ on the left, including the rectangle for comparison.*

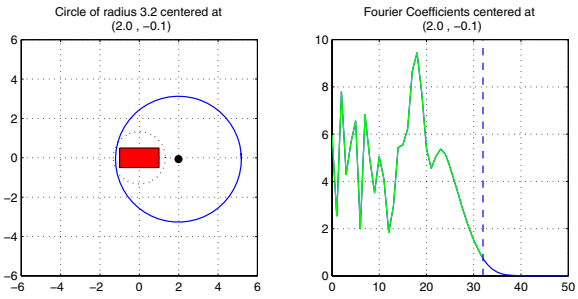


FIG. 3. *We choose a new center, indicated by the black dot in the plot on the left. We compute the modulus of the translated Fourier coefficients, plot them on the right, and locate the transition to zero at* $n = 32$. *We draw the deduced circle of radius* $\frac{n}{2k} = \frac{32}{2\times 5}$ *in the plot on the left. The dashed line represents the circle from Figure* 2.

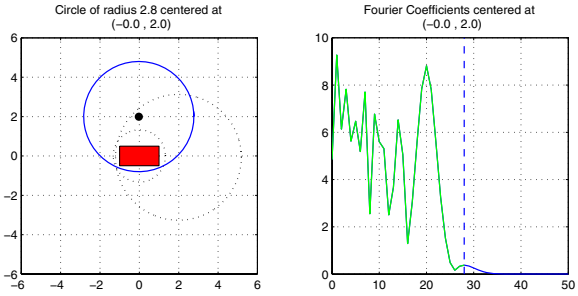


FIG. 4. *We choose another new center, indicated by the black dot. We again plot the modulus of the translated coefficients, locate the transition to zero at* $n = 28$, *and draw the new circle. The dashed lines represents the previous circles from Figures* 2 *and* 3. *Recall that the back-scattering support must lie in their intersection.*

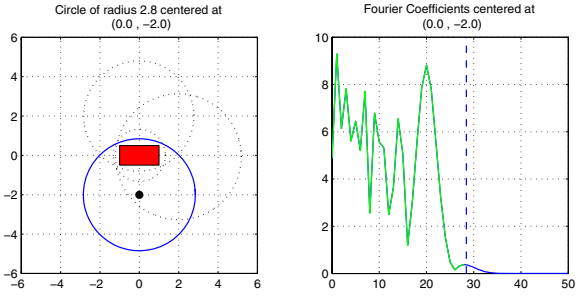


FIG. 5. *We chose the sequence of centers so that we could approximate the rectangle with just a few circles and thus illustrate the method with only a few plots. Of course, we based this choice on a priori knowledge. We don't discuss strategies for efficiently choosing the centers.*

be—somewhat artificially—chosen to be identically zero or to mimic the transition we seek and thus foil our test principle. Nonetheless, such an example would always be exposed by increasing the wave number. Theoretically, we can use a wave of any fixed wavelength to probe any medium—even one which varies very rapidly as a function of position on the scale of that wavelength—to discover its convex back-scattering support. Practically speaking, however, we cannot expect to effectively probe a medium

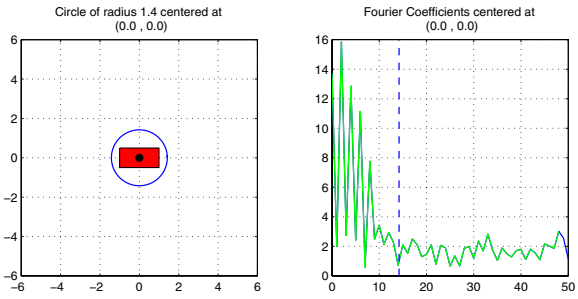


FIG. 6. *The sequence of plots in Figures* 6–9 *uses the same data as the previous sequence, but we have added white noise to both the the amplitude and the phase of the data. We chose the noise level to be* 15% *(i.e., variance equal to* 0.15 *times maximum amplitude* (16) *of the original data for the amplitudes and* $0.15 \times 2\pi$ *for the phases).*

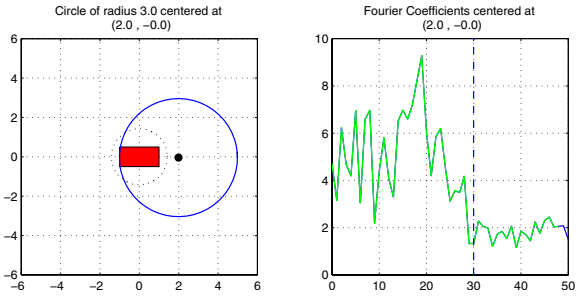


FIG. 7. *We locate the transition to the noise level rather than the transition to zero. If we didn't know that the noise level was* 2.4 *(*0.15×16*), we could readily estimate it from any of the four plots.*

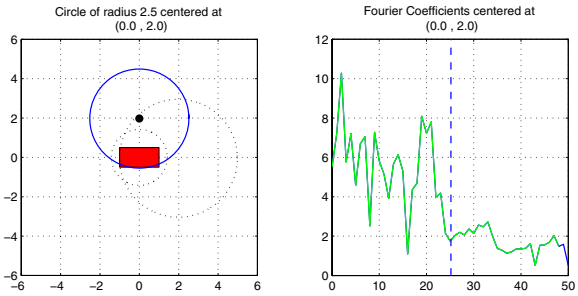


FIG. 8. *Our estimates of the radii in the noisy case are consistently smaller than in the noiseless case.*

whose features vary rapidly on the scale of a wavelength. Some dichotomy between theory and practice is inevitable. One goal for future work will be a more accurate description of those media that we can confidently test with this data and this method.

We should mention that there are certainly other reasonable, and perhaps better, methods for deducing R than our proposed Test Principle 5.1. For instance, one might attempt to sum a regularized version of the series given in (1.8) over various values of R and seek the value of R where the sum becomes bigger than some prescribed threshold. We use Test Principle 5.1 since the very steep transition remained easily visible in the

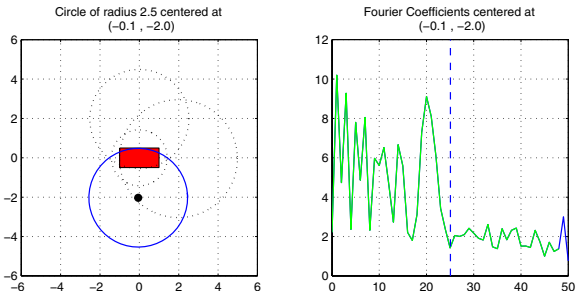


FIG. 9.

presence of noise. This is not a property shared by the partial summations we tried. For our proposed test scheme, we did not need to choose a regularization parameter; however, we did need to decide where the transition to zero, or to the observable noise level, occurred. We found the transition by eye.

REFERENCES

[1] S. AGMON, *Spectral properties of Schrödinger operators and scattering theory*, Ann. Scu. Norm. Super. Pisa, 2 (1975), pp. 151–218.
[2] D. COLTON AND R. KRESS, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, Berlin, 1998.
[3] L. HÖRMANDER, *Notions of Convexity*, Progr. Mathematics 127, Birkhäuser, Boston, 1994.
[4] M. IKEHATA, *On reconstruction in the inverse conductivity problem with one measurement*, Inverse Problems, 16 (2000), pp. 785–793.
[5] M. IKEHATA, *The probe method and its applications*, in Inverse Problems and Related Topics, Res. Notes Math. 419, Chapman & Hall/CRC, Boca Raton, FL, 2000, pp. 57–68.
[6] M. IKEHATA, *A regularized extraction formula in the enclosure method*, Inverse Problems, 18 (2002), pp. 435–440.
[7] A. KIRSCH, *Factorization of the far-field operator for the inhomogeneous medium case and an application in inverse scattering theory*, Inverse Problems, 15 (1999), pp. 413–429.
[8] S. KUSIAK AND J. SYLVESTER, *The scattering support*, Comm. Pure Appl. Math., 56 (2003), pp. 1525–1548.
[9] S. KUSIAK AND J. SYLVESTER, *The scattering support in a background medium*, SIAM J. Math. Anal., 36 (2005), pp. 1142–1158.
[10] D. R. LUKE AND R. POTTHAST, *The no response test—a sampling method for inverse scattering problems*, SIAM J. Appl. Math., 63 (2003), pp. 1292–1312.
[11] C. MÜLLER, *Analysis of Spherical Symmetries in Euclidean Spaces*, Springer-Verlag, Berlin, 1998.
[12] R. POTTHAST, J. SYLVESTER, AND S. KUSIAK, *A 'range test' for determining scatterers with unknown physical properties*, Inverse Problems, 19 (2003), pp. 533–547.
[13] R. POTTHAST, *A set-handling approach for the no-response test and related methods*, Math. Comput. Simulation, 66 (2004), pp. 281–295.
[14] E. P. WIGNER, *Group Theory: And Its Application to the Quantum Mechanics of Atomic Spectra*, Pure and Applied Physics 5, Academic Press, New York, 1959.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 616–626

NONEXISTENCE OF SYNCHRONOUS ORBITS AND CLASS COEXISTENCE IN MATRIX POPULATION MODELS*

RYUSUKE KON†

Abstract. Existence of synchronous orbits in a general class of matrix population models is considered. Our results show that a matrix population model does not possess a synchronous orbit if the associated directed graph is primitive. Furthermore, it is also shown that if there are no synchronous orbits, then all classes coexist. To illustrate these results, the density dependent Leslie matrix model is analyzed.

Key words. synchronous phenomena, periodic insects, permanence, Leslie matrix models, discrete dynamical systems

AMS subject classifications. 39A11, 92D25

DOI. 10.1137/05062353X

1. Introduction.

In this paper, we consider the dynamics of structured populations that are modeled by the following difference equation:

$$\mathbf{x}(t+1) = A_{\mathbf{x}(t)}\mathbf{x}(t), \quad t \in \mathbb{Z}_+, \tag{1}$$

where $\mathbb{Z}_+ = \{0, 1, 2, \dots, \}$, $\mathbf{x}(t) = (x_1(t), x_2(t), \dots, x_n(t))^\top$, and $A_{\mathbf{x}} = (a_{ij}(\mathbf{x}))$ is an $n \times n$ matrix function of $\mathbf{x}$. This equation is a general framework for matrix population models in which a population is divided into n classes (e.g., by chronological age, developmental stage, or habitat position) and the density (or number) of individuals in the ith class is denoted by x_i. Therefore, our interest concentrates on solutions in the nonnegative cone $\mathbb{R}^n_+ := \{\mathbf{x} \in \mathbb{R}^n : x_1 \geq 0, x_2 \geq 0, \dots, x_n \geq 0\}$.

The following equation is a specific example of (1):

$$\begin{pmatrix} x_1(t+1) \\ x_2(t+1) \\ x_3(t+1) \\ \vdots \\ x_n(t+1) \end{pmatrix} = \begin{pmatrix} f_1(\mathbf{x}(t)) & f_2(\mathbf{x}(t)) & \cdots & f_{n-1}(\mathbf{x}(t)) & f_n(\mathbf{x}(t)) \\ p_1(\mathbf{x}(t)) & 0 & \cdots & 0 & 0 \\ 0 & p_2(\mathbf{x}(t)) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & p_{n-1}(\mathbf{x}(t)) & 0 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ \vdots \\ x_n(t) \end{pmatrix}.$$

This equation is the (density dependent) Leslie matrix model for the dynamics of an age-structured population. The variables x_i, $i = 1, 2, \dots, n$, denote the densities (or numbers) of individuals of age i. The functions $f_i(\mathbf{x})$, $i = 1, 2, \dots, n$, denote the numbers of offspring produced by one individual of age i, and $p_i(\mathbf{x})$, $i = 1, 2, \dots, n-1$, denote the probabilities of surviving the ith age-class in one unit of time. This model assumes that the length of life cycle is fixed at n. In addition to the Leslie matrix model, we can find many examples of matrix population models in the literature

*Received by the editors January 30, 2005; accepted for publication (in revised form) August 25, 2005; published electronically December 30, 2005. This research was partially supported by the 21st Century COE Program "Development of Dynamic Mathematics with High Functionality (Kyushu University)" and the Grant-in-Aid for Young Scientists (B), 17740060, 2005, from The Ministry of Education, Culture, Sports, Science and Technology, Japan.
http://www.siam.org/journals/siap/66-2/62353.html
†Faculty of Mathematics, Kyushu University, Hakozaki 6-10-1, Higashi-ku, Fukuoka 812-8581 (kon-r@math.kyushu-u.ac.jp).

[5, 7]. For example, we can find matrix population models incorporating stage or spatial structure (e.g., see [20, 22]).

One of the interesting topics in the study of matrix population models is synchronization. A typical example of synchronous phenomena is found in a density dependent Leslie matrix model with a single reproductive age-class (e.g., see [1, 4, 8, 9, 10, 12, 19, 24]). More precisely, in the Leslie matrix model with $f_1 = \cdots = f_{n-1} = 0$, we can find an orbit $\{\mathbf{x}(t)\}_{t\in\mathbb{Z}_+}$ such that each $\mathbf{x}(t)$ consists of a single nonzero entry whose position moves to the right in a unit of time (e.g., $\mathbf{x}(0) = (+, 0, 0, \dots, 0)^\top, \mathbf{x}(1) = (0, +, 0, \dots, 0)^\top, \dots$). This kind of behavior is called single year class (SYC) dynamics (e.g., see Davydova, Diekmann, and van Gils [12]) since all but one year class are missing. Recently, the concept of SYC dynamics was extended, and the term "multiple year class (MYC) dynamics" was introduced by Mjølhus, Wikan, and Solberg [19]. Synchronous phenomena are also observed in natural insect populations (e.g., see [14, 17, 18, 21]). Periodical cicadas, inhabiting the eastern United States, are typical examples. Their nymphs remain underground for precisely 17 years (or, in the south, 13 years) before emerging from the ground synchronously and in tremendous numbers. Mature nymphs become adults, mate, lay their eggs, and die within the few weeks (see [17]). Therefore, the lengths of their life cycles are fixed (17 or 13 years), and all individuals in each population have the same age; i.e., all but one year class are missing (this phenomenon corresponds to SYC dynamics). In order to explain this synchronization, the Leslie matrix model with $f_1 = \cdots = f_{n-1} = 0$ has been studied.

Although it is known that the structure of the Leslie matrix with a single reproductive age-class can give rise to synchronous phenomena, it is unknown whether this is the only structure leading to synchronous phenomena. For example, some insect undoubtedly has two or more reproductive age-classes, but it is not clear whether the additional reproductive age-classes dissipate synchronous phenomena. In order to clarify this relationship between the structure of the life cycle and the phenomenon of synchronization, we will investigate a structure that eliminates synchronous phenomena from a general class of matrix population models.

Synchronous phenomena are characterized by an orbit on the boundary of the nonnegative cone $\mathrm{bd}\mathbb{R}^n_+ := \{\mathbf{x} \in \mathbb{R}^n_+ : x_1 x_2 \cdots x_n = 0\}$. Hence, following Cushing [8], we define a synchronous orbit as follows.

DEFINITION 1.1 (synchronous orbits). *An orbit* $\{\mathbf{x}(t)\}_{t\in\mathbb{Z}_+}$ *of system* (1) *is said to be synchronous if* $\mathbf{x}(t) \in \mathrm{bd}\mathbb{R}^n_+$ *for all* $t \geq 0$. *A synchronous orbit is said to be nontrivial if* $\mathbf{x}(0) \neq 0$.

Notice that a synchronous orbit does not have to be periodic. It is clear that an SYC dynamics pattern does not appear as long as there are no nontrivial synchronous orbits. Moreover, we see that a nontrivial synchronous orbit always includes some missing classes.

In this paper, we will show that, under certain assumptions, the primitivity of the matrix $A_{\mathbf{x}}$ determines the existence of nontrivial synchronous orbits. It is worth mentioning that Cull and Vogt [6] have addressed the primitivity of a density independent Leslie matrix model to study its periodic behavior of age distributions, i.e., the periodicity of $\mathbf{x}(t)/\sum_{i=1}^n x_i(t)$. As in the study by Cull and Vogt [6], the theory of nonnegative matrices is very useful in our study, although we are concerned not with the periodicity of age distributions but with the existence of synchronous orbits. Since our system involves nonlinear terms, unlike the system of Cull and Vogt [6], we will obtain a result on class coexistence with bounded population densities due to the nonlinearity. That is, we will show that, under certain assumptions, nonexistence

of nontrivial synchronous orbits ensures coexistence of all classes in the sense of *c-permanence*, which is defined as follows. (The definition of *p-permanence* is introduced below to distinguish population survival from class coexistence.)

DEFINITION 1.2 (c-permanence). *System* (1) *is said to be c-permanent if there exist positive constants* $\delta > 0$ *and* $D > 0$ *such that*

$$\delta \le \liminf_{t\to\infty} x_i(t) \le \limsup_{t\to\infty} x_i(t) \le D, \quad i = 1, 2, \dots, n,$$

for all solutions $\{\mathbf{x}(t)\}_{t\in\mathbb{Z}_+}$ *with* $\mathbf{x}(0) \in \mathbb{R}^n_+ \backslash \{(0, 0, \dots, 0)\}$.

The remainder of this paper is organized as follows. In section 2, we introduce some notation and assumptions. That section also includes a new result on the boundedness of solutions, which will be used to prove permanence of a specific matrix population model in section 5. In section 3, we review a known result on permanence for population survival (i.e., p-permanence), which is used to consider c-permanence in the subsequent sections. In section 4, we consider existence of nontrivial synchronous orbits and class coexistence. That section includes the main results of this paper. In section 5, we apply our results to the density dependent Leslie matrix model, which is introduced above, to illustrate our main results. The final section discusses future problems.

2. Preliminaries. In this section, we introduce some notation, assumptions, and preliminary results.

For vectors $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)^\top$, we write $\mathbf{x} \ge \mathbf{y}$ if $x_i \ge y_i$ for all i, and $\mathbf{x} > \mathbf{y}$ if $\mathbf{x} \ge \mathbf{y}$ and $\mathbf{x} \ne \mathbf{y}$. A vector $\mathbf{x}$ is called *nonnegative* if $\mathbf{x} \ge 0$, where 0 denotes the zero vector. A matrix $A = (a_{ij})$ is called *nonnegative* if $a_{ij} \ge 0$ for all i, j. Some important properties of nonnegative matrices are listed in the appendix, which also includes the definitions of *irreducibility* and *primitivity* of matrices and their characteristics. These properties of nonnegative matrices are extensively used in this paper. For matrices $A = (a_{ij})$ and $B = (b_{ij})$, we write $\operatorname{sign}(A) = \operatorname{sign}(B)$ if a_{ij} and b_{ij} have the same sign $-, 0$, or $+$; i.e., the sign pattern of A is identical with that of B. We also write $\operatorname{sign}(\mathbf{x}) = \operatorname{sign}(\mathbf{y})$ for vectors $\mathbf{x}$ and $\mathbf{y}$ if they have the same sign pattern. The set consisting of only the origin is denoted by O.

Throughout this paper, we always assume that system (1) satisfies the following conditions (H1)–(H4):

(H1) each $a_{ij}(\mathbf{x})$ is continuous,
(H2) $A_{\mathbf{x}}\mathbf{x} \ge 0$ for all $\mathbf{x} \ge 0$,
(H3) $A_{\mathbf{x}}\mathbf{x} > 0$ for all $\mathbf{x} > 0$,
(H4) system (1) is *dissipative*; i.e., there exists a positive constant $D > 0$ such that $\limsup_{t\to\infty} \sum_{i=1}^n x_i(t) \le D$ for all solutions $\{\mathbf{x}(t)\}_{t\in\mathbb{Z}_+}$ with $\mathbf{x}(0) \ge 0$.

Assumption (H1) ensures that the map $f(\mathbf{x}) := A_{\mathbf{x}}\mathbf{x}$, which is the right-hand side of (1), is continuous. Assumption (H2) implies that all solutions of (1) with $\mathbf{x}(0) \ge 0$ are always nonnegative. Hence, the nonnegative cone $\mathbb{R}^n_+$ is forward invariant; i.e., $f(\mathbb{R}^n_+) \subset \mathbb{R}^n_+$. Notice that (H2) holds if $A_{\mathbf{x}}$ is nonnegative for all $\mathbf{x} \ge 0$. Assumption (H3) implies that no points $\mathbf{x} > 0$ are mapped to the origin. Therefore, assumption (H3) ensures that $\mathbb{R}^n_+ \backslash O$ is forward invariant; i.e., $f(\mathbb{R}^n_+ \backslash O) \subset \mathbb{R}^n_+ \backslash O$. We can show that (H3) holds if $A_{\mathbf{x}}$ is nonnegative and irreducible for all $\mathbf{x} \ge 0$ as follows. Since $A_{\mathbf{x}}$ is nonnegative for all $\mathbf{x} > 0$, $A_{\mathbf{x}}\mathbf{x} \ge 0$ holds for all $\mathbf{x} > 0$. Suppose that $A_{\mathbf{y}}\mathbf{y} = 0$ for some $\mathbf{y} > 0$ with $y_k > 0$. The irreducibility of $A_{\mathbf{y}}$ ensures that $a_{ik}(\mathbf{y}) > 0$ for some i. Otherwise, there are no paths from the vertices P_k to the other vertices in

the directed graph of $A_{\mathbf{y}}$. This implies that $A_{\mathbf{y}}$ is not strongly connected and hence not irreducible (see Definition A.3 and Theorem A.4 of the appendix). Therefore, $A_{\mathbf{x}}\mathbf{x} > 0$ holds for all $\mathbf{x} > 0$. Assumption (H4) implies that the total population density does not explode. We can find many matrix population models that satisfy assumptions (H1)–(H4) (e.g., see [5, 7]).

In comparison with (H1)–(H3), it is not always easy to check whether system (1) satisfies (H4). In the rest of this section, we obtain a sufficient condition for the dissipativity of system (1). To obtain the sufficient condition in Theorem 2.2, we need the following lemma on dynamical systems.

LEMMA 2.1 (Hutson [15, Lemma 2.1]). *Let (X, d) be a metric space, and let $f : X \to X$ be a continuous function. Let $\gamma^+(\mathbf{x}) = \{\mathbf{x}, f(\mathbf{x}), f^2(\mathbf{x}), \dots\}$ be a semi-orbit of the discrete dynamical system $f : X \to X$. Let $Y \subset X$ be open, and let N be open with a compact closure $\overline{N} \subset Y$. Assume that Y is forward invariant and that $\gamma^+(\mathbf{x}) \cap N \neq \emptyset$ for every $\mathbf{x} \in Y$. Then $M = \gamma^+(\overline{N})$ is a compact absorbing set for Y; i.e., M is a forward invariant compact subset of Y and $\gamma^+(\mathbf{x}) \cap M \neq \emptyset$ for every $\mathbf{x} \in Y$.*

By using this lemma, under assumptions (H1)–(H3), we can obtain the following theorem of dissipativity.

THEOREM 2.2. *Assume that* (H1)–(H3) *hold. Suppose that there exist positive constants $K > 0$ and $\lambda_\infty > 0$ such that the inequalities $\sum_{i=1}^n a_{ij}(\mathbf{x}) \le \lambda_\infty$, $j = 1, 2, \dots, n$, hold for all $\mathbf{x} \in \mathbb{R}^n_+$ with $\sum_{i=1}^n x_i \ge K$. Then system* (1) *is dissipative if $\lambda_\infty < 1$.*

Proof. Let $\{\mathbf{x}(t)\}_{t\in\mathbb{Z}_+}$ be a solution of (1) with $\mathbf{x}(0) \in \mathbb{R}^n_+$. Suppose that $\sum_{i=1}^n x_i(t) \ge K$ for all $t \ge 0$. Then, from (1), we have

$$\begin{aligned}\sum_{i=1}^n x_i(t) &= \sum_{i=1}^n \sum_{j=1}^n a_{ij}(\mathbf{x}(t-1))x_j(t-1)\\ &\le \lambda_\infty \sum_{i=1}^n x_i(t-1)\\ &\vdots\\ &\le \lambda_\infty^t \sum_{i=1}^n x_i(0).\end{aligned}$$

Since $\lambda_\infty < 1$, we have $\mathbf{x}(t) \to 0$ as $t \to \infty$. This is a contradiction. Hence, for every $\mathbf{x}(0) \in \mathbb{R}^n_+$ there exists a $T \ge 0$ such that $\sum_{i=1}^n x_i(T) < K$.

Let $X = Y = \mathbb{R}^n_+$ and $N = \{\mathbf{x} \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i < K\}$. Then it is clear that Y is a forward invariant open subset of X, and N is an open set with a compact closure $\overline{N} \subset Y$. By the above argument, we see that $\gamma^+(\mathbf{x}) \cap N \neq \emptyset$ for every $\mathbf{x} \in Y$. Therefore, Lemma 2.1 implies that $\gamma^+(\overline{N})$ is a compact absorbing set for Y, that is, every solution eventually enters the compact set $\gamma^+(\overline{N})$ and remains there. This implies that system (1) is dissipative. □

Remark. It is straightforward to see that this theorem improves a result by Cushing [7] (cf. Theorem 1.2.2 of [7]). In Theorem 1.2.1 of [7], we can find a sufficient condition that ensures global extinction, i.e., $\lim_{t\to\infty} \mathbf{x}(t) = 0$ for all $\mathbf{x}(0) \in \mathbb{R}^n_+$. In this case, the system is certainly dissipative.

3. P-permanence. In this section, we introduce a known result on the *p-permanence* of system (1), which is defined as follows.

DEFINITION 3.1 (p-permanence). *System* (1) *is said to be p-permanent if there exist positive constants* $\delta > 0$ *and* $D > 0$ *such that*

$$\delta \le \liminf_{t\to\infty} \sum_{i=1}^{n} x_i(t) \le \limsup_{t\to\infty} \sum_{i=1}^{n} x_i(t) \le D$$

for all solutions $\{\mathbf{x}(t)\}_{t\in\mathbb{Z}_+}$ *with* $\mathbf{x}(0) \in \mathbb{R}^n_+\backslash O$.

A result on p-permanence shall be used to consider class coexistence, i.e., c-permanence, in sections 4 and 5. We see that if system (1) is p-permanent, then the total population density $\sum_{i=1}^{n} x_i(t)$ is eventually bounded within some positive interval. Therefore, p-permanence is a mathematical term corresponding to population survival.

The recent study by Kon, Saito, and Takeuchi [16] provides a sufficient condition for the p-permanence of system (1) as follows.

THEOREM 3.2 (see [16]). *Assume that* (H1)–(H4) *hold. Suppose that the matrix* $A_{\mathbf{x}}$ *at the origin, which is denoted by* A_0*, is irreducible. Then system* (1) *is p-permanent if the dominant eigenvalue* λ_0 *of* A_0 *satisfies* $\lambda_0 > 1$.

Remark. Since A_0 corresponds to the Jacobian matrix of (1) evaluated at the origin, $\lambda_0 > 1$ implies that the origin is unstable. Moreover, $\lambda_0 < 1$ implies that the origin is stable, i.e., that system (1) is not p-permanent. Therefore, the magnitude of λ_0 determines whether or not system (1) is p-permanent except in the critical case $\lambda_0 = 1$.

4. Synchronous orbits and class coexistence. In this main section, we consider the existence of synchronous orbits and the possibility of class coexistence, i.e., c-permanence.

The following theorem provides a necessary and sufficient condition for the existence of a nontrivial synchronous orbit.

THEOREM 4.1. *Assume that* (H1)–(H4) *hold. Suppose that* A_0 *is irreducible and* $\mathrm{sign}(A_{\mathbf{x}}) = \mathrm{sign}(A_0)$ *holds for all* $\mathbf{x} \in \mathrm{bd}\mathbb{R}^n_+$. *Then system* (1) *has a nontrivial synchronous orbit if and only if* A_0 *is imprimitive.*

Proof. Suppose that A_0 is imprimitive with index of imprimitivity $h > 1$. Then, by Theorem A.7 of the appendix, A_0^h can be rearranged into quasi-diagonal form by renumbering the indices of rows and columns. So, without loss of generality, we can assume

$$A_0^h = \mathrm{diag}\{B_1, B_2, \ldots, B_h\},$$

where $B_1, B_2, \ldots, B_h$ are primitive matrices. Hence, we can choose a $\mathbf{z} \in \mathrm{bd}\mathbb{R}^n_+\backslash O$ such that $A_0^{kh}\mathbf{z} \in \mathrm{bd}\mathbb{R}^n_+\backslash O$ for all $k \in \mathbb{Z}_+$ (e.g., if B_1 is an $n_1 \times n_1$ matrix, then choose $\mathbf{z} = (z_1, z_2, \ldots, z_n)^\top$ with $z_i > 0$ for $i = 1, \ldots, n_1$ and $z_i = 0$ for $i = n_1 + 1, \ldots, n$). Since A_0 is irreducible and nonnegative, once $A_0^T\mathbf{z} \in \mathrm{int}\mathbb{R}^n_+ := \mathbb{R}^n_+\backslash\mathrm{bd}\mathbb{R}^n_+$ holds for some $T \ge 0$, $A_0^t\mathbf{z} \in \mathrm{int}\mathbb{R}^n_+$ holds for all $t \ge T$. Otherwise, A_0 has a row with only zero entries, so that A_0 is reducible. Therefore, for the $\mathbf{z} \in \mathrm{bd}\mathbb{R}^n_+\backslash O$ chosen above, $A_0^t\mathbf{z} \in \mathrm{bd}\mathbb{R}^n_+\backslash O$ holds for all $t \ge 0$.

It is clear that if $\mathrm{sign}(A) = \mathrm{sign}(B)$ and $\mathrm{sign}(\mathbf{x}) = \mathrm{sign}(\mathbf{y})$ hold for some nonnegative matrices A, B and some nonnegative vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_+$, then $\mathrm{sign}(A\mathbf{x}) = \mathrm{sign}(B\mathbf{y})$ holds. Therefore, if we let $\{\mathbf{x}(t)\}_{t\in\mathbb{Z}_+}$ be a solution of system (1) with $\mathbf{x}(0) = \mathbf{z}$, then $\mathrm{sign}(A_{\mathbf{x}(0)}\mathbf{x}(0)) = \mathrm{sign}(A_0\mathbf{z})$ holds, and inductively $\mathrm{sign}(A_{\mathbf{x}(t)}) = \mathrm{sign}(A_0)$ and $\mathrm{sign}(A_{\mathbf{x}(t-1)}\mathbf{x}(t-1)) = \mathrm{sign}(A_0^t\mathbf{z})$ hold for all $t \ge 0$. This implies that $\mathbf{x}(t) \in \mathrm{bd}\mathbb{R}^n_+\backslash O$ for all $t \in \mathbb{Z}_+$, and then $\{\mathbf{x}(t)\}_{t\in\mathbb{Z}_+}$ is a nontrivial synchronous orbit.

Suppose that A_0 is primitive. Then, by Theorem A.6 of the appendix, there exists an integer $k \geq 1$ such that $A_0^k > 0$. Suppose that there exists a solution $\{\mathbf{x}(t)\}_{t\in\mathbb{Z}_+}$ such that $\mathbf{x}(t) \in \mathrm{bd}\mathbb{R}^n_+ \backslash O$ for all $t \geq 0$. Then we have $A_{\mathbf{x}(k-1)}A_{\mathbf{x}(k-2)}\cdots A_{\mathbf{x}(0)} > 0$. This is a contradiction. Therefore, there are no nontrivial synchronous orbits. □

This theorem ensures that if A_0 is primitive, then there are no orbits remaining in $\mathrm{bd}\mathbb{R}^n_+$; that is, every orbit starting in $\mathrm{bd}\mathbb{R}^n_+$ leaves there and enters the interior of $\mathbb{R}^n_+$ after finite iterations. In the rest of this section, we consider whether or not an interior orbit approaches $\mathrm{bd}\mathbb{R}^n_+$ and show that primitivity implies c-permanence. The following lemma is used below to consider such a problem.

LEMMA 4.2. *Let (X, d) be a compact metric space, and let f and $\gamma^+(\mathbf{x})$ be the same as in Lemma* 2.1. *Let Y be a compact subset of X. Suppose that $\gamma^+(\mathbf{x}) \cap (X\backslash Y) \neq \emptyset$ for every $\mathbf{x} \in X$ and that $X\backslash Y$ is forward invariant. Then there exists a compact absorbing set M for X with $d(M, Y) > 0$.*

Proof. Define $U_t = \{\mathbf{x} \in X : f^t(\mathbf{x}) \in X\backslash Y\}$. Let $\mathbf{x} \in U_t$. Then $f^t(\mathbf{x}) \in X\backslash Y$. By the continuity of f, there exists an open neighborhood $V(\mathbf{x})$ of $\mathbf{x}$ such that $f^t(V(\mathbf{x})) \subset X\backslash Y$. Hence, $V(\mathbf{x}) \subset U_t$. This implies that U_t is open. Since $\gamma^+(\mathbf{x})\cap(X\backslash Y) \neq \emptyset$ for every $\mathbf{x} \in X$, the family of open sets U_t forms an open cover for X. Then, by the compactness of X, there exists a finite subcover $\{U_{t_1}, U_{t_2}, \ldots, U_{t_m}\}$. The forward invariance of $X\backslash Y$ implies $U_t \subset U_{t+1}$. Hence, $X \subset U_T$ holds for $T = \max\{t_1, t_2, \ldots, t_m\}$; i.e., $f^T(X) \subset X\backslash Y$.

Since f is continuous and X is compact, $f^T(X)$ is compact. Let $N = f^T(X)$. Then $\gamma^+(N) = \bigcup_{t=0}^{T-1} f^t(N)$ holds and is compact. Since $\gamma^+(N)$ and Y are compact and $\gamma^+(N) \cap Y = \emptyset$, $d(\gamma^+(N), Y) > 0$ holds. Therefore, we see that $\gamma^+(N)$ is a compact absorbing set for X with $d(\gamma^+(N), Y) > 0$. □

By using this lemma, we can show that if A_0 is primitive, i.e., there are no nontrivial synchronous orbits, then there are no interior orbits converging to $\mathrm{bd}\mathbb{R}^n_+$, as follows.

THEOREM 4.3. *Assume that* (H1)–(H4) *hold. Suppose that $A_{\mathbf{x}}$ is irreducible for all $\mathbf{x} \in \mathbb{R}^n_+$, $\mathrm{sign}(A_{\mathbf{x}}) = \mathrm{sign}(A_0)$ holds for all $\mathbf{x} \in \mathrm{bd}\mathbb{R}^n_+$, and system* (1) *is p-permanent. Then system* (1) *is c-permanent if and only if A_0 is primitive.*

Proof. By Theorem 4.1, the ($\Rightarrow$) part is clear since an imprimitive A_0 leads to a nontrivial synchronous orbit.

Suppose that A_0 is primitive. Since system (1) is p-permanent, by using Lemma 2.1, we can construct a compact absorbing set X for $\mathbb{R}^n_+\backslash O$ such that $X \cap O = \emptyset$. Let $Y = \mathrm{bd}\mathbb{R}^n_+ \cap X$. By Theorem 4.1, for every $\mathbf{x}(0) \in Y$ there exists a $T \geq 0$ such that $\mathbf{x}(T) \in X\backslash Y$. Furthermore, since $A_{\mathbf{x}(t)}$ is irreducible for all $t \geq 0$, $\mathbf{x}(t) \in X\backslash Y$ holds for all $t \geq T$. Otherwise, $A_{\mathbf{x}(t)}$ has a row with only zero entries, and thus $A_{\mathbf{x}(t)}$ is reducible. This fact implies that $X\backslash Y$ is forward invariant. Hence, Lemma 4.2 shows that there exists a compact absorbing set M for X with $d(M, Y) > 0$. This completes the proof. □

Remark. Notice that $A_{\mathbf{x}}$ is assumed to be irreducible not only at $\mathbf{x} = 0$ but also at $\mathbf{x} \in \mathbb{R}^n_+$. If $A_{\mathbf{x}}$ is assumed to be irreducible only at $\mathbf{x} = 0$, then we can construct a matrix function $A_{\mathbf{x}}$ such that (1) has a periodic orbit that visits alternately an interior point and a boundary point. For instance, consider the following example:

$$A_{\mathbf{x}} = \begin{pmatrix} 0 & 16\sigma(x_1, x_2)\exp(-x_1 - x_2) \\ 0.5 & 0.5\sigma(x_1, x_2) \end{pmatrix},$$

where $\sigma(x_1, x_2)$ is the continuous function defined by

$$\sigma(x_1, x_2) = \begin{cases} -x_1x_2 + 1, & 0 \leq x_1x_2 < 1, \\ 0, & x_1x_2 \geq 1. \end{cases}$$

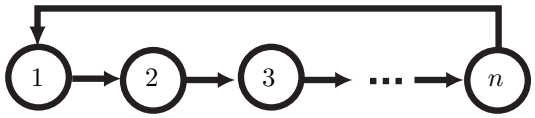


FIG. 1. *The graph of $A_{\mathbf{x}}$ for a semelparous population. This graph has a loop $\{1, 2, \ldots, n, 1\}$, whose length is n. Since there are only loops whose lengths are multiples of n, the greatest common divisor of the lengths are equal to n. Hence, this graph is imprimitive with index of imprimitivity n.*

Note that A_0 is irreducible (and primitive), but $A_{\mathbf{x}}$ is reducible if $x_1 x_2 \geq 1$. Moreover, $\operatorname{sign}(A_{\mathbf{x}}) = \operatorname{sign}(A_0)$ holds for all $\mathbf{x} \in \operatorname{bd}\mathbb{R}^2_+$. We see that $\{(6\ln 2, (3/2)\ln 2), (0, 3\ln 2)\}$ is a periodic orbit of this example.

5. Applications. In this section, we apply the results obtained in the preceding sections to the (density dependent) Leslie matrix model, which was introduced in section 1.

For the functions $f_i(\mathbf{x})$, $i = 1, 2, \ldots, n$, and $p_i(\mathbf{x})$, $i = 1, 2, \ldots, n-1$, we assume the following:

(A1) All $f_i(\mathbf{x})$ and $p_i(\mathbf{x})$ are continuous. $f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_{n-1}(\mathbf{x})$ are nonnegative, and $f_n(\mathbf{x})$ and $p_1(\mathbf{x}), p_2(\mathbf{x}), \ldots, p_{n-1}(\mathbf{x})$ are positive for all $\mathbf{x} \in \mathbb{R}^n_+$.

In order to emphasize that the irreducibility of $A_{\mathbf{x}}$ is determined solely by its sign pattern, in (A1) we do not assume that the functions $p_i(\mathbf{x})$ are less than one. However, from a biological point of view, they must be less than one since they are survival probabilities. In the example studied below, we assume the specific functions $p_i(\mathbf{x})$ that satisfy $0 < p_i(\mathbf{x}) < 1$ for all $\mathbf{x} \in \mathbb{R}^n_+$. It is clear that (A1) ensures that (H1) and (H2) hold. Since $f_n(\mathbf{x})$ and $p_1(\mathbf{x})$, $p_2(\mathbf{x})$, $\ldots$, $p_{n-1}(\mathbf{x})$ are positive for every $\mathbf{x} \geq 0$, the graph $G(A_{\mathbf{x}})$ of $A_{\mathbf{x}}$ has a loop along which we can run through every vertex of the graph (see Figure 1), so that $G(A_{\mathbf{x}})$ is strongly connected (see Theorem A.6 of the appendix). This implies that $A_{\mathbf{x}}$ is irreducible for every $\mathbf{x} \geq 0$. Therefore, (A1) also ensures that (H3) holds. It is clear that dissipativity of the Leslie matrix model is dependent on the forms of the functions f_i and p_i. In fact, if they are all constants, the system becomes linear and hence can exhibit exponential growth. As a nonlinear example, consider the functions $f_i(\mathbf{x})$ and $p_i(\mathbf{x})$:

$$
\begin{aligned}
f_i(\mathbf{x}) &= \frac{\phi_i}{1 + (\sum_{i=1}^n \mu_{ij} x_j)^{\alpha_i}}, \qquad i = 1, 2, \ldots, n, \\
p_i(\mathbf{x}) &= \frac{\sigma_i}{1 + (\sum_{i=1}^n \nu_{ij} x_j)^{\beta_i}}, \qquad i = 1, 2, \ldots, n-1,
\end{aligned} \tag{2}
$$

where the parameters satisfy $\phi_1, \phi_2, \ldots, \phi_{n-1} \geq 0$, $\phi_n > 0$, $0 < \sigma_i < 1$, $\mu_{ij} > 0$, $\nu_{ij} \geq 0$, $\alpha_i > 0$, $\beta_i > 0$ for all i, j. Note that this specific example satisfies the condition (A1) and that $p_1(\mathbf{x}), p_2(\mathbf{x}), \ldots, p_{n-1}(\mathbf{x}) < 1$ hold for all $\mathbf{x} \in \mathbb{R}^n_+$. In this specific case, we can choose $K > 0$ and $0 < \lambda_\infty < 1$ such that

$$
\begin{aligned}
f_1(\mathbf{x}) + p_1(\mathbf{x}) &\leq \lambda_\infty \\
&\vdots \\
f_{n-1}(\mathbf{x}) + p_{n-1}(\mathbf{x}) &\leq \lambda_\infty \\
f_n(\mathbf{x}) &\leq \lambda_\infty
\end{aligned}
$$

hold for all $\mathbf{x} \in \mathbb{R}^n_+$ with $\sum_{i=1}^n x_i \geq K$. Therefore, Theorem 2.2 ensures that the Leslie matrix model with such functions is dissipative; i.e., the assumption (H4) holds.

Let us consider p-permanence of the Leslie matrix model. As shown in Theorem 3.2, the magnitude of the dominant eigenvalue of A_0 plays a crucial role for p-permanence of system (1). The dominant eigenvalue λ_0 of A_0 usually has a strong relationship with the so-called *inherent net reproductive number* $\mathcal{R}_0$, which is defined to be the expected number of offspring per individual per lifetime evaluated by the constant matrix A_0 (see Theorem 1.1.3 of Cushing [7] and Theorem 3 and section 3.1 of Cushing and Yicang [11]). The inherent net reproductive number $\mathcal{R}_0$ of the Leslie matrix is given by

$$\mathcal{R}_0 = \sum_{i=1}^{n} f_i(0) \prod_{j=1}^{i} p_{j-1}(0),$$

where for notational convenience $p_0(0)$ is defined to be 1. For the Leslie matrix model, it is known that $\mathcal{R}_0 > 1$ (resp., $\mathcal{R}_0 < 1$) if and only if $\lambda_0 > 1$ (resp., $\lambda_0 < 1$) (see Cushing [7] and Cushing and Yicang [11]). Therefore, Theorem 3.2 implies that the Leslie matrix model is p-permanent if it is dissipative and $\mathcal{R}_0 > 1$.

Let us consider primitivity of $A_{\mathbf{x}}$ under the assumption (A1). As mentioned above, under the assumption (A1), $A_{\mathbf{x}}$ is irreducible for every $\mathbf{x} \geq 0$. The graph of $A_{\mathbf{x}}$ with $f_1(\mathbf{x}) = f_2(\mathbf{x}) = \cdots = f_{n-1}(\mathbf{x}) = 0$ is depicted in Figure 1. A population with this life cycle is called *semelparous*. In a semelparous population, individuals can reproduce only once in their lives. By Theorem A.5 of the appendix, we see that the index of imprimitivity of $A_{\mathbf{x}}$ for a semelparous population is equal to n, the order of the matrix $A_{\mathbf{x}}$; that is, $A_{\mathbf{x}}$ is not primitive for all $\mathbf{x} \geq 0$. Therefore, Theorem 4.1 ensures that a semelparous population has a nontrivial synchronous orbit. On the other hand, consider the case where $f_1(\mathbf{x}) = f_2(\mathbf{x}) = \cdots = f_{n-1}(\mathbf{x}) = 0$ does not hold. By Theorem A.5, we see that if there are two consecutive fertile age-classes such that $f_i(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathrm{bd}\mathbb{R}^n_+$, then $A_{\mathbf{x}}$ is primitive for all $\mathbf{x} \in \mathrm{bd}\mathbb{R}^n_+$. Hence, if $f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathrm{bd}\mathbb{R}^n_+$, i.e., all age-classes are fertile, then $A_{\mathbf{x}}$ is clearly primitive for all $\mathbf{x} \in \mathrm{bd}\mathbb{R}^n_+$. In such a primitive case, Theorem 4.3 ensures that all age-classes coexist if system (1) is p-permanent.

Figure 2 considers the dynamics of the Leslie matrix model with four age-classes. We use the functions f_i and p_i defined by (2). Figure 2(a) shows the population dynamics of the fourth age-class in an imprimitive Leslie matrix model. From this figure, we see that the orbit converges to a nontrivial synchronous orbit, where all but one year class are missing. If individuals in the third age-class are also fertile, then the orbit stays in the interior of the nonnegative cone. So, we see that all classes coexists as ensured by Theorem 4.3 (see Figures 2(b) and (c)).

6. Discussion. In this paper, we have considered the existence of nontrivial synchronous orbits in a general class of matrix population models. In Theorem 4.1, we showed that the primitivity of the matrix $A_{\mathbf{x}}$ on the boundary $\mathrm{bd}\mathbb{R}^n_+$ is essential for this existence. Furthermore, in Theorem 4.3, we showed that if there are no nontrivial synchronous orbits, then all classes coexist in the sense of c-permanence. By using the specific Leslie matrix model, we confirmed these results in section 5.

Since Theorem 4.1 ensures only existence of a nontrivial synchronous orbit, that orbit's stability is unknown. However, in our example in Figure 2(a), the nontrivial synchronous orbit seems to be stable. It is a future problem to consider the relationship between stability of synchronous orbits and structure of matrix population

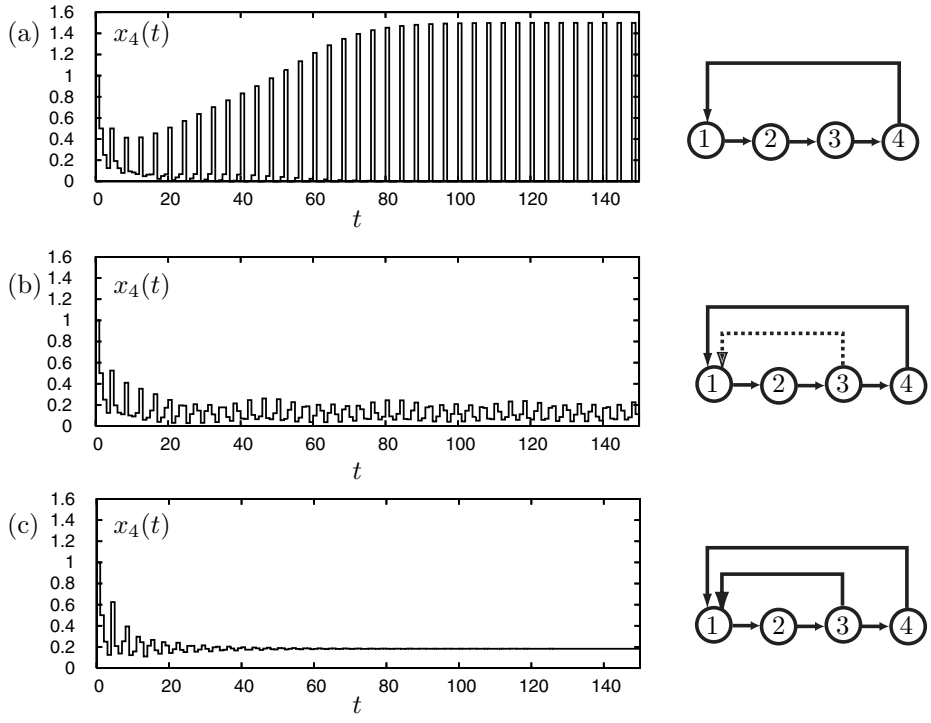


FIG. 2. *The population dynamics of the Leslie matrix model with four age-classes. The left figures show temporal fluctuations of the fourth age-class density. The parameters are* $\phi_1 = 0$, $\phi_2 = 0$, $\phi_4 = 20$, $\sigma_1 = \sigma_2 = \sigma_3 = 0.5$, $\mu_{ij} = \nu_{ij} = \alpha_i = \beta_i = 1$ *for all* i, j, *and the initial condition satisfies* $x_1(0) = x_2(0) = x_3(0) = x_4(0) = 1$. *The parameter* ϕ_3 *is chosen as follows:* (a) $\phi_3 = 0$, (b) $\phi_3 = 1$, (c) $\phi_3 = 5$.

models (see [4, 9, 10, 12, 19, 24] for stability of synchronous orbits in semelparous populations).

In the definition of c-permanence (Definition 1.2), all nonzero orbits are required to be attracted by some compact set in the interior of the nonnegative cone, $\mathrm{int}\mathbb{R}^n_+$. However, we often observe the case where all positive orbits are attracted by some compact set in $\mathrm{int}\mathbb{R}^n_+$ even if the system has a nontrivial synchronous orbit; i.e., the system is not c-permanent. For example, in the Leslie matrix model for a semelparous population, we can find this type of class coexistence. Therefore, it is an important future problem to study class coexistence involving synchronous orbits.

Appendix. In this section, we list some useful theorems of nonnegative matrices. There are several books which discuss the properties of such matrices (e.g., see [2, 3, 5, 13, 23]).

One of the most important properties of nonnegative matrices is irreducibility, which is defined as follows.

DEFINITION A.1 (irreducibility). *A square matrix* A *is said to be irreducible if it can be rearranged into the following form by renumbering the indices of rows and columns:*

$$\begin{pmatrix} B & 0 \\ C & D \end{pmatrix},$$

where B and D are square matrices and 0 *denotes the matrix with only zero entries. Otherwise A is called irreducible.*

An irreducible nonnegative matrix can have multiple eigenvalues whose magnitudes are equal to the magnitude of the dominant eigenvalue λ. By the number of such eigenvalues, irreducible nonnegative matrices are classified as follows.

DEFINITION A.2. *Let A be an irreducible nonnegative matrix that has h eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_h$, whose magnitudes are equal to the magnitude of the dominant eigenvalue $\lambda = \lambda_1$. A is called primitive if $h = 1$, and imprimitive if $h > 1$. h is called the index of imprimitivity of A.*

The theory of nonnegative matrices has a strong relationship with a graph theory.

DEFINITION A.3. *The associated directed graph, $G(A)$, of an $n \times n$ matrix A consists of n vertices $P_1, P_2, \ldots, P_n$, where an edge leads from P_j to P_i if $a_{ij} \neq 0$. A directed graph G is said to be strongly connected if for any ordered pair (P_i, P_j) of vertices of G there exists a path which leads from P_i to P_j. Let $P = \{P_{i_0}, P_{i_1}, \ldots, P_{i_\ell}\}$ be a path in a graph G. Then ℓ is the length of P. P is a loop if $P_{i_0} = P_{i_\ell}$.*

Irreducibility and the index of imprimitivity are characterized by directed graphs as follows.

THEOREM A.4 (e.g., see Theorem 2.2.7 of [3]). *A matrix A is irreducible if and only if $G(A)$ is strongly connected.*

THEOREM A.5 (e.g., see Theorem 2.2.30 of [3]). *Let A be an irreducible nonnegative matrix. The index of imprimitivity of A is equal to the greatest common divisor of the lengths of loops in $G(A)$.*

Remark. This theorem shows that indices of imprimitivity h (like irreducibility) depend only on the pattern of a matrix; i.e., every irreducible nonnegative matrix that has positive entries in exactly the same positions has the same index of imprimitivity.

The following two theorems are utilized in obtaining Theorem 4.1.

THEOREM A.6 (e.g., see Theorem 13.8 of [13]). *A nonnegative square matrix A is primitive if and only if there exists an integer $k \geq 1$ such $A^k > 0$.*

THEOREM A.7 (e.g., see Corollary 13.2 of [13]). *If A is an imprimitive matrix with index of imprimitivity h, then A^h can be rearranged into the following quasi-diagonal form by renumbering the indices of rows and columns:*

$$\operatorname{diag}\{A_1, A_2, \ldots, A_h\} = \begin{pmatrix} A_1 & 0 & \ldots & 0 \\ 0 & A_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & A_h \end{pmatrix}, \tag{3}$$

where $A_1, A_2, \ldots, A_h$ are primitive matrices with the same dominant eigenvalue and 0 *denotes the matrix with only zero entries.*

Acknowledgments. I would like to thank two anonymous referees for their helpful comments.

REFERENCES

[1] L. J. S. ALLEN, *A density-dependent Leslie matrix model*, Math. Biosci., 95 (1989), pp. 179–187.
[2] A. BERMAN, M. NEUMANN, AND R. J. STERN, *Nonnegative Matrices in Dynamic Systems*, Pure Appl. Math., John Wiley and Sons, New York, 1989.
[3] A. BERMAN AND R. J. PLEMMONS, *Nonnegative matrices in the mathematical sciences*, Classics in Appl. Math. 9, SIAM, Philadelphia, PA, 1994.
[4] M. G. BULMER, *Periodical insects*, Amer. Natur., 111 (1977), pp. 1099–1117.
[5] H. CASWELL, *Matrix Population Models*, 2nd ed., Sinauer Associates, Sunderland, MA, 2001.

[6] P. Cull and A. Vogt, *The periodic limit for the Leslie model*, Math. Biosci., 21 (1974), pp. 39–54.
[7] J. M. Cushing, *An introduction to structured population dynamics*, CBMS-NSF Reg. Conf. Ser. Appl. Math. 71, SIAM, Philadelphia, PA, 1998.
[8] J. M. Cushing, *Cycle chains and the LPA model*, J. Difference Equations Appl., 9 (2003), pp. 655–670.
[9] J. M. Cushing and J. Li, *Intra-specific competition and density dependent juvenile growth*, Bull. Math. Biol., 54 (1992), pp. 503–519.
[10] J. M. Cushing and J. Li, *The dynamics of a size-structured intraspecific competition model with density dependent juvenile growth rates*, in Individual-based Models and Approaches in Ecology: Populations, Communities, and Ecosystems, D. L. DeAngelis and L. J. Gross, eds., Routledge, Chapman, and Hall, New York, 1992, Chapter 6, pp. 112–125.
[11] J. M. Cushing and Z. Yicang, *The net reproductive value and stability in matrix population models*, Natur. Resource Modeling, 8 (1994), pp. 297–333.
[12] N. V. Davydova, O. Diekmann, and S. A. van Gils, *Year class coexistence or competitive exclusion for strict biennials?*, J. Math. Biol., 46 (2003), pp. 95–131.
[13] F. R. Gantmacher, *The theory of matrices*, Vol. 1, translated from the Russian by K. A. Hirsch, AMS Chelsea Publishing, Providence, RI, 1998 (reprint of the 1959 translation).
[14] K. Heliövaara, R. Väisänen, and C. Simon, *Evolutionary ecology of periodical insects*, Trends in Ecology and Evolution, 9 (1994), pp. 475–480.
[15] V. Hutson, *A theorem on average Liapunov functions*, Monatsh. Math., 98 (1984), pp. 267–275.
[16] R. Kon, Y. Saito, and Y. Takeuchi, *Permanence of single-species stage-structured models*, J. Math. Biol., 48 (2004), pp. 515–528.
[17] M. Lloyd and H. S. Dybas, *The periodical cicada problem.* I. *Population ecology*, Evolution, 20 (1966), pp. 133–149.
[18] R. M. May, *Periodical cicadas*, Nature, 277 (1979), pp. 347–349.
[19] E. Mjølhus, A. Wikan, and T. Solberg, *On synchronization in semelparous populations*, J. Math. Biol., 50 (2005), pp. 1–21.
[20] M. G. Neubert and H. Caswell, *Density-dependent vital rates and their population dynamic consequences*, J. Math. Biol., 41 (2000), pp. 103–121.
[21] T. M. Powledge, *The* 17*-year itch*, Scientific American, 290 (2004), pp. 32–33.
[22] J. A. L. Silva, M. L. de Castro, and D. A. R. Justo, *Synchronism in a metapopulation model*, Bull. Math. Biol., 62 (2000), pp. 337–349.
[23] R. S. Varga, *Matrix Iterative Analysis*, 2nd ed., Springer Ser. Comput. Math. 27, Springer-Verlag, Berlin, 2000.
[24] A. Wikan and E. Mjølhus *Overcompensatory recruitment and generation delay in discrete age-structured population models*, J. Math. Biol., 35 (1996), pp. 195–239.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 627–647

MODELING INTERVENTION MEASURES AND SEVERITY-DEPENDENT PUBLIC RESPONSE DURING SEVERE ACUTE RESPIRATORY SYNDROME OUTBREAK*

SZE-BI HSU† AND YING-HEN HSIEH‡

Abstract. The 2003 severe acute respiratory syndrome (SARS) epidemic came and left swiftly, resulting in more than 8,000 probable cases worldwide and 774 casualties. It is generally believed that quarantine of those individuals suspected of being infected was instrumental in quick containment of the outbreaks. In this work we propose a differential equation model that includes quarantine and other intervention measures implemented by the health authority, including those to prevent nosocomial infections and decrease frequency of contacts among the general public. We also consider the possible behavior change by the general populace to avoid infection, in response to the severity of the outbreak in general and to these intervention measures in particular. Complete analysis is given for the model without quarantine. For the general model with quarantine, a basic reproduction number is derived and full description of its dynamics is provided. We will show that introducing quarantine measures in the model could produce bistability in the system, thus changing the basic dynamics of the model. We give numerical examples of parameter values with which bistable steady states, where one is disease-free and the other endemic, could exist. However, realistic parameter values indicate that, assuming limited imported cases, the occurrence of the stable endemic steady state or bistability is unlikely. The modeling results indicate that for an infectious disease with infectivity and patterns of transmission typical of SARS, the outbreak can always be eradicated by implementing border control of imported cases and limited quarantine, along with the public's social response to avoid infections. Moreover, the results also suggest that quarantine measures will be effective in reducing infections only if the quarantined/isolated SARS patients and their potential contacts can successfully reduce their contact rate and/or transmission probabilities. Hence the effectiveness of quarantine for infectious diseases like SARS, for which no infection is being prevented during the quarantine period, can only be indirect and therefore must be combined with other intervention measures in order to quickly contain the outbreaks.

Key words. SARS, mathematical model, basic reproduction number, quarantine, bistable steady states, Taiwan

AMS subject classifications. 92D25, 92D30, 34D23, 93D20

DOI. 10.1137/040615547

1. Introduction.

The worldwide severe acute respiratory syndrome (SARS) epidemic outbreak of November 2002–July 2003 accounted for more than 8,000 infections with 774 fatalities directly attributable to SARS [1]. It is generally believed [2] that the experience of affected regions showed that the transmission of SARS-Coronavirus (SARS-CoV) can be effectively controlled by adherence to basic public health measures, including rapid case detection, case isolation, contact tracing, and good infection control such as hand-washing and use of personal protective equipment. Another measure believed to be instrumental in breaking the transmission chain is the quarantine of well but potentially infective individuals to prevent infections [3, 4, 5, 6]. During the outbreak in Taiwan from April to June 2003, the

*Received by the editors September 22, 2004; accepted for publication (in revised form) June 7, 2005; published electronically January 6, 2006.

http://www.siam.org/journals/siap/66-2/61554.html

†Department of Mathematics, National Tsing Hua University, Hsinchu, Taiwan (sbhsu@math.nthu.edu.tw). This author was supported by National Science Council of Taiwan grant NSC 93-2115-M-007-002.

‡Corresponding author. Department of Applied Mathematics, National Chung Hsing University, Taichung, Taiwan (hsieh@amath.nchu.edu.tw). This author was supported by National Science Council of Taiwan grant NSC 93-2751-B005-001-Y.

health authority attempted to quarantine more than 150,000 people who either had possible contact with a suspected SARS case or had just arrived from an affected area as determined by the World Health Organization (WHO). Of these quarantined individuals, only 17 were later officially confirmed as SARS cases. Hence questions remain as to the effectiveness of the quarantine.

During the outbreak, two distinct levels of quarantine were implemented in Taiwan. Level A quarantine, aimed at people suspected of having close contact with a suspected SARS case, was implemented on March 18, 2003. Level B quarantine, aimed at travelers from affected areas, was initiated on April 28 in the aftermath of the first SARS fatality in Taiwan on April 26. Details of the implementation of quarantine measures in Taiwan were described in [7]. By the end of the summer, a total of more than 150,000 people had been quarantined during the SARS outbreak.

There were 346 officially confirmed SARS cases as defined by WHO during the outbreak of 2003 in Taiwan, among which were 37 direct SARS casualties and 36 SARS-related deaths. In addition, 180 patients, who either had a previous negative PCR or antibody test or had been suspected or ruled-out cases, tested SARS antibody positive. However, Level B quarantine detected no confirmed SARS cases, while Level A quarantined persons included 17 officially confirmed SARS cases and 7 suspected or ruled-out cases with positive antibody tests [8]. Using the case data of the 480 laboratory-confirmed SARS cases, [8] showed that, compared to all other patients, previously quarantined persons had a significantly shorter onset-to-diagnosis time, i.e., the time it took a person with onset of symptoms to be diagnosed with suspected SARS and hospitalized. Hence quarantine had at least been useful in attaining more rapid detection and hospitalization of cases.

Rapid case definition also depends on knowledge regarding the clinical and molecular aspect of the disease, an inherently difficult task when facing a newly emerging disease like SARS. Contact tracing and quarantine of the traced contacts is another effective but difficult measure especially in an established democratic society, due to the ethical and legal ramifications [9]. Adherence to infection control, in the hospital or in the community, by the health care workers or the general populace, depends very much on the individual. The personal decision whether to diligently avoid contacts and infection is often based on the circumstances, i.e., whether there is any perceived cause for behavior change by the individual. The increasing severity of an outbreak or the implementation of massive intervention measures, e.g., the images of everyone wearing a face mask while in public places, is surely a cause for behavior change to avoid infection. This perhaps critically important factor will also be considered in our model.

In this work we will focus on three types of interventions evident during the past SARS outbreak: quarantine of potential infectives, isolation of suspected cases, and behavior change of the general public (including health care workers) in response to the increasing severity of the outbreak in an effort to avoid contacts which might lead to SARS infection. The focus is to study the roles played by intervention measures and social response in the quick containment of the outbreak. Previous modeling work of the SARS epidemic includes the early modeling of SARS by [10, 11] to obtain the all-important basic reproduction number for SARS, [12] on modeling the community and hospital transmission of SARS, and [13, 14] on models for data-fitting of SARS in Taiwan. Also see [15] for a review of mathematical models of SARS. Recent modeling work of epidemics with intervention measures (quarantine, vaccination, evacuation, etc.) includes [16, 17, 18, 19, 20] on smallpox, [17] on flu, [21] on bubonic plague, [22] on measles and whooping cough, [23] on optimal intervention strategies, and

[18] on a class of infectious disease models with quarantine.

This article is organized as follows: In section 2 we describe the general model with the intervention measures to be considered and the computation of basic reproduction numbers. In section 3 we give the complete analysis of the model with severity-dependent public response but without quarantine. Section 4 gives analytical results for the full model with quarantine and full description of its dynamics. Finally in section 5 we discuss the biological significance of our results.

2. The model. In this work, we propose a general model with Level A and B quarantines, as well as imported cases who entered the exposed class upon their arrival before April 28, but were quarantined (Level B) as they entered from the affected areas after April 28. The model variables are given as follows; note that the time unit is in days:

S—the number of susceptible individuals at time t;

E—the number of infected asymptomatic persons at time t;

Q_A—the number of asymptomatic infected persons at time t under Level A quarantine;

Q_B—the number of imported asymptomatic infected persons at time t (who were under B quarantine if arriving from affected areas after April 28);

I—the number of infective persons with onset of symptoms not isolated or quarantined at time t;

P—the number of isolated probable SARS cases at time t;

D—the cumulative number of SARS deaths at time t;

R—the cumulative number of discharged SARS patients at time t.

The key assumptions used are as follows:

1. A SARS-infected person is infective after onset of symptoms.

2. A quarantined person is quarantined without symptoms (hence is not infective), becoming infective with reduced contact rates due to quarantine, and is isolated upon diagnosis.

3. An infected person can infect others unless quarantined or isolated as a probable case with reduced contact rate depending on the effectiveness of the isolation. The underlying assumption here is that once diagnosed as a probable SARS case and hospitalized, a patient cannot infect others.

4. A probable case is removed from isolation either by death or discharge.

5. As behavior change by individuals occurs as a result of public response to the severity of the outbreak, the infection rate (or the product of transmission probability and contact rate) decreases with the increasing cumulative number of probable cases. Similarly, the effectiveness of quarantine and isolation also increases with the increasing number of probable cases, resulting in a decreased number of infections. To account for this decrease, we make use of a rational function $\frac{1}{1+a[P(t)+R(t)+D(t)]}$, where $P+R+D$ is the cumulative number of probable cases. We note that the decreasing rational function used, which resembles Holling's functional response in predator-prey models [24], is not the only choice of function to portray the phenomenon in question. A decreasing exponential function, for example, could do just as well.

6. We assume homogeneous mixing with quarantine-adjusted incidence.

7. Quarantine for Level A is proportionate to the number of infected asymptomatic persons.

8. Imported cases are a function of time ($Q(t) = 0$, 1, or 2 as deduced from data), with Level B quarantine after April 28.

The model parameters are as follows:

λ—infection rate due to contact with infective class;

q_1—proportion of recruitment of asymptomatic infected persons for Level A quarantine;

γ_3—isolation rate of infectives not under quarantine;

μ—progression rate from exposure to onset of symptoms;

γ_i, $i = 1, 2$—isolation rates of Q_A and Q_B, respectively;

α_A, α_B, α_P—the proportionate reduction in infectivity of quarantined persons due to Level A and B quarantines (before isolation) and probable cases, respectively;

ρ_i, $i = 1, 2$—respective fatality rates of infective cases and isolated probable SARS patients;

σ_i, $i = 1, 2$—respective discharge rates of infective cases and isolated probable SARS patients;

c—contact rate in absence of an outbreak;

a—the effect of behavior change in reduction of contact due to the cumulative number of probable cases;

β—transmission probability per effective contact.

The flowchart for the model is given in Figure 2.1.

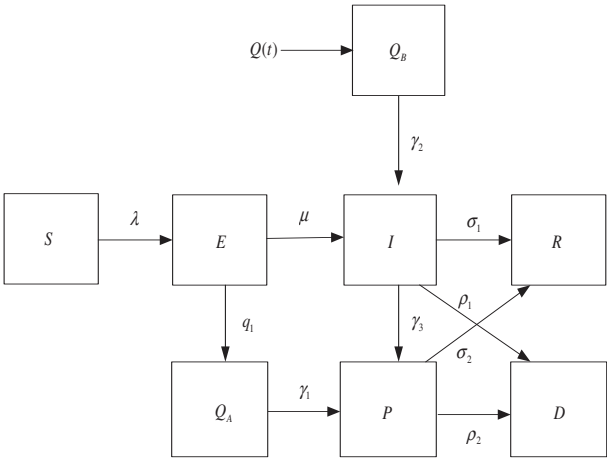


FIG. 2.1. *Flowchart for the model.*

Finally, the model equations with imported cases, Level A and B quarantines, and behavior change are as follows:

$$S' = -\lambda(S, E, I, Q_A, Q_B, P, R, D)S, \tag{2.1}$$

$$E' = \lambda(S, E, I, Q_A, Q_B, P, R, D)S - \mu E - q_1 E, \tag{2.2}$$

$$Q_A' = q_1 E - \gamma_1 Q_A, \tag{2.3}$$

$$Q_B' = Q(t) - \gamma_2 Q_B, \tag{2.4}$$

$$I' = \mu E + \gamma_2 Q_B - (\sigma_1 + \rho_1 + \gamma_3)I, \tag{2.5}$$

$$P' = \gamma_1 Q_A + \gamma_3 I - (\sigma_2 + \rho_2)P, \tag{2.6}$$

$$R' = \sigma_1 I + \sigma_2 P, \tag{2.7}$$

$$D' = \rho_1 I + \rho_2 P, \tag{2.8}$$

where the incidence of infection with quarantine is given by

$$\lambda(S,E,I,Q_A,Q_B,P,R,D) = \beta\frac{c}{1+a(P+R+D)} \times \frac{I+\alpha_A Q_A+\alpha_B Q_B+\alpha_P P}{S+E+I+\alpha_A Q_A+\alpha_B Q_B+\alpha_P P}.$$

Note that $S+E+Q_A+Q_B+I+P+R+D = S_0+I_0$, where S_0 and I_0 are the initial susceptible population sizes. The system is nonautonomous, due to the imported case term $Q(t)$ in the right-hand side of the equation for Q_B'. Brauer and van den Driessche [25] have shown that, if there is a positive flow of infectives into the population, disease-free equilibrium might not exist. However, during the SARS outbreak border control was implemented in Taiwan as well as all other affected areas. Therefore we can reasonably assume that $Q(t)$ has compact support and subsequently the asymptotic properties of the nonautonomous system given in (2.1)–(2.8) are the same as the corresponding autonomous system, i.e., with $Q(t) = 0$. Hence we need only consider the autonomous system hereafter.

For the model without quarantine, the model equations become

$$S' = -\lambda(S,E,I,P,D)S, \tag{2.9}$$

$$E' = \lambda(S,E,I,P,D)S - \mu E, \tag{2.10}$$

$$I' = \mu E - (\sigma_1+\rho_1+\gamma_3)I, \tag{2.11}$$

$$P' = \gamma_3 I - (\sigma_2+\rho_2)P, \tag{2.12}$$

$$R' = \sigma_1 I + \sigma_2 P, \tag{2.13}$$

$$D' = \rho_1 I + \rho_2 P, \tag{2.14}$$

where

$$\lambda(S,E,I,P,R,D) = \beta\left[\frac{c}{1+a(P+R+D)}\right]\frac{I+\alpha_P P}{S+E+I+\alpha_P P}.$$

The disease-free equilibrium (DFE) for the six-dimensional system in (S,E,I,P,R,D) is $(S^*,0,0,0,R^*,D^*)$ with $S^*+R^*+D^* = S_0+I_0$; the endemic equilibrium is $(0,0,0,0,R^{\#},D^{\#})$ with $R^{\#}+D^{\#} = S_0+I_0$.

Making use of the method in [26], we obtain the expression for the basic reproduction number R_0 of this case:

$$R_0 = \frac{\beta c}{(\sigma_1+\rho_1+\gamma_3)[1+a(R^*+D^*)]} + \frac{\beta c\alpha_P\gamma_3}{(\sigma_1+\rho_1+\gamma_3)[1+a(R^*+D^*)](\sigma_2+\rho_2)}. \tag{2.15}$$

Similarly as for the original model with quarantine, we have the more general expression for the effective basic reproduction number with quarantine R_Q, again using the procedure developed in [26]:

$$R_Q = \beta\frac{c}{[1+a(R^*+D^*)]}\left\{\frac{\mu}{(\sigma_1+\rho_1+\gamma_3)[\mu+q_1]}+\frac{\alpha_A q_1}{\gamma_1[\mu+q_1]}\right\} + \beta\frac{c}{[1+a(R^*+D^*)]}\frac{\alpha_P}{(\sigma_2+\rho_2)}\left\{\frac{\gamma_3}{(\sigma_1+\rho_1+\gamma_3)}\frac{\mu}{[\mu+q_1]}+\frac{q_1}{\mu+q_1}\right\}. \tag{2.16}$$

Note that both R_0 and R_Q have very clear biological interpretations which will be discussed in section 5.

3. Analysis for model without quarantine. In this section we provide full analysis for the model without quarantine. To simplify, we let $\alpha_p = 0$. That is, hospitalized and isolated probable cases do not make a significant contribution to the infections, as indicated by the result in a data-motivated modeling study of Taiwan's SARS outbreak in [27]. We note that while it is true that nosocomial infections played a crucial role during the SARS outbreak in all affected, as confirmed by the fact that nearly 80% of SARS infections in Taiwan occurred nosocomially [28], most had occurred *before* the infective individuals had been diagnosed with SARS and hospitalized with adequate isolation. Only a small number of infections in Taiwan as well as in other affected areas have been documented as being caused by a confirmed or probable SARS patient who most likely had been isolated.

Hence the system in (2.9)–(2.14) becomes

$$S' = -\frac{\beta IS}{E+I+S}\frac{c}{1+a(P+R+D)}, \tag{3.1}$$

$$E' = \frac{\beta IS}{E+I+S}\frac{c}{1+a(P+R+D)} - \mu E, \tag{3.2}$$

$$I' = \mu E - (\sigma_1+\rho_1+\gamma_3)I, \tag{3.3}$$

$$P' = \gamma_3 I - (\sigma_2+\rho_2)P, \tag{3.4}$$

$$R' = \sigma_1 I + \sigma_2 P, \tag{3.5}$$

$$D' = \rho_1 I + \rho_2 P, \tag{3.6}$$

with $S(0) = S_0 > 0$, $I(0) = I_0 > 0$, $E(0) = P(0) = R(0) = D(0) = 0$.

We first give the following lemma, the proof of which is in [29].

LEMMA 3.1. *Let f be continuously differentiable. If $f(t) \longrightarrow$ constant as $t \to \infty$ and $|f''(t)| \le M$ for all t, then $f'(t) \to 0$ as $t \to \infty$.*

THEOREM 3.2. *We have the following asymptotic properties: $S(t) \longrightarrow S_\infty \ge 0$, $R(t) \longrightarrow R_\infty > 0$, $D(t) \longrightarrow D_\infty > 0$, and $\lim_{t\to\infty} I(t) = 0$, $\lim_{t\to\infty} E(t) = 0$, $\lim_{t\to\infty} P(t) = 0$.*

Proof. Obviously $S(t)$ is monotone decreasing and bounded below; hence $\lim_{t\to\infty} S(t) = S_\infty \ge 0$, $0 \le S_\infty < S_0$. Moreover, $(E+S)' = -\mu E$, and therefore $E(t)+S(t)$ is monotone decreasing for $t \ge 0$. Hence $E(t) \longrightarrow E_\infty \ge 0$ as $t \to \infty$. Since $S(t)+E(t)+I(t)+P(t)+R(t)+D(t) \equiv N = S_0 + I_0$ for all t, $R' \ge 0$, $D' \ge 0 \Longrightarrow$ $R(t) \longrightarrow R_\infty > 0$, $D(t) \longrightarrow D_\infty > 0$. Obviously $R'' = \sigma_1 I' + \sigma_2 P'$, $|I'|$ and $|P'|$ are bounded, and hence $|R''| \le M$ for some $M > 0$. Consequently by Lemma 3.1 $I(t) \longrightarrow I_\infty = 0$, $P(t) \longrightarrow P_\infty = 0$ as $t \to \infty$.

Claim: $E_\infty = 0$. Suppose $E_\infty > 0$; then $I' = \mu E - (\sigma_1+\rho_1+\gamma_3)I \ge \mu(E_\infty - \varepsilon) - (\sigma_1+\rho_1+\gamma_3)\varepsilon > 0$ for ε small, t large. It follows that $I(t)$ becomes unbounded. This is a contradiction. □

Next, we let $q = \sigma_1 + \rho_1 + \gamma_3$ and $c = 1$ (i.e., β denotes contact rate times transmission probability) for the sake of simplicity. We also, for the moment, assume $a = 0$, i.e., no behavior change. We will return to discuss the case with behavior change later in this section. Subsequently, we consider the following simplified system:

$$S' = -\frac{\beta IS}{E+I+S}, \tag{3.7}$$

$$E' = \frac{\beta IS}{E+I+S} - \mu E, \tag{3.8}$$

$$I' = \mu E - qI, \tag{3.9}$$

with $S(0) = S_0 > 0$, $I(0) = I_0 > 0$, $E(0) = 0$.

Now, we let $W_1 = S/I$, $W_2 = E/I$. Then (3.7)–(3.9) become

$$W_1' = -\frac{\beta W_1}{1 + W_1 + W_2} - W_1(\mu W_2 - q), \tag{3.10}$$

$$W_2' = \frac{\beta W_1}{1 + W_1 + W_2} - \mu W_2 - W_2(\mu W_2 - q), \tag{3.11}$$

with $W_1(0) > 0$, $W_2(0) = 0$.

To study the flow of (3.10)–(3.11) in the W_2W_1-phase plane, we first consider the isoclines $W_1' = 0$ and $W_2' = 0$. Clearly, $W_1' < 0$ if $W_2 > q/\mu$. Moreover,

$$\begin{aligned} W_1' \geq 0 &\Longleftrightarrow q - \mu W_2 \geq \frac{\beta}{1 + W_1 + W_2} \\ &\Longleftrightarrow 1 + W_1 + W_2 \geq \frac{\beta}{q - \mu W_2} \quad \text{for} \quad q - \mu W_2 \geq 0 \\ &\Longleftrightarrow W_1 \geq \frac{\beta}{q - \mu W_2} - (1 + W_2) = f(W_2) \quad \text{for} \quad q - \mu W_2 \geq 0, \end{aligned}$$

and

$$\begin{aligned} W_2' \geq 0 &\Longleftrightarrow \frac{\beta W_1}{1 + W_1 + W_2} \geq [\mu + (\mu W_2 - q)]W_2 \\ &\Longleftrightarrow \beta W_1 \geq [\mu + (\mu W_2 - q)]W_2(1 + W_1 + W_2) \\ &\Longleftrightarrow \beta - W_2[\mu + (\mu W_2 - q)]W_1 > [\mu + (\mu W_2 - q)]W_2(1 + W_2) \\ &\Longleftrightarrow W_1 > \frac{W_2(1 + W_2)[\mu + (\mu W_2 - q)]}{\beta - W_2[\mu + (\mu W_2 - q)]} = g(W_2). \end{aligned}$$

Consequently, it is easy to verify that the curves $W_1 = f(W_2)$ and $W_1 = g(W_2)$ do not intersect.

There are four cases to be considered:

1. $q < \beta$, $q < \mu$.
   Let $\widetilde{W_2}$ be the positive root of
   $$h(W_2) = \beta - W_2[\mu + (\mu W_2 - q)] = 0.$$
   Clearly $h(\frac{q}{\mu}) = \beta - q > 0$. Hence $\frac{q}{\mu} < \widetilde{W_2}$. In the first quadrant of the W_2W_1-phase plane, the isocline $W_1 = 0$, $(W_1 = f(W_2))0 \leq W_2 < \frac{q}{\mu}$, satisfies $f(0) = \frac{\beta}{q} - 1 > 0$ and $f((\frac{q}{\mu})^-) = \infty$. The isocline $\dot{W}_2 = 0 (W_1 = g(W_2))$ satisfies $g(0) = 0$, $g(\widetilde{W_2}-) = \infty$. We note that the isocline $\dot{W}_1 = 0$ is above that of $\dot{W}_2 = 0$. Every trajectory converges to the endemic equilibrium $(S/I, E/I) = (0, 0)$ as $t \to \infty$.
2. $\beta > q$, $\mu < q$.
   There are two equilibria $(0, 0)$ and $(W_2^*, 0)$, where $W_2^* = \frac{q}{\mu} - 1$. Similar to case 1, every trajectory converges to $(W_2^*, 0)$.
3. $\beta < q$, $\mu < q$.
   Clearly $f(W_2^*) = f(\frac{q}{\mu} - 1) = \frac{\beta}{\mu} - \frac{q}{\mu} < 0$. Observe that $h(\frac{q}{\mu}) = \beta - q < 0$ and we have $\widetilde{W_2} < \frac{q}{\mu}$. Since the isocline $\dot{W}_2 = 0$ is above that of $\dot{W}_1 = 0$, it follows that, as $t \longrightarrow \infty$, $h(W_2(t)) \longrightarrow \widetilde{W_2}$, the positive root of $h(W_2) = 0$, and $W_1(t) \longrightarrow \infty$. Moreover, $(W_1(t), W_2(t))$ approaches the curve $W_1 = g(W_2)$, i.e., $W_2 = 0$, as $t \to \infty$.

4. $\beta < q$, $q < \mu$.
   Obviously, this case is similar to the previous case with $W_2(t) \to \widetilde{W_2}$ and $W_1(t) \to \infty$ as $t \to \infty$.

We then have the following theorem.

THEOREM 3.3. *For system* (3.7)–(3.9), *if* $\beta > q$, *then* $S(t) \to 0$ *as* $t \to \infty$. *If* $\beta < q$, *then* $S(t) \to S_\infty > 0$ *as* $t \to \infty$.

Proof. For cases 1 and 2 as described earlier, $\frac{S(t)}{I(t)} = W_1(t) \to 0$ as $t \to \infty$. Since $I(t) \longrightarrow 0$ and $S(t) \longrightarrow S_\infty \geq 0$, we have $S_\infty = 0$. Hence, $\beta > q$ implies $S(t) \longrightarrow 0$ as $t \to \infty$. If $\beta < q$, then, by cases 3 and 4, we have $W_1(t) \longrightarrow \infty$ and $W_2(t) \longrightarrow \widetilde{W_2} > 0$ as $t \to \infty$.

Claim: $S_\infty > 0$. If $S_\infty = 0$, then

$$\int_0^\infty \frac{1}{1 + W_1(t) + W_2(t)} dt = \infty \quad \text{and} \quad \int_0^\infty \frac{1}{W_1(t)} dt \geq \int_0^\infty \frac{1}{1 + W_1(t) + W_2(t)} dt = \infty.$$

Therefore we have $\int_0^\infty \frac{1}{W_1(t)} dt = \infty$.

From (4.10),

$$W_1' = \frac{-\beta W_1}{1 + W_1 + W_2} - W_1(\mu W_2 - q) \geq \frac{-\beta W_1}{1 + W_1 + \widetilde{W_2} - \epsilon} + W_1[q - \mu(\widetilde{W_2} + \epsilon)],$$

where $\widetilde{W_2} - \epsilon < W_2(t) < \widetilde{W_2} + \epsilon$ for $t \geq t_0$.

For $t \geq t_0$,

$$\frac{W_1'}{W_1} \geq \frac{-\beta}{1 + W_1 + \widetilde{W_2} - \epsilon} + [q - \mu(\widetilde{W_2} + \epsilon)].$$

Because $W_1(t) \to \infty$ as $t \to \infty$,

$$\frac{W_1'}{W_1} \geq \frac{1}{2}[q - \mu(\widetilde{W_2} + \epsilon)] > 0 \qquad \text{for } t \geq T, \text{ for some } T \text{ large.}$$

Therefore $W_1 \to \infty$ exponentially, and

$$W_1(t) \geq W_1(T) \exp\left\{\frac{1}{2}[q - \mu(\widetilde{W_2} + \epsilon)](t - T)\right\}.$$

But

$$\infty = \int_T^\infty \frac{1}{W_1(t)} dt \leq \int_T^\infty \frac{1}{W_1(T) \exp\{\frac{1}{2}[q - \mu(\widetilde{W_2} + \epsilon)(t - T)]\}} dt < \infty.$$

This is a contradiction. □

Now we return to consider system (3.1)–(3.6) with behavior change.

THEOREM 3.4. *Let* $\widetilde{\beta} = \frac{\beta c}{1 + aN}$, *where* $N = R_\infty^* + D_\infty^*$.

(i) *If* $q = \sigma_1 + \rho_1 + \gamma_3 > \widetilde{\beta}$, *then the solution of system* (3.1)–(3.6) *satisfies* $S(t) \longrightarrow S_\infty > 0$ *as* $t \to \infty$.

(ii) *If* $q < \widetilde{\beta}$, *then* $S(t) \longrightarrow 0$ *as* $t \to \infty$.

Note that the condition $q < \widetilde{\beta}$ in the above theorem is equivalent to $R_0 > 1$ with R_0 as defined in (2.15) and $\alpha_P = 0$. Hence with this theorem we have shown that the asymptotic result for R_0 is global.

Proof. (i) Suppose not; then $\lim_{t\to\infty} S(t) = S^*_\infty = 0$. Consider the limiting system of

$$S' = -\frac{\beta IS}{E+I+S}\frac{c}{1+a(P+R+D)}a \geq 0, \tag{3.12}$$

$$E' = \frac{\beta IS}{E+I+S}\frac{c}{1+a(P+R+D)} - \mu E, \tag{3.13}$$

$$I' = \mu E - (\sigma_1+\rho_1+\gamma_3)I. \tag{3.14}$$

Since $P(t) \longrightarrow 0$, $R(t) \longrightarrow R^*_\infty$, and $D_(t) \longrightarrow D^*_\infty$ as $t \to \infty$, we have the limiting system as follows:

$$S' = -\frac{\beta IS}{E+I+S}\frac{c}{1+aN}, \quad a \geq 0, \tag{3.15}$$

$$E' = \frac{\beta IS}{E+I+S}\frac{c}{1+aN} - \mu E, \tag{3.16}$$

$$I' = \mu E - (\sigma_1+\rho_1+\gamma_3)I. \tag{3.17}$$

From the analysis of system (3.7)–(3.9) with

$$\widetilde{\beta} = \frac{\beta c}{1+aN},$$

if $q > \widetilde{\beta}$, then $S(t) \to S_\infty > 0$. This is a contradiction.

(ii) Assuming $q < \widetilde{\beta}$, we want to show $\lim_{t\to\infty} S(t) = 0$. If not, then $\lim_{t\to\infty} S(t) = S_\infty > 0$. In this case, we have $R_\infty + D_\infty + S_\infty = N$, where $R_\infty = \lim_{t\to\infty} R(t)$, $D_\infty = \lim_{t\to\infty} D(t)$.

Since $S_\infty > 0$, $D_\infty + R_\infty < N$. The limiting system of (3.12)–(3.14) is system (3.15)–(3.17) with D^*_∞, R^*_∞ replaced by D_∞, R_∞. From the analysis of system (3.7)–(3.9) and the assumption $q < \widetilde{\beta}$, we have

$$\widehat{\beta} = \frac{c\beta}{1+a(D_\infty+R_\infty)} > \frac{c\beta}{1+aN} = \widetilde{\beta} > q;$$

hence $S(t) \longrightarrow 0$ as $t \to \infty$. This is a contradiction. □

4. Analysis for model with quarantine. We now give some analytical results on the model with Level A quarantine only. We then have the system

$$S' = -\lambda(S,E,I,Q_A,P,R,D)S, \tag{4.1}$$

$$E' = \lambda(S,E,I,Q_A,P,R,D)S - \mu E - q_1 E, \tag{4.2}$$

$$Q'_A = q_1 E - \gamma_1 Q_A, \tag{4.3}$$

$$I' = \mu E - (\sigma_1+\rho_1+\gamma_3)I, \tag{4.4}$$

$$P' = \gamma_1 Q_A + \gamma_3 I - (\sigma_2+\rho_2)P, \tag{4.5}$$

$$R' = \sigma_1 I + \sigma_2 P, \tag{4.6}$$

$$D' = \rho_1 I + \rho_2 P, \tag{4.7}$$

where the incidence of infection with quarantine rates is given by

$$\lambda(S,E,I,Q_A,P,R,D) = \beta\frac{I+\alpha_A Q_A}{S+E+I+\alpha_A Q_A}\frac{c}{1+a(P+R+D)}. \tag{4.8}$$

Again $S+E+Q_A+I+P+R+D = S_0+I_0 \equiv N$, where $S(0) = S_0 > 0$, $I(0) = I_0 > 0$.

As in Theorem 3.2, we have $S(t) \longrightarrow S_\infty \geq 0$, $R(t) \longrightarrow R_\infty > 0$, $D(t) \longrightarrow D_\infty > 0$ and $I(t) \longrightarrow 0$, $E(t) \longrightarrow 0$, $P(t) \longrightarrow 0$, $Q_A(t) \longrightarrow 0$ as $t \longrightarrow \infty$.

Next we let $\widetilde{q} = \sigma_1 + \rho_1 + \gamma_3$, $c = 1$, and $a = 0$ and consider the limiting system

$$\begin{aligned} S' &= \frac{-\beta(I + \alpha_A Q_A)}{S + E + I + \alpha_A Q_A} S, \\ E' &= \frac{\beta(I + \alpha_A Q_A)}{S + E + I + \alpha_A Q_A} S - \mu E - q_1 E, \\ Q_A' &= q_1 E - \gamma_1 Q_A, \\ I' &= \mu E - \widetilde{q} I. \end{aligned} \tag{4.9}$$

Let $W_1 = \frac{S}{I}$, $W_2 = \frac{E}{I}$, $W_3 = \frac{Q_A}{I}$. Then we have

$$\begin{aligned} W_1' &= -\frac{\beta(1 + \alpha_A W_3)}{1 + W_1 + W_2 + \alpha_A W_3} W_1 - (\mu W_2 - \widetilde{q}) W_1, \\ W_2' &= \frac{\beta(1 + \alpha_A W_3)}{1 + W_1 + W_2 + \alpha_A W_3} W_1 - (\mu + q_1) W_2 - (\mu W_2 - \widetilde{q}) W_2, \\ W_3' &= (q_1 W_2 - \gamma_1 W_3) - (\mu W_2 - \widetilde{q}) W_3. \end{aligned} \tag{4.10}$$

We note that if $\alpha_A = 0$, $q_1 > 0$, then (4.10) is reduced to the two-dimensional system

$$\begin{aligned} W_1' &= \frac{-\beta W_1}{1 + W_1 + W_2} - (\mu W_2 - \widetilde{q}) W_1, \\ W_2' &= \frac{\beta W_1}{1 + W_1 + W_2} - (\mu + q_1) W_2 - (\mu W_2 - \widetilde{q}) W_2. \end{aligned}$$

As in Theorem 3.4, it can be shown that (i) if $\widetilde{\beta} < \widetilde{q}(\frac{\mu + q_1}{\mu})$, then $S(t) \to S_\infty > 0$ as $t \to \infty$ and (ii) if $\widetilde{\beta} > \widetilde{q}(\frac{\mu + q_1}{\mu})$, then $S(t) \to 0$ as $t \to \infty$.

We give the following equilibria and their respective stability analyses:

1. $E_0 = (0, 0, 0)$ is an equilibrium of (4.10). Then the variational matrix at E_0 is

$$M_0 = \begin{bmatrix} -\beta + \widetilde{q} & 0 & 0 \\ \beta & -(\mu + q_1) + \widetilde{q} & 0 \\ 0 & q_1 & -\gamma_1 + \widetilde{q} \end{bmatrix}.$$

We then have the following trivial lemma.

LEMMA 4.1. *$E_0 = (0, 0, 0)$ is locally asymptotically stable if $\beta > \widetilde{q}$, $\mu + q_1 > \widetilde{q}$, $\gamma_1 > \widetilde{q}$.*

2. $E_{23}^* = (0, W_2^*, W_3^*)$, where

$$\begin{aligned} W_2^* &= \frac{\widetilde{q} - (\mu + q_1)}{\mu} > 0 \Longleftrightarrow \widetilde{q} > \mu + q_1, \\ W_3^* &= \frac{q_1 W_2^*}{(\mu W_2^* - \widetilde{q}) + \gamma_1} > 0 \Longleftrightarrow \gamma_1 > \mu + q_1. \end{aligned}$$

Note that $E_{23}^* = (0, W_2^*, W_3^*)$ exists if $\widetilde{q} > \mu + q_1$, $\gamma_1 > \mu + q_1$.

The variational matrix at E_{23}^* is

$$M^* = \begin{bmatrix} \frac{-\beta(1+\alpha_A W_3^*)}{1+W_2^*+\alpha_A W_3^*} - (\mu W_2^* - q) & 0 & 0 \\ \frac{\beta(1+\alpha_A W_3^*)}{1+W_2^*+\alpha_A W_3^*} & -(\mu+q_1) - (2\mu W_2^* - \widetilde{q}) & 0 \\ 0 & q_1 - \mu W_3^* & -\gamma_1 - (\mu W_2^* - \widetilde{q}) \end{bmatrix}.$$

The local stability result is given below, the proof of which is also trivial.

LEMMA 4.2. *E_{23}^* is locally asymptotically stable if $\beta > \frac{1+W_2^*+\alpha_A W_3^*}{1+\alpha_A W_3^*}(\mu+q_1)$.*

3. $E_\infty = (+\infty, \widetilde{W_2}, \widetilde{W_3})$, where $\widetilde{W_2} < \frac{q}{\mu}$. From the first equation of (4.10), we have $\frac{W_1'}{W_1} \le -(\mu W_2 - \widetilde{q})$. If $\lim_{t\to\infty} W_1(t) = \infty$, then we must have $\widetilde{W_2} < \frac{q}{\mu}$. From the second and third equations of (4.10) and $\lim_{t\to\infty} W_1(t) = \infty$, it can be shown that $(\widetilde{W_2}, \widetilde{W_3})$ is the solution of

$$\begin{aligned} \beta(1+\alpha_A W_3) - (\mu W_2 - \widetilde{q})W_2 - (\mu+q_1)W_2 = 0, \\ (q_1 W_2 - \gamma_1 W_3) - (\mu W_2 - \widetilde{q})W_3 = 0. \end{aligned}$$

By the Poincare transform, $Z_1 = \frac{1}{W_1}$, $Z_2 = \frac{W_2}{W_1}$, $Z_3 = \frac{W_3}{W_1}$. Consequently, system (4.10) becomes

$$\begin{aligned} Z_1' &= \beta Z_1 \frac{Z_1 + \alpha_A Z_3}{1 + Z_1 + Z_2 + \alpha_A Z_3} + \mu Z_2 - \widetilde{q} Z_1, \\ Z_2' &= (1+Z_2)\frac{\beta(Z_1 + \alpha_A Z_3)}{1 + Z_1 + Z_2 + \alpha_A Z_3} - (\mu+q_1)Z_2, \\ Z_3' &= q_1 Z_2 - \gamma_1 Z_3 + Z_3 \frac{\beta(Z_1 + \alpha_A Z_3)}{1 + Z_1 + Z_2 + \alpha_A Z_3}. \end{aligned} \tag{4.11}$$

The local stability of E_∞ for system (4.10) is equivalent to the local stability of $\widehat{E_0} = (0,0,0)$ for system (4.11). The variational matrix of $\widehat{E_0}$ for system (4.11) is computed as

$$\begin{bmatrix} -\widetilde{q} & \mu & 0 \\ \beta & -(\mu+q_1) & \beta\alpha_A \\ 0 & q_1 & -\gamma_1 \end{bmatrix}.$$

From the Routh–Hurwitz criterion, the stability conditions can be rewritten as

1. $\beta < A_1$, $A_1 = \frac{\gamma_1\widetilde{q} + (\mu+q_1)(\gamma_1+\widetilde{q})}{\mu + \alpha_A \widetilde{q}}$.
2. $\beta < A_2$, $A_2 = \frac{(\mu+q_1)\gamma_1\widetilde{q}}{\mu\gamma_1 + \alpha_A q_1 \widetilde{q}}$.
3. $\beta < A_3$, $A_3 = \frac{(\mu+q_1)^2(\gamma_1+\widetilde{q}) + \gamma_1\widetilde{q} + (\mu+q_1)(\gamma_1+\widetilde{q})^2}{\mu(\mu+q_1) + \mu\widetilde{q} + \alpha_A q_1((\mu+q_1)+\gamma_1)}$.

We now have the following trivial results.

LEMMA 4.3. *E_∞ is stable if $\beta < A_2$.*

Proof. By routine computation, we have $A_2 < A_1$, $A_2 < A_3$. ∎

LEMMA 4.4. *If $\gamma_1 > \widetilde{q}$, then $A_2 > \widetilde{q}$.*

Proof. Clearly, $A_2 > \widetilde{q} \Longleftrightarrow \gamma_1 > \alpha_A \widetilde{q}$, $0 \le \alpha_A \le 1$. ∎

Thus, when $\gamma_1 > \widetilde{q}$, we have the stability of E_0 and E_∞ diagramed as in Figure 4.1.

We note that if $q_1 = 0$, $\alpha_A = 0$, then $A_2 = \widetilde{q}$. If $\alpha_A = 0$, then $A_2 = \widetilde{q}\frac{\mu+q_1}{\mu}$. If $0 < \alpha_A < 1$, then

$$A_2 = \gamma_1 \widetilde{q} \frac{\mu + q_1}{\mu\gamma_1 + \alpha_A q_1 \widetilde{q}} \to \frac{\gamma_1}{\alpha_A} \quad \text{as } q_1 \longrightarrow \infty.$$

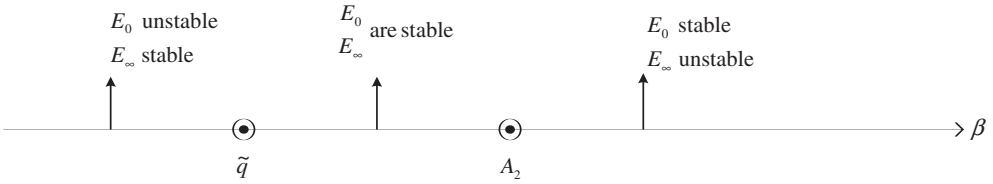


FIG. 4.1. *Stability of E_0 and E_∞ when $\gamma_1 > \widetilde{q}$.*

Thus for the model with quarantine measure, there is a region of bistability. The smaller α_A and larger q_1 give a large region of bistability. According to analysis, if the contact rate $\beta < \widetilde{q}$, we would have only the DFE *no matter if the quarantine measure is implemented or not*. However, if $\beta > A_2$, then we have the endemic case, i.e., $S(t) \longrightarrow 0$ as $t \longrightarrow \infty$.

The following lemma is trivial to prove.

LEMMA 4.5. *If $\mu + q_1 > \widetilde{q}$, $\gamma_1 > \widetilde{q}$, $\widetilde{q} < \beta < A_2$, then both E_0 and E_∞ are locally asymptotically stable. Furthermore, there exists a unique interior equilibrium $E_c = (W_{1c}, W_{2c}, W_{3c})$, $0 < W_{2c} < \frac{\widetilde{q}}{\mu}$.*

Remark. We conjecture that E_c is a saddle point with two-dimensional stable manifold, although we are unable to give a rigorous proof. Instead we will give a full description of the dynamics for the model with quarantine.

Next we consider the case $\widetilde{q} > \mu + q_1$, $\gamma_1 > \mu + q_1$, which guarantees the existence of E_{23}^*. The following inequality is also easy to obtain.

LEMMA 4.6. $A_2 > (\mu + q_1)\frac{1+W_2^*+\alpha_A W_3^*}{1+\alpha_A W_3^*}$.

An illustration of the stability of E_{23} and E_∞ in this case is given in Figure 4.2.

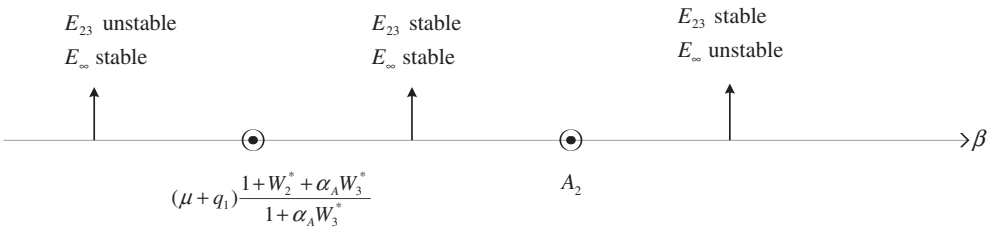


FIG. 4.2. *Stability of E_{23} and E_∞ when $A_2 > (\mu + q_1)\frac{1+W_2^*+\alpha_A W_3^*}{1+\alpha_A W_3^*}$.*

We note that if $\alpha_A = 0$, $q_1 = 0$, then

$$(\mu + q_1)\frac{1 + W_2^* + \alpha_A W_3^*}{1 + \alpha_A W_3^*} = \widetilde{q} \quad \text{and} \quad A_2 = \widetilde{q}.$$

It is also easy to show that

$$(\mu + q_1)\frac{1 + W_2^* + \alpha_A W_3^*}{1 + \alpha_A W_3^*} > \widetilde{q} \Longleftrightarrow \frac{1}{\alpha_A} > \frac{\widetilde{q} - (\mu + q_1)}{\gamma_1 - (\mu + q_1)}.$$

Thus if $\frac{1}{\alpha_A} > \frac{\widetilde{q}-(\mu+q_1)}{\gamma_1-(\mu+q_1)}$ (e.g., if $\gamma_1 > \widetilde{q}$), we have a diagram for the relative sizes of parameters in Figure 4.3.

Thus, for the model with the quarantine measures and the contact rate β, $\widetilde{q} < \beta < (\mu + q_1)(1 + \frac{W_2^*}{1+\alpha_A W_3^*})$ yields the DFE. On the other hand, we would have the

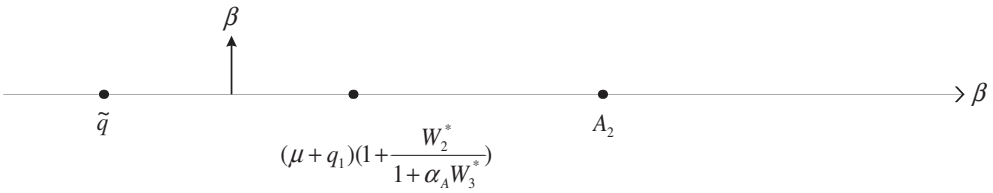


FIG. 4.3. *Diagram for the relative sizes of parameters when* $\frac{1}{\alpha_A} > \frac{\tilde{q}-(\mu+q_1)}{\gamma_1-(\mu+q_1)}$ *(e.g., if* $\gamma_1 > \tilde{q}$*).*

endemic steady state if no quarantine action is taken (i.e., $\beta > \tilde{q} \Longrightarrow$ endemic steady state in the case of no quarantine). The diagram for the case $\frac{1}{\alpha_A} > \frac{\tilde{q}-(\mu+q_1)}{\gamma_1-(\mu+q_1)}$, which is possible if $\gamma_1 < \tilde{q}$, has been deleted for brevity.

For the case $(\mu + q_1)(1 + \frac{W_2^*}{1+\alpha_A W_3^*}) < \beta < \tilde{q}$, E_{23}^* and E_∞ are both stable. Consequently, having quarantine measures implemented might lead to an adverse effect. More precisely, since $\beta < \tilde{q}$, we have the system approaching DFE when there is no quarantine, but with quarantine, we would have the bistable case where the system could approach an endemic steady state, given that the appropriate initial population is in the stability region of the endemic equilibrium.

Remark. In the bistable case, we again conjecture that the interior equilibrium (E_c) exists and (E_c) is a saddle point with two-dimensional stable manifold.

Now we return to the original system (4.1)–(4.7). Consider the limiting system

$$
\begin{aligned}
S' &= -\frac{\beta(I+\alpha_A Q_A)}{S+E+I+\alpha_A Q_A}\frac{cS}{1+a(D_\infty+R_\infty)},\\
E' &= \frac{\beta(I+\alpha_A Q_A)}{S+E+I+\alpha_A Q_A}\frac{cS}{1+a(D_\infty+R_\infty)} - (\mu+q_1)E,\\
Q_A' &= q_1 E - \gamma_1 Q_A,\\
I' &= \mu E - \tilde{q} I.
\end{aligned}
\tag{4.12}
$$

Letting $\widehat{\beta} = \frac{\beta c}{1+a(D_\infty^*+R_\infty^*)}$, we have the following theorem, the proof of which is similar to that of the case without quarantine.

LEMMA 4.7. *If* $W_1(t) \to \infty$ *as* $t \to \infty$*, then we have* $W_2(t) \to \widetilde{W}_2 < \infty$*,* $Z_3 = \frac{W_3}{W_1} \to 0$*, and* $S(t) \to S_\infty > 0$*.*

We now have the main theorem.

THEOREM 4.8. *Let* $\widetilde{\beta} = \frac{\beta c}{1+aN}$*.*

1. *If* $W_1(t) \to \infty$ *as* $t \to \infty$*,* $\widetilde{\beta} < A_2$*, and* E_0, E_{23}^* *are unstable, then* $S(t) \to S_\infty > 0$*.*
2. *If* $W_1(t) \to 0$ *as* $t \to \infty$*,* $\widetilde{\beta} > A_2$*, and one of the two equilibria,* E_0 *or* E_{23}^**, is asymptotically stable, then* $S(t) \to 0$ *as* $t \to \infty$*.*
3. *The bistable case occurs when* $\tilde{q} < \widetilde{\beta} < A_2$*, or* $(\mu+q_1)(1+\frac{W_2^*}{1+\alpha_A W_3^3}) < \widetilde{\beta} < A_2$*.*

Proof.

1. If not, $S(t) \to 0$ as $t \to \infty$, i.e., $S_\infty^* = 0$. Consider the limiting system (4.11) where we have

$$\widehat{\beta} = \frac{\beta c}{1+a(D_\infty^*+R_\infty^*)} = \frac{\beta c}{1+aN} = \widetilde{\beta} < A_2.$$

It follows that $\lim_{t\to\infty} W_1(t) = +\infty$ (assuming the convergence is global) $\Longrightarrow$ $S(t) \to S_\infty > 0$. Hence we have a contradiction.

2. If not, assume $S(t) \to S_\infty^* > 0$. Then $S_\infty^* + D_\infty^* + R_\infty^* = N$. Consequently,

$$\widehat{\beta} = \frac{\beta c}{1 + a(D_\infty^* + R_\infty^*)} > \frac{\beta C}{1 + a(S_\infty^* + D_\infty^* + R_\infty^*)} = \frac{\beta C}{1 + aN} = \widetilde{\beta} > A_2$$

and $\lim_{t\to\infty} S(t) = S_\infty = 0$, again a contradiction. □

Remark. It can be shown that the local stability condition for the effective reproduction number with quarantine $R_Q < 1$ is equivalent to condition 1 of Theorem 4.8, namely, $\widetilde{\beta} = \frac{\beta c}{1+a(S_\infty^* + D_\infty^* + R_\infty^*)} < A_2$.

We further consider the case where $E_0 = (0, 0, 0)$ is unstable and $E_{23}^* = (0, W_2^*, W_3^*)$ does not exist. We note that $E_0 = (0, 0, 0)$ is stable $\iff \beta > \widetilde{q}$, $\mu + q_1 > \widetilde{q}$, $\gamma_1 > \widetilde{q}$ and E_{23}^* exists $\iff \mu + q_1 < \widetilde{q}$, $\gamma_1 > \mu + q_1$.

We consider the case $W_3(t) \to \infty$ as $t \to \infty$. Let $U_1 = W_1$, $U_2 = W_2$, $U_3 = \frac{1}{W_3}$. Then system (4.10) becomes

$$\begin{aligned} U_1' &= \frac{-\beta(U_3 + \alpha_A)}{U_3(1 + U_1 + U_2) + \alpha_A} U_1 - (\mu U_2 - \widetilde{q}) U_1, \\ U_2' &= \frac{\beta(U_3 + \alpha_A)}{U_3(1 + U_1 + U_2) + \alpha_A} U_1 - (\mu + q_1) U_2 - (\mu U_2 - \widetilde{q}) U_2, \\ U_3' &= -q_1 U_2 U_3^2 + \gamma_1 U_3 + U_3(\mu U_2 - \widetilde{q}). \end{aligned} \tag{4.13}$$

We give the equilibria and stability analysis of (4.13):

1. $\widetilde{E_0} = (0, 0, 0)$ always exists.
2. $\widetilde{E_2} = (0, \widetilde{U}_2, 0)$, where $\widetilde{U}_2$ satisfies $-(\mu + q_1)\widetilde{U}_2 - (\mu\widetilde{U}_2 - \widetilde{q})\widetilde{U}_2 = 0$. Therefore $\widetilde{U}_2 = \frac{\widetilde{q} - (\mu + q_1)}{\mu} > 0 \iff \widetilde{q} > \mu + q_1$. Subsequently, $\widetilde{E_2}$ exists if and only if $\widetilde{q} > \mu + q_1$.
3. $\widetilde{E_{12}} = (U_1^*, U_2^*, 0)$.
   From the first equation of (4.13), $U_2^* > 0$ satisfies $-\beta - (\mu U_2^* - \widetilde{q}) = 0$. Therefore $U_2^* = \frac{\widetilde{q} - \beta}{\mu} > 0$.
   From the second equation of (4.13), U_1^* satisfies $\beta U_1^* - (\mu + q_1)U_2^* - (\mu U_2^* - \widetilde{q})U_2^* = 0$. Therefore $U_1^* = \frac{1}{\beta}((\mu + q_1) - \beta)U_2^* > 0$.
   It follows that $\widetilde{E_{12}}$ exists $\iff \widetilde{q} > \beta$, $\mu + q_1 > \beta$.

(i) Stability of $\widetilde{E_0}$. The variational matrix of (4.13) at $\widetilde{E_0}$ is

$$M_0 = \begin{bmatrix} -\beta + \widetilde{q} & 0 & 0 \\ \beta & -(\mu + q_1) + \widetilde{q} & 0 \\ 0 & 0 & \gamma_1 - \widetilde{q} \end{bmatrix}.$$

Thus $\widetilde{E_0}$ is stable $\iff \widetilde{q} < \beta$, $\widetilde{q} < \mu + q_1$, $\gamma_1 < \widetilde{q}$.

(ii) Stability of $\widetilde{E_2}$. The variational matrix of (4.13) at $\widetilde{E_2}$ is

$$\begin{bmatrix} -\beta + (\mu + q_1) & 0 & 0 \\ \beta & -\widetilde{q} + (\mu + q_1) & 0 \\ 0 & 0 & \gamma_1 - (\mu + q_1) \end{bmatrix}.$$

Thus $\widetilde{E_2}$ is stable $\iff \mu + q_1 < \beta$, $\widetilde{q} > \mu + q_1$, $\gamma_1 < \mu + q_1$.

(iii) Stability of $\widetilde{E_{12}}$. The variational matrix of (4.13) at $\widetilde{E_{12}}$ is

$$\begin{bmatrix} 0 & -\mu U_1^* & * \\ \beta & -(\mu + q_1) - 2\mu U_2^* + \widetilde{q} & * \\ 0 & 0 & \gamma_1 + (\mu U_2^* - \widetilde{q}) \end{bmatrix}.$$

The eigenvalue λ satisfies

$$(\lambda - (\gamma_1 + (\mu U_2^* - \widetilde{q})))(\lambda^2 + (\mu + q_1 + 2\mu U_2^* - \widetilde{q})\lambda + \beta\mu U_1^*) = 0, \tag{4.14}$$

$$\gamma_1 + \mu U_2^* - \widetilde{q} = \gamma_1 - \beta \text{ implies } \mu + q_1 + 2\mu U_2^* - \widetilde{q} = \widetilde{q} + (\mu + q_1) - 2\beta.$$

Since $\widetilde{E_{12}}$ exists $\Longrightarrow \widetilde{q} > \beta$, $\mu + q_1 > \beta$, we have $\mu + q_1 + 2\mu U_2^* - \widetilde{q} > 0$ and $\widetilde{E_{12}}$ is stable $\Longleftrightarrow \gamma_1 < \beta$, $\widetilde{q} > \beta$, $\mu + q_1 > \beta$.

Now, when we assume that $E_0 = (0, 0, 0)$ is unstable and E_{23}^* does not exist, there are three cases:

Case 1. $\widetilde{q} > \mu + q_1 > \gamma_1$, and we have three subcases:

Subcase 1. $A_2 > \widetilde{q} \Longleftrightarrow \gamma_1 > \alpha_A \widetilde{q}$.

Subcase 2. $A_2 > \gamma_1 \Longleftrightarrow (\widetilde{q} - \gamma_1) + \widetilde{q} q_1 > \alpha_A q_1 \widetilde{q}$.

Subcase 3. $A_2 > \mu + q_1 \Longleftrightarrow \gamma_1(\widetilde{q} - \mu) > \alpha_A q_1 \widetilde{q}$.

Figure 4.4 illustrates the possibilities for the stability of $\widetilde{E_{12}}$ and E_∞ in Case 1.

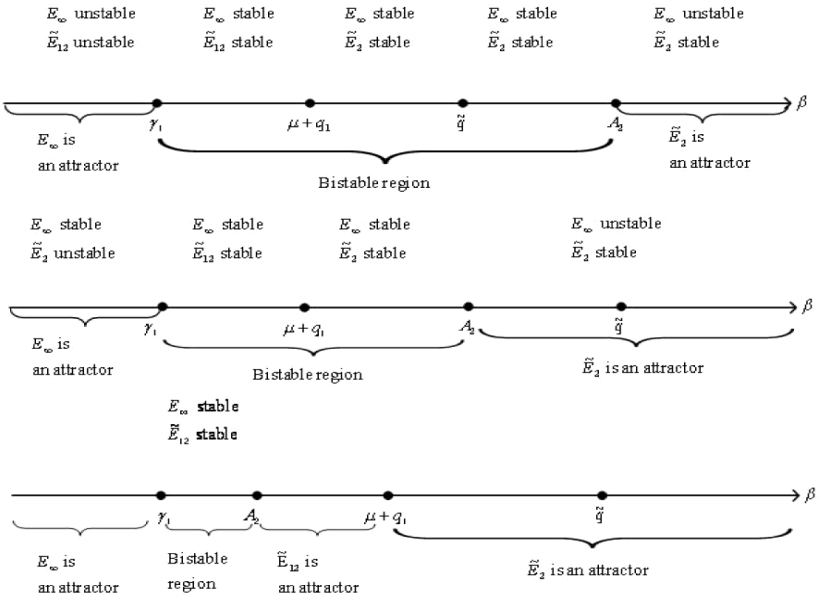


FIG. 4.4. *Stability of* $\widetilde{E_{12}}$ *and* E_∞ *when* $\widetilde{q} > \mu + q_1 > \gamma_1$.

Case 2. $\widetilde{q} < \mu + q_1$, $\gamma_1 < \mu + q_1$, and we have two subcases:

Subcase (i). $\gamma_1 < \widetilde{q}$. The stability of $\widetilde{E_0}$, $\widetilde{E_{12}}$, and E_∞ is as follows.

- $\gamma_1 < q < A_2 < \mu + q_1$.
  If $\beta < \gamma_1$, then E_∞ is an attractor; if $\gamma_1 < \beta < \widetilde{q}$, then $E_\infty, \widetilde{E}_{12}$ are stable; if $\widetilde{q} < \beta < A_2$, then $E_\infty, \widetilde{E}_0$ are stable; if $\beta > A_2$, then $\widetilde{E}_0$ is an attractor.
- $\gamma_1 < \widetilde{q} < \mu + q_1 < A_2$.
  Same as above.
- $\gamma_1 < A_2 < \widetilde{q} < \mu + q_1$.
  If $\beta < \gamma_1$, then E_∞ is an attractor; if $\gamma_1 < \beta < A_2$, then $\widetilde{E}_{12}, E_\infty$ are stable; if $A_2 < \beta < \widetilde{q}$, then $\widetilde{E}_{12}$ is an attractor; if $\beta > \widetilde{q}$, then $\widetilde{E}_0$ is an attractor.

Subcase (ii). $\gamma_1 > \widetilde{q}$. Then $A_2 > \widetilde{q}$. If $\beta < \widetilde{q}$, then E_∞ is an attractor. If $\widetilde{q} < \beta < A_2$, then $\widetilde{E}_0$ and E_∞ are stable. If $\beta > A_2$, then $\widetilde{E}_0$ is an attractor.

Case 3. $\widetilde{q} < \mu + q_1 < \gamma_1$, and $\gamma_1 > \widetilde{q}$. Then $A_2 > \widetilde{q}$, and we have a result similar to that in Subcase (ii).

5. Concluding remarks. For the model without quarantine but with behavior change due to public response to the severity of the disease, we have shown that the local stability condition in Theorem 3.4 is equivalent to the condition that the basic reproduction number R_0 given in (2.15) with $\alpha_P = 0$ is less than 1. If $\beta c > \sigma_1 + \rho_1 + \gamma_3$, the epidemic would persist without public response; however, if the magnitude of public response, as measured by the parameter a, is sufficiently large so that $R_0 < 1$, the reduction of infections through public response will be large enough to drive the epidemic down to a disease-free state.

For the model with both quarantine and behavior change, the dynamics is much more complicated. The effective reproduction number with quarantine R_Q in (2.16) gives local stability of DFE when $R_Q < 1$. However, there are ranges of the parameters which would lead to bistable steady states, i.e., one locally stable DFE and another locally stable endemic equilibrium. In such cases, we conjecture that there is a saddle point with two-dimensional stable manifold. As an illustration, we give the following numerical example.

We let $\alpha_A = 0.1$, $\widetilde{\beta} = 0.5$, $\widetilde{q} = 0.3$, $\gamma_1 = 0.4$, $q_1 = 0.2$, $\mu = 0.2$, and use the initial values of $S(0) = 1$, $E(0) = 0$, $Q_A(0) = 0$, $I(0) = 1$, i.e., one infective case entering a totally susceptible population of one individual so that $(W_1(0), W_2(0), W_3(0)) = (1, 0, 0)$. The result is given in Figure 5.1, where the system goes to the endemic equilibrium.

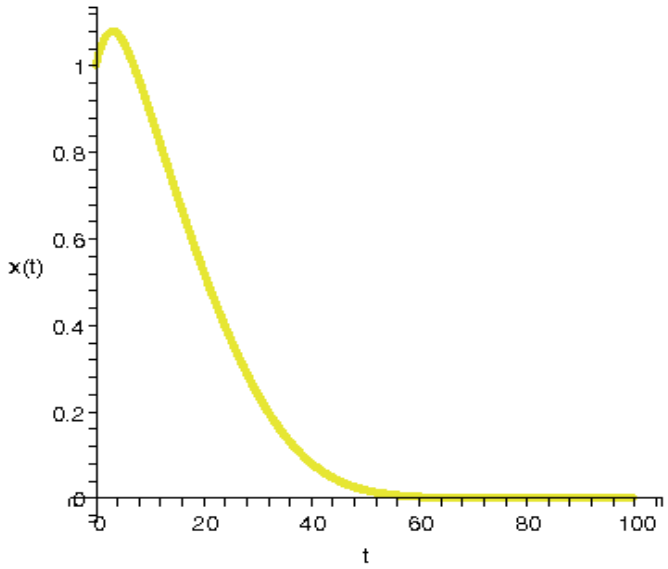


FIG. 5.1. *Numerical example with* $\alpha_A = 0.1$, $\widetilde{\beta} = 0.5$, $\widetilde{q} = 0.3$, $\gamma_1 = 0.4$, $q_1 = 0.2$, $\mu = 0.2$, *and initial population* $S(0) = 1$, $E(0) = 0$, $Q_A(0) = 0$, $I(0) = 1$, *where system approaches endemic equilibrium.* $X(t)$ *is* $W_1(t) = S(t)/I(t)$, *which goes to zero.*

However, if we let the initial values be $S(0) = 10$, $E(0) = 0$, $Q_A(0) = 0$, $I(0) = 1$, i.e., 10 infective cases entering a totally susceptible population of 100 individuals, so that $(W_1(0), W_2(0), W_3(0)) = (10, 0, 0)$, the system will approach DFE as shown in

Figures 5.2 and 5.3. Note that for this set of parameters, $A_2 = 0.56 > \widetilde{\beta} = 0.5 > \widetilde{q} = 0.3$, $\mu + q_1 = 0.2 + 0.2 > \widetilde{q} = 0.3$, $\gamma_1 = 0.4 > \widetilde{q} = 0.3$. Moreover, Case 3 of Theorem 4.8 holds for this data. Therefore we have bistability where both $E_0 = (0, 0, 0)$ and E_∞ are locally stable.

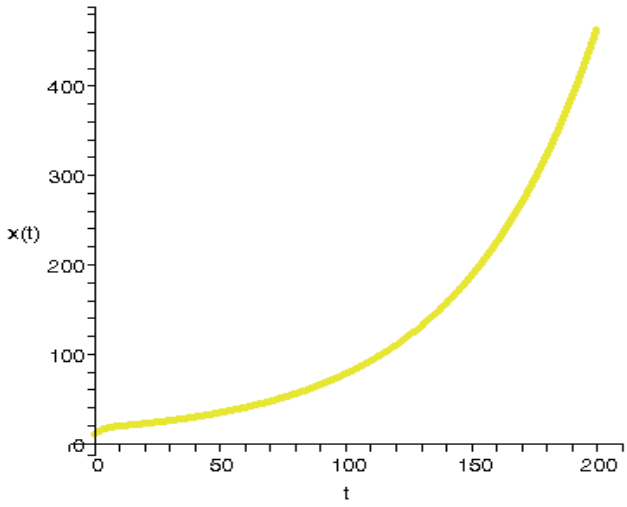


FIG. 5.2. *Numerical example with* $\alpha_A = 0.1$, $\widetilde{\beta} = 0.5$, $\widetilde{q} = 0.3$, $\gamma_1 = 0.4$, $q_1 = 0.2$, $\mu = 0.2$, *and initial population* $S(0) = 10$, $E(0) = 0$, $Q_A(0) = 0$, $I(0) = 1$, *where the system approaches DFE.* $X(t)$ *is* $W_1(t) = S(t)/I(t)$, *which goes to a nonzero equilibrium, and* $S(t) \to 14.58322$.

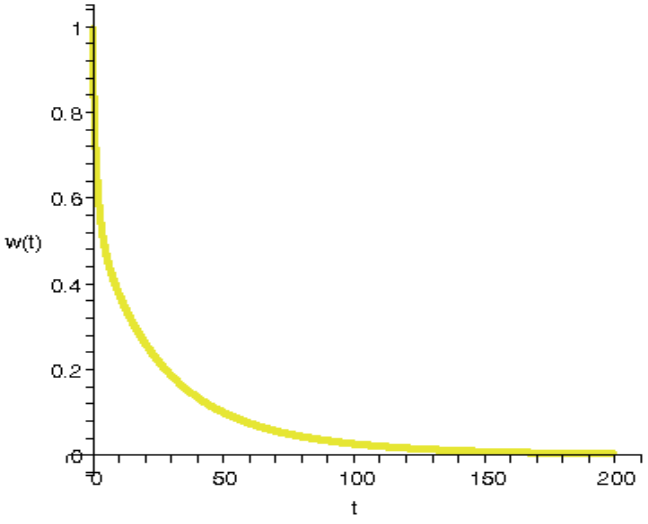


FIG. 5.3. *Numerical example with* $\alpha_A = 0.1$, $\widetilde{\beta} = 0.5$, $\widetilde{q} = 0.3$, $\gamma_1 = 0.4$, $q_1 = 0.2$, $\mu = 0.2$, *and initial population* $S(0) = 10$, $E(0) = 0$, $Q_A(0) = 0$, $I(0) = 1$, *where the system approaches DFE.* $W(t)$ *is* $I(t)$, *which goes to zero.*

The epidemiological interpretation is most interesting. In an epidemic where there is public response to the increasing severity of the epidemic to cut down infections

through individual behavior change, sufficient decrease of the infection rate through the parameter a would be enough to contain the epidemic, *regardless of the initial population sizes at the onset of outbreak.*

Suppose that, in addition to the severity-dependent public response, quarantine is implemented. If the adherence to quarantine to decrease contact rate c and transmission probability β is satisfactory, the effective infection rate $\widetilde{\beta} = \frac{\beta c}{1+aN}$ is sufficiently lowered. Case 1 in Theorem 4.8 will be satisfied along with $\widetilde{\beta} < \widetilde{q}$, and the epidemic can be successfully contained, again regardless of the initial population sizes at the onset of outbreak. If, on the other hand, $\widetilde{\beta}$ is not sufficiently lowered and we have Case 2 of Theorem 4.8, the epidemic will persist and the susceptible population $S(t)$ will be depleted eventually. There is the third scenario, where $\widetilde{\beta}$ is decreased but not sufficiently so, to less than A_2 but greater than $\widetilde{q}$. Then the system could approach either the DFE or the endemic steady state, depending on the initial values of the system.

In the worst case scenario, if the quarantine were not adhered to faithfully, the false sense of security brought about by the quarantine that all infective persons were in quarantine could lead to increased $\widetilde{\beta}$. Thus the quarantine might in fact have an adverse effect by contributing to the persistence of the epidemic. For illustration, we consider the hypothetical case where $\beta < \widetilde{q} = \sigma_1 + \rho_1 + \gamma_3$. Since $\beta > \widetilde{\beta}$ when $c = 1$, from Theorem 3.4 we know the system goes to DFE with no quarantine implemented. However, for Case 1 (Figure 4.4), Subcase 3, when $\mu + q_1 < \beta < \widetilde{q}$ the system will converge to the endemic equilibrium. This demonstrates the (distinct) possibility that, under appropriate parameter values, a quarantine program which is not sufficiently comprehensive ($q_1 < \beta - \mu$) could have the adverse effect of causing a system which would have approached DFE without quarantine to converge to the endemic equilibrium instead.

We should note, however, that $\widetilde{\beta} = \frac{\beta c}{1+aN}$, where N is the constant total population size, which numbers in at least millions whether in Hong Kong, Singapore, Taipei, Beijing, or Toronto. In [27], it was determined that $a = 0.0013$ and $c\beta = 0.429$ for Taiwan's SARS outbreak. Subsequently the real value for $\widetilde{\beta}$ is of the order $0.429/(1+0.0013*10^6) \ll 1$, making it most unlikely for either Case 2 (asymptotically stable endemic equilibrium) or Case 3 (bistability) in Theorem 4.8 to prevail. In other words, the modeling results indicate that for an infectious disease with infectivity and patterns of transmission typical of SARS, the outbreak can always be eradicated by implementing border control of imported cases and limited quarantine, along with the public's social response to avoid infections.

It is also interesting to note that if $\alpha_A = 0$, the stability condition for E_∞ becomes $\beta < \widetilde{q}\mu/(\mu + q_1)$. Hence quarantine is always beneficial and an effective Level A quarantine is always helpful in containing the epidemic. However, if for some disease unlike SARS in its ability to infect during the asymptomatic stage, some fraction of the quarantined population is not fully isolated and can still infect others (i.e., $\alpha_A > 0$), then quarantine might also affect the outbreak adversely. A numerical example of this scenario is as follows: Let $\alpha_A = 0.1$, $\beta = 0.5$, $\mu = 0.5$, $\widetilde{q} = 0.7$, $q_1 = 0.1$, $\gamma_1 = 0.01$, $A_2 = 0.35$. Here $\widetilde{E}_{12} = (0.08, 0.4, +\infty)$ is a global attractor as in Case 1, Subcase 3 in Figure 4.4. However, if there is no quarantine (i.e., $q_1 = 0$ and hence $\alpha_A = 0$), DFE is the global attractor. Additional examples of this type can be observed in Case 1, Subcases 2 and 3 (Figure 4.4), where $\widetilde{E}_2$ can become the global attractor for the appropriate parameter range of β, as well as in Case 2, Subcase (i), where again it is possible for $\widetilde{E}_{12}$ to become a global attractor. Note that

a condition for these cases to emerge is $\widetilde{q} > A_2$ or, equivalently, $\gamma_1/\widetilde{q} < \alpha_A$. Hence if there is a nonzero reduction in the infection rate of the quarantined class α_A which is larger than the ratio of the progression rate of the quarantined persons γ_1 to the removal rate of the unquarantined infectives $\widetilde{q}$, an adverse effect could take place with implementation of quarantine. To keep this possibility from occurring, one would need either (i) significant reduction of infection by the quarantined individuals (small α_A) or (ii) quick isolation of quarantined persons at onset (large γ_1) compared to the removal of the infective class (small $\widetilde{q}$). Similar possible adverse effects of intervention measures have also been observed in other theoretical models of infectious diseases (e.g., [30, 31]).

Going back to the quarantine for SARS, we assume that $\alpha_P = \alpha_A = 0$. If all other pertinent parameters remain the same, we have $R_Q = R_0\mu/(\mu+q_1)$ from (2.15)–(2.16). That is, the implementation of quarantine would give the mean reproduction number of an infective individual a factor of $\mu/(\mu + q_1)$, where q_1 is the effective quarantine rate. That is, through quarantine alone, the mean reproduction number of an infective individual is reduced by a factor of $1 - \mu/(\mu + q_1)$.

In a data-based modeling study where the pertinent parameters were estimated from the Taiwan SARS data [27], the quarantine rate q_1 was estimated to be 0.0277. It was reported by Donnelly et al. [32] that the maximum likelihood estimate for the mean time from exposure to onset of symptoms is 6.37 days. Hence the mean progression rate from exposure to onset is approximately $\mu = 1/6.37 = 0.157$. Making use of the two estimates, we conclude that *if all other parameters remain unchanged*, the quarantine in Taiwan would result in a reduction of 15% ($\mu/(\mu+q_1) = 0.850$) in the mean reproduction number by an infective individual. Given that current studies of SARS indicate that the basic reproduction numbers R_0 in all the SARS-affected areas were greater than 2 at the beginning of the outbreak in 2003, one can conclude that quarantine alone would not have been able to contain the epidemic (i.e., reduce R_0 to less than 1) in Taiwan. For a given affected area with a basic reproduction number R_0, we need to have an effective quarantine rate of $q_1 > q_1^* = 0.157(R_0 - 1)$ for R_Q to be less than 1. Using the estimated values of R_0 for Hong Kong, Toronto, and Taiwan in current literature, we give in Table 5.1 the effective quarantine rate q_1^* needed in the affected areas to reduce the reproduction number to less than 1, if all else remains the same. Note that the estimate for Taiwan [13] assumes that a symptomatic patient is infective from onset to classification as a probable case followed by isolation. If we assume the patient is not infective during the first two days of onset as suggested by some studies (see [2]), the reproduction number is reduced to 3.56 and subsequently $q_1^* = 0.402$.

TABLE 5.1

Affected area	Reproduction number [literature cited]	Effective quarantine rate q_1^* needed to contain outbreak
Hong Kong	2.7 [32]	0.267
	3 [10]	0.314
Toronto	3.3 [33]	0.361
Taiwan	4.23 [13]	0.507

Since the SARS-CoV virus does not appear to be infective before onset of symptoms [2], quarantine does not directly prevent infections by the exposed individuals during the quarantine period. However, studies on SARS quarantine data in Taiwan [8] indicate that quarantined persons are significantly more quickly diagnosed

and hospitalized as compared to the unquarantined individuals. Hence the effectiveness of quarantine for infectious diseases like SARS, for which no infection is being prevented during the quarantine period, can only be indirect and therefore must be combined with other intervention measures in order to fully contain the outbreaks.

Acknowledgments. The authors are grateful for constructive discussions with Roy Anderson, John Glasser, Chwan-Chuan King, and Fred Brauer which helped formulate some of the ideas for this work. Ying-Hen Hsieh would like to thank MITACS (Canada) for their generous financial support while attending the MITACS SARS meetings in Banff, Canada, where several of the abovementioned discussions took place.

REFERENCES

[1] World Health Organization, *Summary of Probable SARS Cases with Onset of Illness from* 1 *November* 2002 *to* 31 *July* 2003, http://www.who.int/csr/sars/country/table2003_09_23/en/ (26 September 2003).

[2] World Health Organization, *Consensus Document on the Epidemiology of Severe Acute Respiratory Syndrome (SARS)*, http://www.who.int/csr/sars/en/WHOconsensus.pdf (17 October 2003).

[3] M. Enserink, *SARS: A pandemic prevented*, Science, 302 (2003), p. 2045.

[4] J. Ou, Q. Li, and G. Zeng, *Efficiency of quarantine during an epidemic of severe acute respiratory syndrome—Beijing, China,* 2003, MMWR, 52 (2003), pp. 1037–1040.

[5] X. Pang, Z. Zhu, F. Xu, J. Guo, X. Gong, et al., *Evaluation of control measures implemented in the severe acute respiratory syndrome outbreak in Beijing,* 2003, JAMA, 290 (2003), pp. 3215–3221.

[6] B. Diamond, *SARS spreads new outlook on quarantine models*, Nat. Med., 9 (2003), p. 1441.

[7] M. L. Lee, C. J. Chen, I. J. Su, K. T. Chen, C. C. Yeh, C. C. King, et al., *Use of quarantine to prevent transmission of severe acute respiratory syndrome—Taiwan* 2003, MMWR, 52 (2003), pp. 680–683.

[8] Y. H. Hsieh, C. C. King, M. S. Ho, C. W. S. Chen, J. Y. Lee, F. C. Liu, Y. C. Wu, and J. S. J. Wu, *Quarantine for SARS, Taiwan*, Emerg. Infect. Dis., 11 (2005), pp. 278–282.

[9] L. O. Gostin, R. Bayer, and A. L. Fairchild, *Ethical and legal challenges posed by severe acute respiratory syndrome: Implications for the control of severe infectious disease threats*, JMAM, 290 (2003), pp. 3229–3237.

[10] M. Lipsitch et al., *Transmission dynamics and control of severe acute respiratory syndrome*, Science, 300 (2003), pp. 1966–1970.

[11] S. Riley et al., *Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions*, Science, 300 (2003), pp. 1961–1966.

[12] J. O. Lloyd-Smith, A. P. Galvani, and W. M. Getz, *Curtailing transmission of severe acute respiratory syndrome within a community and its hospital*, Proc. R. Soc. Lond. B Biol. Sci., 270 (2003), pp. 1979–1989.

[13] Y. H. Hsieh, C. W. S. Chen, and S. B. Hsu, *SARS outbreak, Taiwan* 2003, Emerg. Infect. Dis., 10 (2004), pp. 201–206.

[14] Y. H. Hsieh, J. Y. Lee, and H. L. Chang, *On SARS epidemiology, cumulative case curve, and logistic-type model: Ascertaining effectiveness of intervention and predicting case number*, Emerg. Infect. Dis., 10 (2004), pp. 1165–1167.

[15] C. A. Donnelly, M. C. Fisher, C. Fraser, A. C. Ghani, S. Riley, N. M. Ferguson, and R. M. Anderson, *Epidemiological and genetic analysis of severe acute respiratory syndrome*, Lancet Infect. Dis., 4 (2004), pp. 672–683.

[16] E. H. Kaplan, D. L. Craft, and L. M. Wein, *Analyzing bioterror response logistics: The case of smallpox*, Math. Biosci., 185 (2003), pp. 33–72.

[17] L. Sattenspiel and D. A. Herring, *Simulating the effect of quarantine on the spread of the* 1918-19 *flu in central Canada*, Bull. Math. Biol., 65 (2003), pp. 1–26.

[18] H. Hethcote, Z. Ma, and S. Liao, *Effects of quarantine in six endemic models for infectious diseases*, Math. Biosci., 180 (2002), pp. 141–160.

[19] E. H. Kaplan, D. L. Craft, and L. M. Wein, *Emergency response to a smallpox attack: The case for mass vaccination*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 10935–10940.

[20] M. I. Meltzer, I. Damon, J. W. LeDuc, and J. D. Millar, *Modeling potential responses to smallpox as a bioterrorist weapon*, Emerg. Infect. Dis., 7 (2001), pp. 959–969.

[21] M. J. Keeling and C. A. Gilligan, *Metapopulation dynamics of bubonic plague*, Nature, 407 (2000), pp. 903–906.
[22] P. Rohani, C. J. Green, N. B. Mantilla-Beniers, and B. T. Grenfell, *Ecological interference between fatal diseases*, Nature, 422 (2003), pp. 885–888.
[23] D. Clancy, *Optimal intervention for epidemic models with general infection and removal rate functions*, J. Math. Biol., 39 (1999), pp. 309–331.
[24] C. S. Holling, *The functional response of predators to prey density and its role in mimicry and population regulation*, Mem. Entolmol. Soc. Can., 45 (1965), pp. 1–60.
[25] F. Brauer and P. van den Driessche, *Models for transmssion of disease with immigration of infectives*, Math. Biosci., 171 (2001), pp. 143–154.
[26] P. van den Driessche and J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci., 180 (2002), pp. 29–48.
[27] Y. H. Hsieh, C. C. King, C. W. S. Chen, M. S. Ho, S. B. Hsu, and Y. C. Wu, *On the Impact of Intervention Measures and Public Response for Severe Acute Respiratory Syndrome Outbreak*, preprint, National Chung Hsing University, Taichung, Taiwan.
[28] M. S. Ho and I. J. Su, *Preparing to prevent severe acute respiratory syndrome and other respiratory infections*, Lancet Infect. Dis., 4 (2004), pp. 684–689.
[29] W. A. Coppell, *Stability and Asymptotic Behavior of Solutions of Differential Equations*, Heath, Boston, 1965.
[30] R. M. Anderson, S. Gupta, and R. M. May, *Potential of community-wide chemotherapy or immunotherapy to control the spread of HIV*-1, Nature, 350 (1991), pp. 356–359.
[31] Y. H. Hsieh and J. Velaso-Hernanadez, *Community treatment of HIV*-1*: Initial and asymptotic dynamics*, BioSystems, 35 (1995), pp. 75–81.
[32] C. A. Donnelly, A. C. Ghani, G. M. Leung, A. J. Hedley, C. Fraser, S. Riley, L. J. Abu-Raddad, L. M. Ho, T. Q. Thach, P. Chau, K. P. Chan, T. H. Lam, L. Y. Tse, T. Tsang, S. H. Liu, J. H. Kong, E. M. Lau, N. M. Ferguson, and R. M. Anderson, *Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong*, Lancet, 361 (2003), pp. 1761–1766.
[33] J. Wallinga, *Presentation at the Global Meeting on the Epidemiology of SARS*, World Health Organization, Geneva, Switzerland, 2003.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 648–679

IRREVERSIBLE PASSIVE ENERGY TRANSFER IN COUPLED OSCILLATORS WITH ESSENTIAL NONLINEARITY*

GAETAN KERSCHEN†, YOUNG SUP LEE‡, ALEXANDER F. VAKAKIS§, D. MICHAEL McFARLAND¶, AND LAWRENCE A. BERGMAN¶

Abstract. We study numerically and analytically the dynamics of passive energy transfer from a damped linear oscillator to an essentially nonlinear end attachment. This transfer is caused by either fundamental or subharmonic resonance capture, and in some cases is initiated by nonlinear beat phenomena. It is shown that, due to the essential nonlinearity, the end attachment is capable of passively absorbing broadband energy at both high and low frequencies, acting, in essence, as a passive broadband boundary controller. Complicated transitions in the damped dynamics can be interpreted based on the topological structure and bifurcations of the periodic solutions of the underlying undamped system. Moreover, complex resonance capture cascades are numerically encountered when we increase the number of degrees of freedom of the system. The ungrounded essentially nonlinear end attachment discussed in this work can find application in numerous practical settings, including vibration and shock isolation of structures, seismic isolation, flutter suppression, and packaging.

Key words. resonance capture, passive energy transfer, essential nonlinearity, nonlinear energy sinks, energy pumping

AMS subject classifications. 74H10, 74H15, 74H45

DOI. 10.1137/040613706

1. Introduction.

We study passive and irreversible energy transfer from a linear oscillator to an essentially nonlinear attachment, which, in essence, acts as a *nonlinear energy sink (NES)*; such energy transfer we refer to as *nonlinear energy pumping*. In previous works (Vakakis and Gendelman (2001), Vakakis et al. (2003)) grounded and relatively heavy nonlinear attachments were considered, a feature that limits their attractiveness in practical applications. To eliminate these restrictions, an ungrounded and light nonlinear attachment is considered in this work, which, in addition, possesses the feature of modularity. As shown in Lee et al. (2005), even though the system considered has a simple configuration, it possesses a very complicated structure of undamped periodic orbits, which, in turn, give rise to a complicated

*Received by the editors August 20, 2004; accepted for publication July 12, 2005; published electronically January 6, 2006. This work was funded in part by AFOSR Contract 00-AF-B/V-0813.
http://www.siam.org/journals/siap/66-2/61370.html

†Département d'Aérospatiale, Mécanique et Matériaux (ASMA), Université de Liège, B-4000 Liège, Belgium (g.kerschen@ulg.ac.be); National Technical University of Athens, P.O. Box 64042, GR-157 10 Zografos, Athens, Greece; and University of Illinois at Urbana-Champaign, 104 S. Wright St., Urbana, IL 61801. The work of this author was partially supported by grants from the Belgian National Fund for Scientific Research—FNRS, the Belgian Rotary District 1630, and the Fulbright and Duesberg Foundations, which made his visit to the National Technical University of Athens and the University of Illinois possible.

‡Department of Mechanical and Industrial Engineering, University of Illinois at Urbana-Champaign, 104 S. Wright St., Urbana, IL 61801 (yslee4@uiuc.edu).

§Corresponding author. Department of Applied Mathematical and Physical Sciences, National Technical University of Athens, P.O. Box 64042, GR-157 10 Zografos, Athens, Greece (vakakis@central.ntua.gr), and Departments of Mechanical and Industrial Engineering and of Aerospace Engineering, University of Illinois at Urbana-Champaign, 104 S. Wright St., Urbana, IL 61801 (avakakis@uiuc.edu). The work of this author was partially supported by the research grant HRAKLEITOS awarded by the Hellenic Ministry of Development (program EPEAEK II).

¶Department of Aerospace Engineering, University of Illinois at Urbana-Champaign, 104 S. Wright St., Urbana, IL 61801 (dmmcf@uiuc.edu, lbergman@uiuc.edu).

series of transitions and energy exchange phenomena in the damped dynamics. We aim to show that in this system there are at least three different mechanisms for energy pumping, based either on fundamental and subharmonic resonance captures or on nonlinear beat phenomena.

Previous works examined targeted energy transfer in systems of coupled nonlinear oscillators through energy exchanges between donor and acceptor discrete breathers due to nonlinear resonance (Kopidakis, Aubry, and Tsironis (2001), Aubry et al. (2001), Morgante et al. (2002)). In Vainchtein et al. (2004) resonant interactions between monochromatic electromagnetic waves and charged particles were studied, leading to chaotization of particles and transport in phase space. In Khusnutdinova and Pelinovsky (2003) the processes governing energy exchange between coupled Klein–Gordon oscillators were analyzed; the same weakly coupled system was studied in Maniadis, Kopidakis, and Aubry (2004), and it was shown that, under appropriate tuning, total energy transfer can be achieved for coupling above a critical threshold. In related work, localization of modes in a periodic chain with a local nonlinear disorder was analyzed (Cai, Chan, and Cheung (2000)); transfer of energy between widely spaced modes in harmonically forced beams was analytically and experimentally studied (Malatkar and Nayfeh (2003)); and a nonlinear dynamic absorber designed for a nonlinear primary was analyzed (Zhu, Zheng, and Fu (2004)).

In this work we consider the two-degree-of-freedom (DOF) system

$$\begin{aligned} &m_1\ddot{y} + k_1 y + c_1\dot{y} + c_2(\dot{y}-\dot{v}) + k_2(y-v)^3 = P(t) \\ &\qquad\Rightarrow\quad \ddot{y} + \omega_0^2 y + \lambda_1\dot{y} + \lambda_2(\dot{y}-\dot{v}) + C(y-v)^3 = F(t), \\ &m_2\ddot{v} + c_2(\dot{v}-\dot{y}) + k_2(v-y)^3 = 0 \quad\Rightarrow\quad \varepsilon\ddot{v} + \lambda_2(\dot{v}-\dot{y}) + C(v-y)^3 = 0, \end{aligned} \tag{1}$$

where $\omega_0^2 = k_1/m_1$, $C = k_2/m_1$, $\varepsilon = m_2/m_1$, $\lambda_1 = c_1/m_1$, $\lambda_2 = c_2/m_1$, and $F(t) = P(t)/m_1$. Our basic aim is to study the dynamics of irreversible energy transfer ("energy pumping") from the linear oscillator (which will be directly excited) to the nonlinear attachment (which will be assumed to be initially at rest). We show that there are at least three dynamic mechanisms that can initiate or cause such energy transfer in the damped system, and these can be studied and understood by first considering the dynamics of the underlying undamped system.

2. Review of the dynamics of the undamped system (Lee et al., 2005). Since the structure and bifurcations of the periodic orbits of the undamped and unforced system play an essential role in energy transfer phenomena in the damped and forced system, we start with a brief review of the dynamics of system (1) with $\lambda_1 = \lambda_2 = F(t) = 0$; for a more detailed discussion we refer to Lee et al. (2005).

In Figure 1 we present the various branches of periodic solutions in a frequency-energy plot. A periodic orbit is represented by a point in the plot, and a branch, represented by a solid line, is a collection of periodic orbits possessing the same qualitative features. For instance, the branch $S11+$ gathers all the periodic orbits for which the linear and nonlinear oscillators vibrate with the same frequency and in an in-phase fashion. There are two general classes of solutions: *symmetric solutions* $Snm\pm$ correspond to orbits that satisfy the initial conditions $\dot{v}(0) = \pm\dot{v}(T/2)$ and $\dot{y}(0) = \pm\dot{y}(T/2)$, where T is the period, n is the number of half-waves in v, and m the number of half-waves in y in a half-period interval; *unsymmetric solutions* Unm are orbits that fail to satisfy the initial conditions of the symmetric orbits, with the same notation for the two indices. We adopt the following convention regarding the placement of the various branches in the frequency domain: we assign to a specific branch

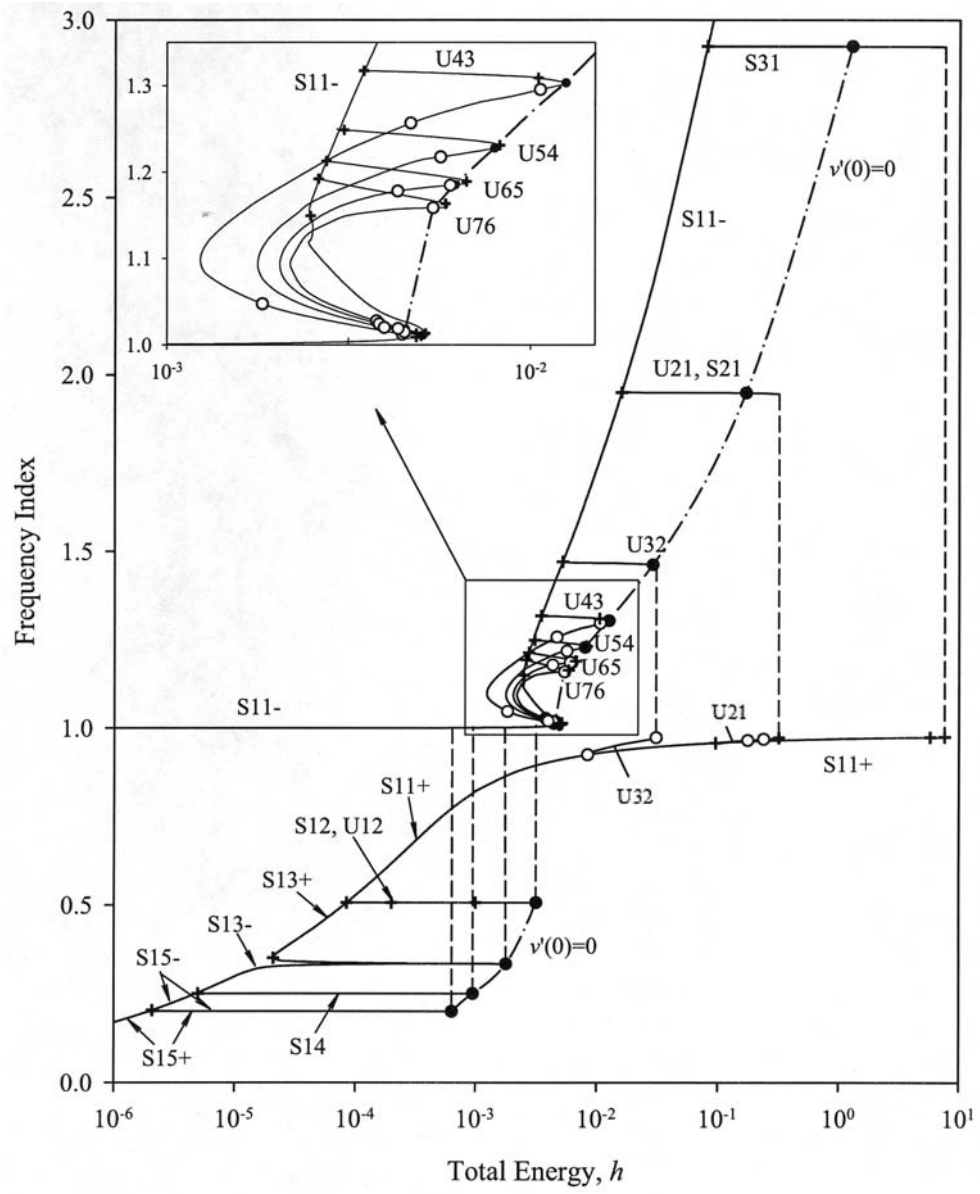


FIG. 1. *Frequency-energy plot of the periodic orbits: for the sake of clarity, no stability is indicated; special orbits are denoted by bullets* (•) *and are connected by dashed-dot lines; other symbols indicate bifurcation points (stability-instability boundaries):* (+) *four Floquet multipliers at* +1*, and* (◯) *two Floquet multipliers at* +1 *and two at* −1 *(see Lee et al.* (2005)*).*

of solutions a frequency index equal to the ratio of its indices; e.g., $S21\pm$ is represented by the frequency index $\omega = 2/1 = 2$, as is $U21$; $S13\pm$ is represented by $\omega = 1/3$; etc. This convention rule holds for every branch except $S11\pm$, which, however, are particular branches forming the basic backbone of the entire plot. On the energy axis we depict the (conserved) total energy of the system when it oscillates in the corresponding periodic motion. Transitions between certain branches seem to involve

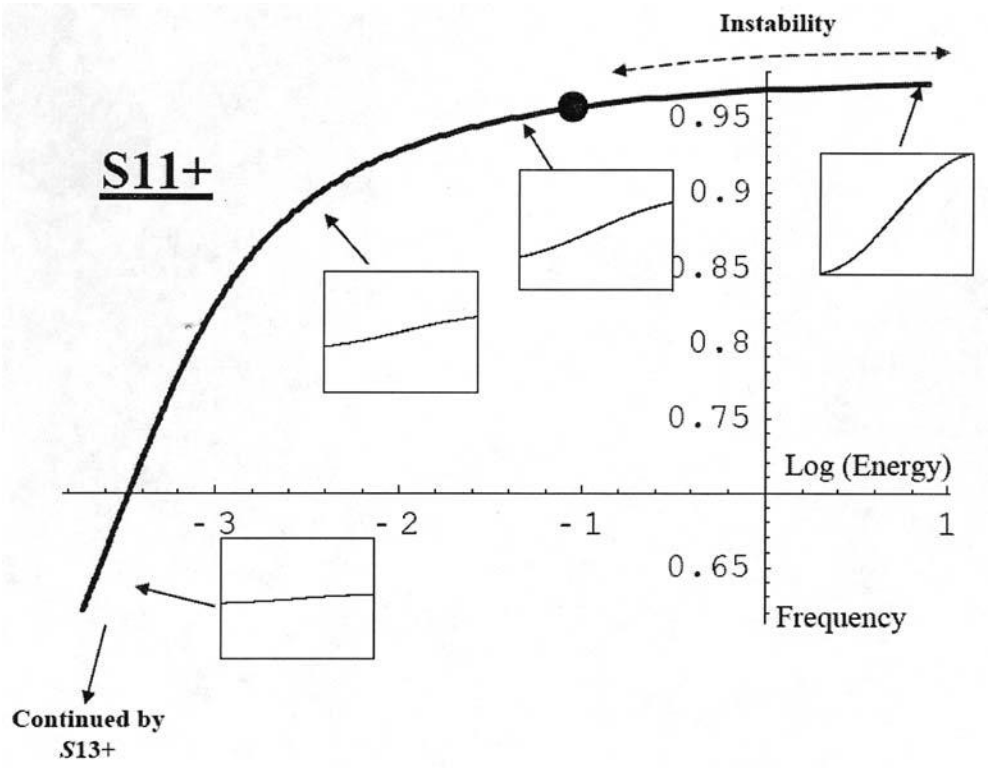


FIG. 2. *Detailed plot of branch* $S11+$ *in the frequency index-logarithm of energy plane. A dot* (•) *represents the initial condition of the motion depicted in Figure* 4. *(At certain points of the branch the corresponding motions in the configuration plane* (y, v) *are depicted.)*

"jumps," but this is only due to the frequency convention adopted, and no actual discontinuities in the dynamics occur. (By definition, branches $S(kn)(km)\pm$, k integer, are identified with $Snm\pm$.) Periodic orbits that correspond to synchronous motions of the two particles of the system, and correspond to curves in the configuration plane (y, v), will be termed *nonlinear normal modes (NNMs)* (Vakakis et al. (1996)).

The main backbone of the frequency-energy plot is formed by the branches $S11\pm$, which represent in- or out-of-phase NNMs possessing one half-wave per half-period. Moreover, the natural frequency of the linear oscillator $\omega_0 = 1$ (which we identify with a frequency index equal to unity, $\omega = 1$) naturally divides the periodic solutions into higher- and lower-frequency modes. A close-up of $S11+$ is presented in Figure 2 together with some modal curves depicted in the configuration plane (y, v) of the system. The horizontal and vertical axes in the plots in the configuration plane are the nonlinear and linear, respectively, oscillator responses, and the aspect ratios in these plots are set so that equal tick mark increments on the horizontal and vertical axes are equal in size, enabling one to directly deduce whether the motion is localized in the linear or the nonlinear oscillator. Figure 2 clearly highlights the energy dependence of the NNMs; the NNMs become strongly localized to the nonlinear attachment as the total energy in the system decreases. This observation shows how useful a frequency-energy plot can be for the interpretation of the dynamics. For the out-of-phase branch $S11-$, the NNMs become localized to y or v as $\omega \to 1+$ or $\omega \gg 1$, respectively.

There is a sequence of higher- and lower-frequency periodic solutions bifurcating

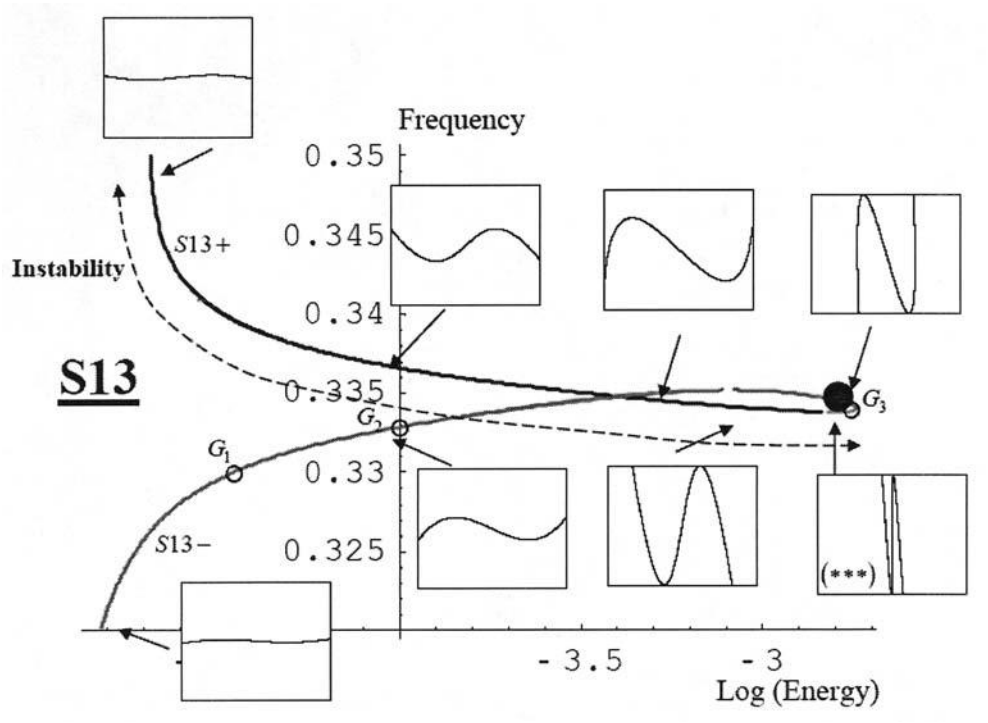


FIG. 3. *Detailed plot of tongues $S13\pm$ in the frequency index-logarithm of the energy plane. Points G_1, G_2, G_3 refer to the text, and the special periodic orbit is represented by triple stars $(***)$; the dot $(\bullet)$ represents the initial condition of the motion depicted in Figure 5. (At certain points of the branch the corresponding motions in the configuration plane (y, v) are depicted.)*

or emanating from branches $S11\pm$, which we will denote as *tongues*. Each tongue occurs in the neighborhood of an internal resonance between the linear oscillator and the nonlinear attachment, and corresponds to either symmetric (S-tongue; e.g., $S13\pm$) or unsymmetric (U-tongue; e.g., $U21\pm$) periodic motion of the system.

Considering first the symmetric solutions, the branches $S1(2k+1)\pm$, $k = 1, 2, \ldots$, appear in the neighborhoods of frequencies $\omega = 1/(2k+1)$, i.e., at progressively lower frequencies with increasing k. For fixed k, each of the two branches $S1(2k+1)\pm$ is linked through a smooth transition with its neighboring branches $S1(2k-1)\pm$ or $S1(2k+3)\pm$ and exists over a finite interval of energy. The pair $S1(2k+1)\pm$ is eliminated through a saddle-node bifurcation at a higher energy value (cf. Figure 3 for branches $S13\pm$). The pairs of branches $S1(2k)\pm$, $k = 1, 2, \ldots$, bifurcate out of $S1(2k+1)\pm$ and exist over finite energy intervals. Branches $Sn1\pm$, $n = 2, 3, \ldots$, appear in the neighborhoods of frequencies $\omega = n$, i.e., at progressively higher frequencies with increasing n; the pair of branches $Sn1\pm$ emanates from $S11-$ and coalesces with $S11+$ through a saddle-node bifurcation. Consider the subharmonic NNMs on tongues $S13\pm$ (similar results hold for the other S-branches), which correspond to motions where the linear oscillator oscillates "three times faster" than the nonlinear attachment. We refer to Figure 3, where a detailed frequency-energy plot for this branch is depicted.

We now discuss the evolution of the motion along $S13-$. As point G_1 is reached in the neighborhood of $\omega = 1/3$, it holds that $v(t) \gg y(t)$, and the nonlinear attachment

vibrates nearly independently, in essence "driving" the linear oscillator; moreover, at that regime of the motion the force generated by the essentially nonlinear coupling spring is approximately equal to that generated by the linear spring. As the energy increases towards point G_2 the nonlinear attachment still "drives" the primary mass, but now the force generated by the linear spring tends to overcome that of the nonlinear spring; this means that the motion of the linear oscillator is less influenced by the motion of the nonlinear attachment. Once point G_2 is reached (with initial displacements $v(0) = 0.0915$, $y(0) = 0.013$ and zero initial velocities), both the linear oscillator and the nonlinear attachment approximately vibrate as a set of *uncoupled linear oscillators* with natural frequencies at ratio 1/3,

$$\ddot{v} + \left(\frac{1}{9}\right) v = 0, \qquad \ddot{y} + y = 0.$$

This means that in the neighborhood of point G_2 of $S13-$ the system oscillates approximately as a system of two uncoupled linear oscillators, a result which explains why the branches $S13\pm$ appear as horizontal straight line segments at frequency index 1/3 of the frequency-energy plot of Figure 1. As energy increases towards point G_3 of Figure 2, the situation is reversed; because the force generated by the nonlinear spring is now negligible compared to that generated by the linear spring, the linear oscillator vibrates nearly independently and drives the nonlinear attachment. Eventually point G_3 is reached, where the periodic motion is approximately given by $y(t) \approx Y \cos \omega t$, $v(t) \approx V \cos \omega t$, and there occurs triple coalescence of branches $S13\pm$ and $S33-$ (which is identical to $S11-$).

Focusing now on the unsymmetric branches, we observe a family of $U(m+1)m$ branches bifurcating from branch $S11-$ that exist over finite energy levels and are eliminated through saddle-node bifurcations with other branches of solutions. The transition of branches $U21$ and $U32$ to $S11+$ seems to involve jumps, but this is only due to the frequency convention adopted, and no actual discontinuities in the dynamics occur. It should be mentioned that periodic motions on the U-tongues are not NNMs because nontrivial phases between the two oscillators are realized. The motion on these tongues is represented by Lissajous curves in the configuration plane, whereas motion on S-tongues corresponds to one-dimensional curves. Localization phenomena are also detected at certain regions of U-tongues (Lee et al. (2005)).

It turns out that certain periodic orbits (termed *special orbits* and depicted by dots in Figure 1) are of particular importance concerning the passive and irreversible energy transfer from the linear to the nonlinear oscillator. These special orbits satisfy the initial conditions $v(0) = \dot{v}(0) = y(0) = 0$ and $\dot{y}(0) \neq 0$, which happen to be identical to the state of the undamped system (1) at $t = 0+$ (being at rest at $t = 0-$) after application of an impulse of magnitude $\dot{y}(0)$ to the linear oscillator. Moreover, certain stable special orbits are localized to the nonlinear oscillator (Lee et al. (2005)) which implies that if the system initially at rest is forced impulsively and one of the stable, localized special orbits is excited, the major portion of the induced energy is channeled directly to the invariant manifold of that special orbit, and hence the motion is rapidly and passively transferred ("pumped") from the linear to the nonlinear oscillator. Therefore, *the impulsive excitation of one of the stable special orbits is one of the triggering mechanisms initiating (direct) passive energy pumping in the system.*

In the following section we discuss in detail three mechanisms for passive energy pumping in system (1). In addition to the mechanism based on excitation of special orbits, we analyze two energy pumping mechanisms that rely on the spatial localiza-

tion of the mode shapes of certain NNMs of Figure 1 as the energy of oscillation of the system decreases due to damping dissipation.

3. Energy pumping mechanisms in the damped system. In this section, the impulsively forced, damped system (1) is considered, and three basic mechanisms for the initiation of nonlinear energy pumping are studied. The first mechanism (*fundamental energy pumping*) is realized when the motion takes place along the backbone curve $S11+$ of the frequency-energy plot of Figure 1, occurring for relatively low frequencies $\omega < \omega_0$. The second mechanism (*subharmonic energy pumping*) resembles the first and occurs when the motion takes place along a lower frequency branch Snm, $n < m$. The third mechanism (*energy pumping initiated by nonlinear beat*), which leads to stronger energy pumping, involves the excitation of a special orbit with main frequency ω_{SO} greater than the natural frequency of the linear oscillator ω_0; in this case energy pumping is initiated by a nonlinear beat phenomenon, as discussed in the previous section. In what follows we discuss each mechanism separately, and provide numerical simulations that demonstrate passive and irreversible energy transfer from the linear oscillator to the nonlinear attachment in each case.

3.1. Fundamental energy pumping. The first mechanism for energy pumping involves excitation of the branch of in-phase synchronous periodic solutions $S11+$, where the linear oscillator and the nonlinear attachment oscillate with identical frequencies in the neighborhood of the fundamental frequency ω_0. Although energy pumping is considered only in the damped system, in order to gain an understanding of the governing dynamics it is necessary to consider the case of no damping.

In Figure 2 we depicted a detailed plot of branch $S11+$ of the undamped system and noted that, at higher energies, the in-phase NNMs are spatially extended (involving finite-amplitude oscillations of both the linear oscillator and the nonlinear attachment). However, *the nonlinear mode shapes of solutions on* $S11+$ *depend essentially on the level of energy, and at low energies they become localized to the attachment.* Considering now the motion in phase space, this low-energy localization is a basic characteristic of the two-dimensional NNM invariant manifold corresponding to $S11+$; moreover, this localization property is preserved in the weakly damped system, where the motion takes place in a two-dimensional damped NNM invariant manifold. This means that when the initial conditions of the damped system are such that they excite the damped analogue of $S11+$, the corresponding mode shape of the oscillation, initially spatially extended, becomes localized to the nonlinear attachment with decreasing energy due to damping dissipation. This, in turn, leads to passive, continuous, and irreversible transfer of energy from the linear oscillator to the nonlinear attachment which acts, in essence, as an NES. The underlying dynamical phenomenon governing fundamental energy pumping was proven to be a *resonance capture on a* 1:1 *resonance manifold* of the system (Vakakis and Gendelman (2001)).

Numerical evidence of fundamental energy pumping is given in Figure 4 for the system with parameters $\varepsilon = 0.05$, $\omega_0^2 = 1$, $C = 1$, and $\lambda_1 = \lambda_2 = 0.0015$. Small damping is considered in order to better highlight the energy pumping phenomenon, and the motion is initiated near the black dot of Figure 2. Comparing the transient responses of Figures 4(a)–(b), we note that the response of the primary system decays faster than that of the NES. The percentage of instantaneous energy captured by the NES versus time is depicted in Figure 4(e) and confirms the assertion that continuous and irreversible transfer of energy from the linear oscillator to the NES takes place; this is more evident by computing the percentage of total input energy that is eventually dissipated by the damper of the NES (cf. Figure 4(f)), which in this

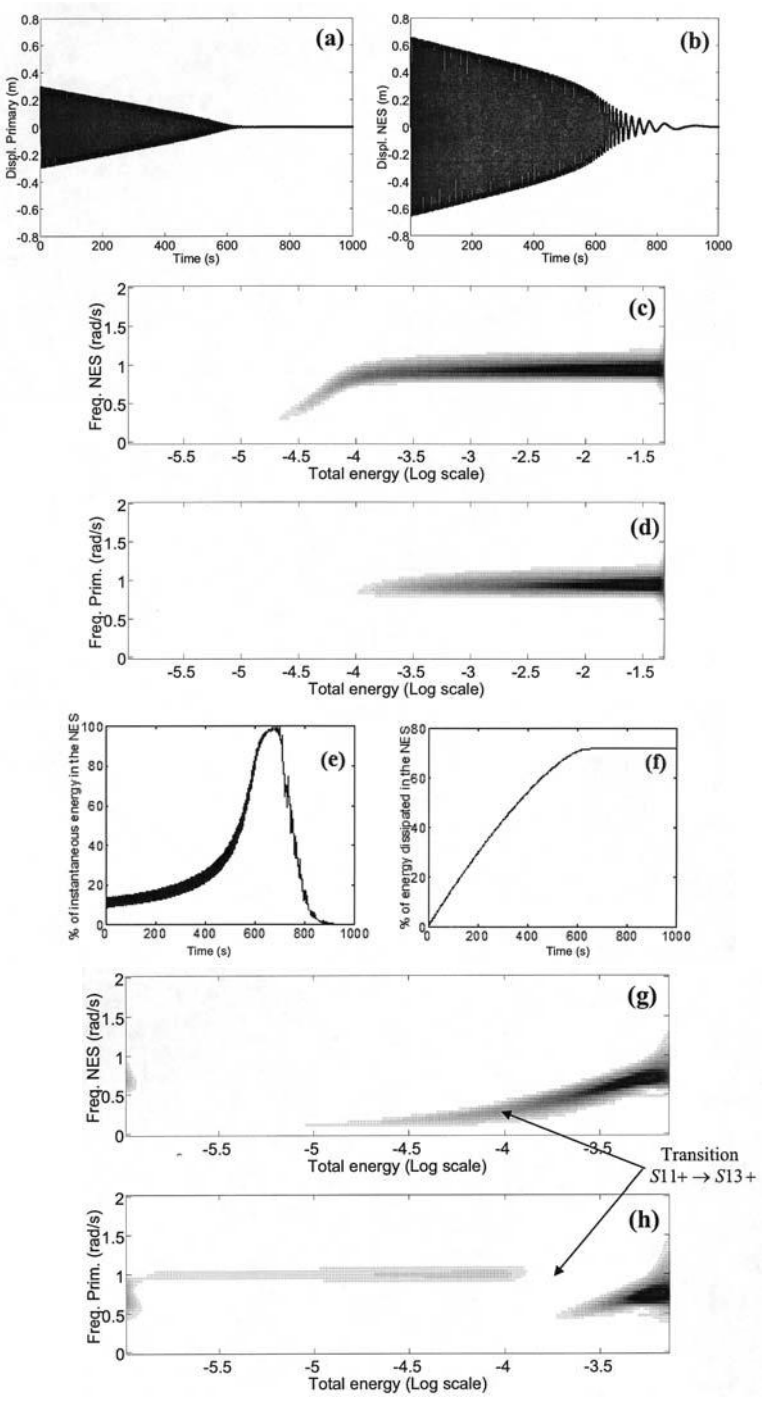


FIG. 4. *Fundamental energy pumping. Shown are the transient responses of the* (a) *linear oscillator and* (b) *NES; WTs of the motion of* (c) *NES and* (d) *linear oscillator;* (e) *percentage of instantaneous total energy in the NES;* (f) *percentage of total input energy dissipated by the NES; transition of the motion from* $S11+$ *to* $S13+$ *at smaller energy levels using the* (g) *NES (observe the settlement of the motion at frequency* 1/3*) and* (h) *linear oscillator.*

particular simulation amounts to 72%; the energy dissipated at the NES is computed by the relation

$$E_{NES}(t) = \lambda_2 \int_0^t \left[\dot{v}(\tau) - \dot{y}(\tau)\right]^2 d\tau.$$

The evolution of the frequency components of the motions of the two oscillators as energy decreases can be studied by numerical wavelet transforms (WTs) (Lee et al. (2005)) of the transient responses, as depicted in Figures 4(c)–(d). These plots highlight that a 1:1 resonance capture is indeed responsible for energy pumping. Below the value of -4 of the logarithm of energy level, the motion of the linear oscillator is too small to be analyzed by the particular windows used in the WT; however, a more detailed WT over smaller energy regimes (cf. Figures 4(g)–(h)) reveals a smooth transition from $S11+$ to $S13+$, in accordance with the frequency-energy plot of Figure 1. This transition manifests itself by the appearance of two predominant frequency components in the responses (at frequencies 1 and 1/3) as energy decreases.

3.2. Subharmonic energy pumping. Subharmonic energy pumping involves excitation of a low-frequency S-tongue. As mentioned previously, by low-frequency tongues we mean the particular regions of the frequency-energy plot where the NES engages in m:n (m, n integers such that $m < n$) resonance captures with the linear oscillator. Another feature of lower tongues is that in these regions the frequency of the motion remains approximately constant with varying energy; as a result, the tongues are represented by horizontal lines in the frequency-energy plot, and the response of system (1) resembles locally that of a linear system (see also discussion about the tongues $S13\pm$ in section 2). In addition, at each specific m:n resonance capture there appears a pair of closely spaced tongues corresponding to in- and out-of-phase oscillations of the two subsystems.

To discuss the dynamics of subharmonic energy pumping we now focus on a particular pair of lower tongues, say $S13\pm$, and refer to Figure 3. As discussed in section 2, at the extremity of a lower pair of tongues, the curve in the configuration plane is strongly localized to the linear oscillator. However, as for the fundamental mechanism for energy pumping, the decrease of energy by viscous dissipation leads to curves in the configuration plane that are increasingly localized to the NES, and nonlinear energy pumping to the NES occurs. In this case, the underlying dynamical phenomenon causing energy pumping is resonance capture in the neighborhood of an m:n resonance manifold of the dynamics. Specifically, for the pair of tongues $S13\pm$, a 1:3 resonance capture occurs that leads to subharmonic energy pumping with the linear oscillator vibrating with a frequency three times that of the NES. It is emphasized that due to the stability properties of the tongues $S13\pm$, subharmonic energy pumping involves excitation of $S13-$ but not of $S13+$.

The transient dynamics when the motion is initiated at the extremity of $S13-$ (cf. the initial condition denoted by the black dot in Figure 3) is displayed in Figure 5. The same parameters as in section 3.1 are considered. Until 500 s, subharmonic energy pumping takes place: despite the presence of viscous dissipation, the NES response grows continuously, with simultaneous rapid decrease of the response of the linear oscillator. A substantial amount of energy is transferred to the NES (cf. Figure 5(e)), and eventually nearly 70% of the energy is dissipated by the NES damper (cf. Figure 5(f)). A prolonged 1:3 resonance capture is nicely evidenced by the WT of Figures 5(c)–(d), and the motion follows the whole lower tongue $S13-$ from the right

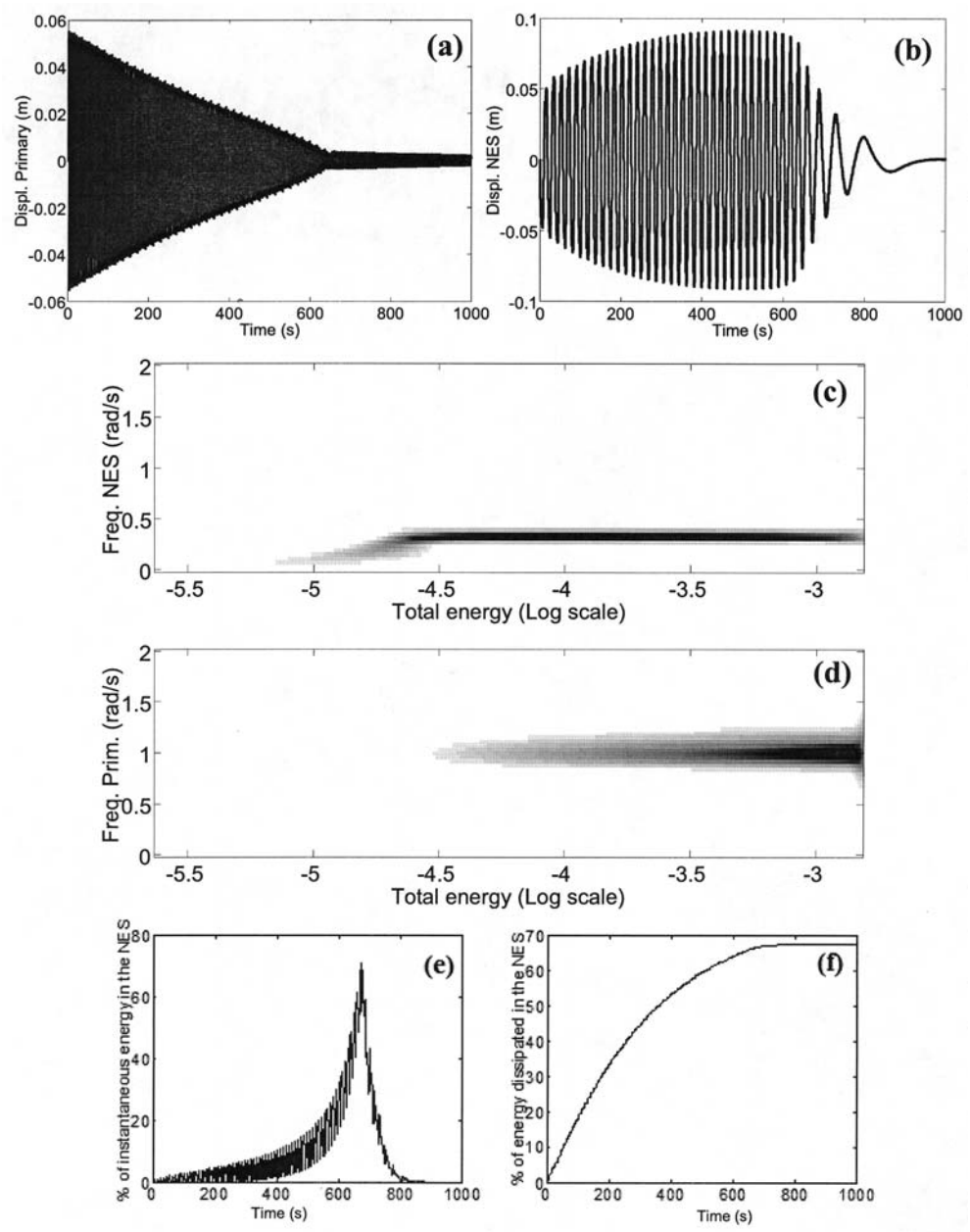


FIG. 5. *Subharmonic energy pumping initiated on S13−. Shown are the transient responses of the* (a) *linear oscillator and* (b) *NES; WTs of the motion of* (c) *the NES and* (d) *the linear oscillator;* (e) *percentage of instantaneous total energy in the NES;* (f) *percentage of total input energy dissipated by the NES.*

to the left. Once escape from resonance capture occurs (around 620–630 s), energy is no longer transferred to the NES.

3.3. Energy pumping initiated by nonlinear beating. The previous two mechanisms cannot be activated with the NES at rest, since in both cases the motion

is initialized from a nonlocalized state of the system. This means that these energy pumping mechanisms cannot be activated directly after the application of an impulsive excitation to the linear oscillator with the NES initially at rest. Such a forcing situation, however, is important from a practical point of view; indeed, this is the situation where local NESs are utilized to confine and passively dissipate unwanted vibrations from linear structures that are forced by impulsive (or broadband) loads.

Hence, it is necessary to discuss an alternative, third energy pumping mechanism capable of initiating passive energy transfer with the NES being initially at rest. This alternative mechanism is based on the excitation of a *special orbit* (as defined and discussed in section 2) that plays the role of a "bridging orbit" for activation of either fundamental or subharmonic energy pumping. Excitation of a special orbit results in the transfer of a substantial amount of energy from the initially excited linear oscillator directly to the NES through a nonlinear beat phenomenon. In that context, the special orbit may be regarded as an initial bridging orbit or trigger, which eventually activates fundamental or subharmonic energy pumping, once the initial nonlinear beat initiates the energy transfer. Indeed, as shown below, *the third mechanism for energy pumping represents an efficient initial (triggering) mechanism for rapid transfer of energy from the linear oscillator to the NES at the crucial initial stage of the motion, before activating either one of the (fundamental or subharmonic) main energy pumping mechanisms through a nonlinear transition (jump) in the dynamics.*

To study the dynamics of this triggering mechanism, we first formulate the following conjecture: *Due to the essential (nonlinearizable) nonlinearity, the NES is capable of engaging in an m:n resonance capture with the linear oscillator, m and n being a set of integers. Accordingly, in the undamped system there exists a sequence of special orbits (corresponding to nonzero initial velocity of the linear oscillator with all other initial conditions zero), aligned along a one-dimensional smooth manifold in the frequency-energy plot.* As a first step to test this conjecture, a nonlinear boundary value problem (NLBVP) was formulated to compute the periodic orbits of system (1) with no forcing and damping, and the additional restriction for the special orbits was imposed. (For a detailed formulation of the NLBVP, we refer to Lee et al. (2005).) The numerical results in the frequency-energy plane are depicted in Figure 6 for parameters $\varepsilon = 0.05$, $\omega_0^2 = 1$, $C = 1$. Each triangle in the plot represents a special orbit, and a one-dimensional manifold appears to connect the special orbits (though a rigorous proof of the existence of this manifold is not given here). In addition, it appears that there exists a countable infinity of special orbits, occurring in the neighborhoods of the countable infinities of internal resonances m:n (m, n integers) of the system, but again no rigorous proof of this conjecture is given in this paper. We note that a subset of high-frequency branches (for $\omega > 1$) possesses two special orbits instead of one (for example, all $U(p+1)p$ branches with $p \geq 3$). To distinguish between the two special solutions in such high-frequency branches we partition them into two subclasses: the a-special orbits, which exist in the neighborhood of $\omega = \omega_0 = 1$, and the b-special orbits, which occur away from this neighborhood (cf. Figure 6); it was numerically proven in Lee et al. (2005) that the a-special orbits are unstable, whereas the b-special orbits are stable. As shown below, it is the excitation of the stable b-special orbits that activates the third mechanism for energy pumping.

Representative special orbits are given in Figure 7. By construction, all special orbits have a common feature; namely, they pass with vertical slope through the origin of the configuration plane (y, v). This feature renders them compatible with an impulse applied to the linear oscillator, which corresponds to a nonzero velocity of the linear oscillator with all other initial conditions zero. The curves corresponding to

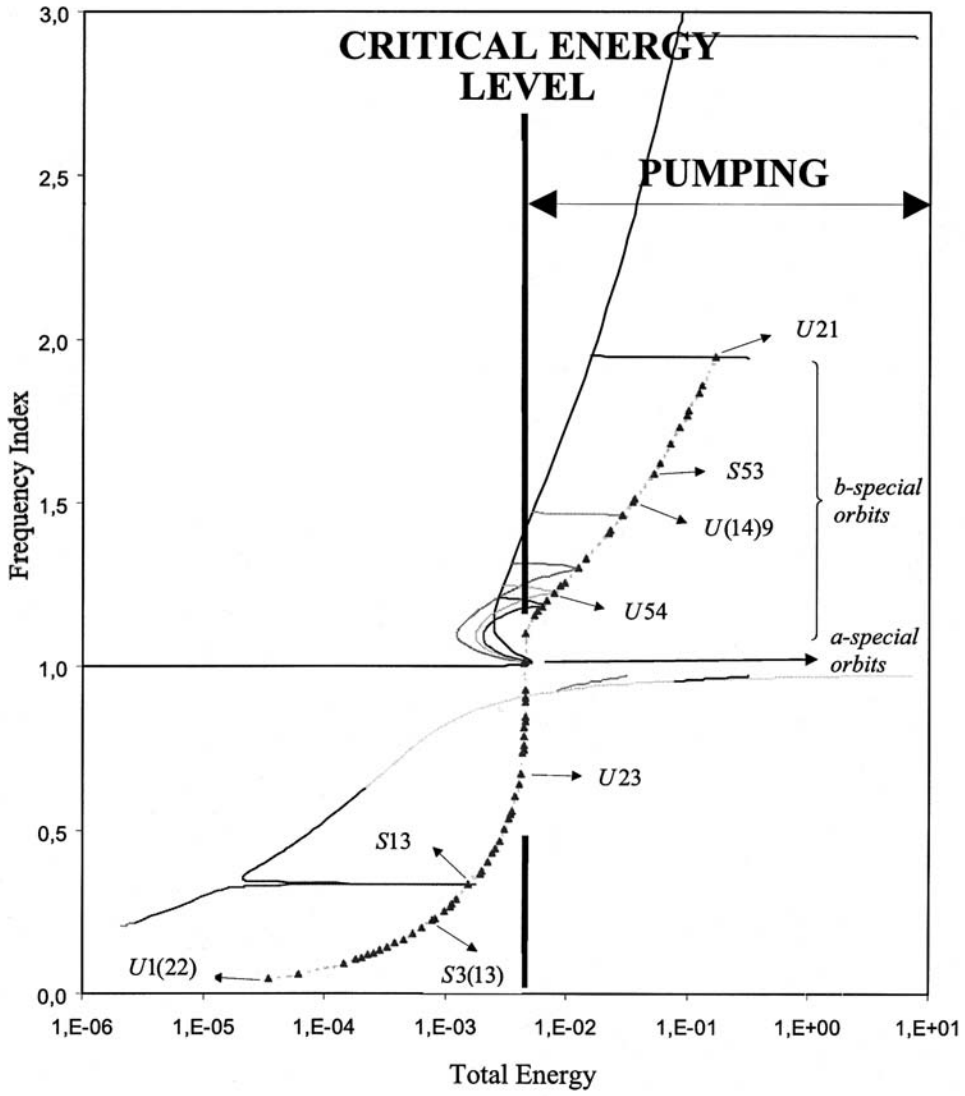


FIG. 6. *Manifold of special orbits (represented by triangles) in the frequency-energy plot.*

the special orbits in the configuration plane can be either closed or open, depending upon the differences between the two indices characterizing the orbits; specifically, odd differences between indices correspond to closed curves in the configuration plane and lie on U-branches, whereas even differences between indices correspond to open curves on S-branches. In addition, higher-frequency special orbits (with frequency index $\omega > \omega_0$) in the upper part of the frequency-energy plot (i.e., $m > n$) are localized to the nonlinear oscillator; conversely, special orbits in the lower part of the frequency-energy plot (with frequency index $\omega < \omega_0$) tend to be localized to the linear oscillator. This last observation is of particular importance since it directly affects the transfer of a significant amount of energy from the linear oscillator to the NES through the mechanism discussed in this section: indeed, there seems to be a well-

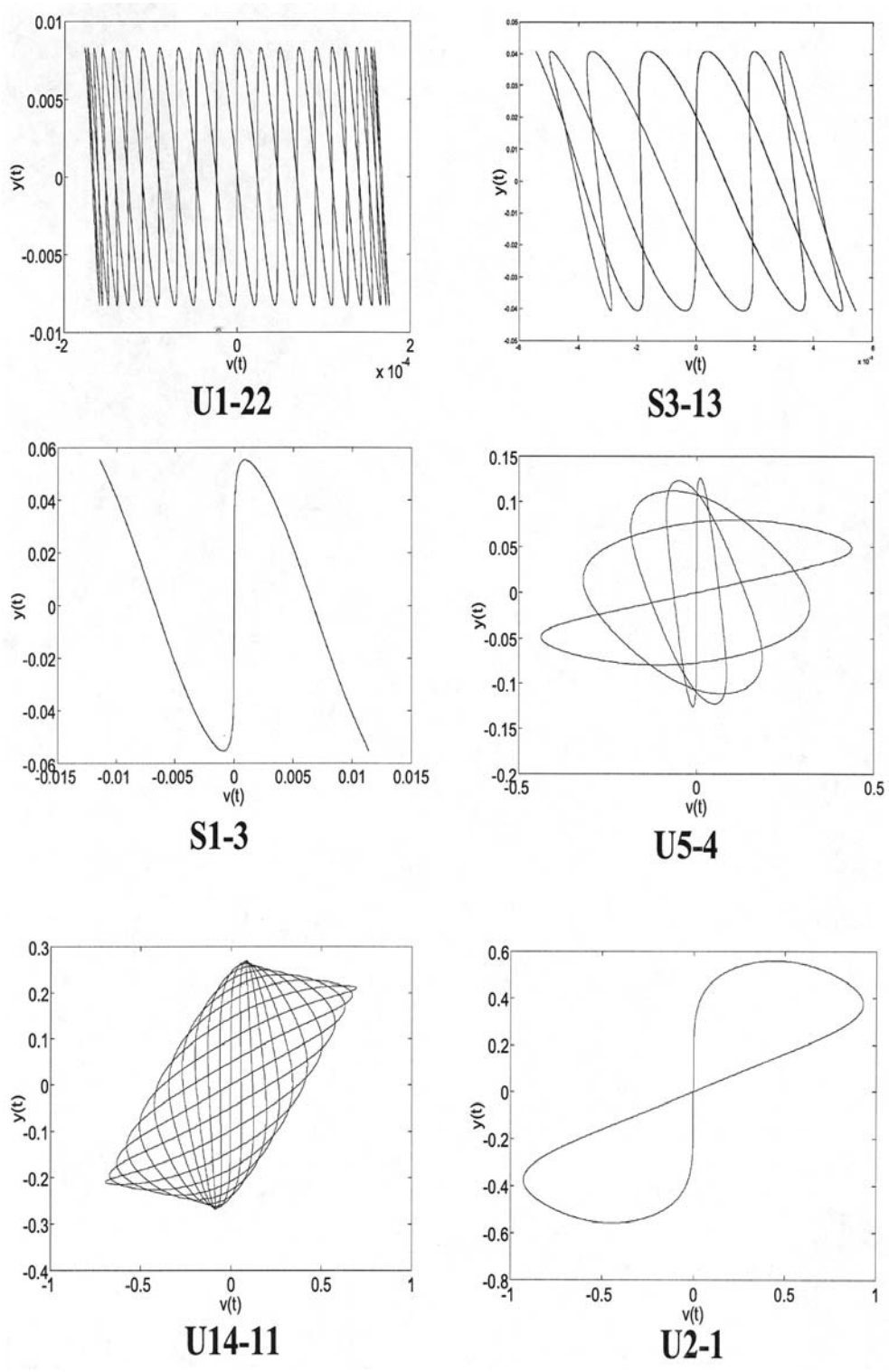


FIG. 7. *Representative special orbits in the configuration plane* (y, v). *Closed curves correspond to special orbits on* U*-branches, and open curves to special orbits on* S*-branches.*

defined critical threshold of energy that separates high- from low-frequency special orbits, i.e., those that do or do not localize, respectively, to the NES (cf. Figure 6). *The third mechanism for energy pumping can be activated only for input energies above the critical threshold*, since below that the (low-frequency) special orbits are incapable of transferring significant amounts of input energy from the linear oscillator to the NES; in other words, the critical level of energy represents a lower bound below which no significant energy pumping can be initiated through activation of a special orbit. Moreover, combining this result with the topology of the one-dimensional manifold of special orbits of Figure 6, it follows that *it is the subclass of stable b-special orbits that is responsible for activating the third energy pumping mechanism, whereas the subclass of unstable a-special orbits does not affect energy pumping.* This theoretical insight will be fully validated by the numerical simulations that follow.

We now proceed to analyze in detail the nonlinear beat phenomenon that takes place when a special orbit is excited by the initial conditions. When the NES engages in an m:n resonance capture with the linear oscillator, a nonlinear beat phenomenon takes place. Due to the essential (nonlinearizable) nonlinearity of the NES and the lack of any preferential frequency, the considered nonlinear beat phenomenon does not require any a priori "tuning" of the nonlinear attachment, since at the specific frequency-energy range of the m:n resonance capture the nonlinear attachment adjusts its amplitude ("tunes itself") to fulfill the necessary conditions of internal resonance. This represents a significant departure from the classical nonlinear beat phenomenon observed in coupled oscillators with linearizable nonlinear stiffnesses (e.g., spring-pendulum systems), where the defined ratios of linearized natural frequencies of the component subsystems dictate the type of internal resonances that can be realized (Golnaraghi (1991), Salemi, Golnaraghi, and Heppler (1997)). As an example, in Figure 8 we depict the exchanges of energy during the nonlinear beat phenomena corresponding to the special orbits of branches $U21$ and $U54$ for parameters $\varepsilon = 0.05$, $\omega_0^2 = 1$, $C = 1$ and no damping. As expected, energy is continuously exchanged between the linear oscillator and the NES, so the energy transfer is not irreversible as is required for energy pumping; *we conclude that excitation of a special orbit can only initiate (trigger) energy pumping, but not cause it in itself.* The amount of energy transferred during each cycle of the beat varies with the special orbit considered; for $U21$ and $U54$, as much as 32% and 86% of energy, respectively, can be transferred to the NES. It can be shown that, for increasing integers m and n with corresponding ratios $m/n \to 1+$, the maximum energy transferred during a cycle of the special orbit tends to 100%; at the same time, however, the resulting period of the cycle of the beat (and, hence, of the time needed to transfer the maximum amount of energy) should increase as the least common multiple of m and n.

We note at this point that the nonlinear beat phenomenon associated with the excitation of the special orbits can be studied analytically using the complexification-averaging method first introduced by Manevitch (1999). To demonstrate the analytical procedure, we analyze in detail the special orbit on branch $U21$ of the system with no damping. In Lee et al. (2005), the periodic motions on this entire branch were studied, and it was shown that the responses of the linear oscillator and the nonlinear attachment can be approximately expressed as

$$\begin{aligned} y(t) &= Y_1 \sin \omega t + Y_2 \sin 2\omega t \equiv y_1(t) + y_2(t), \\ v(t) &= V_1 \sin \omega t + V_2 \sin 2\omega t \equiv v_1(t) + v_2(t), \end{aligned} \tag{2}$$

where

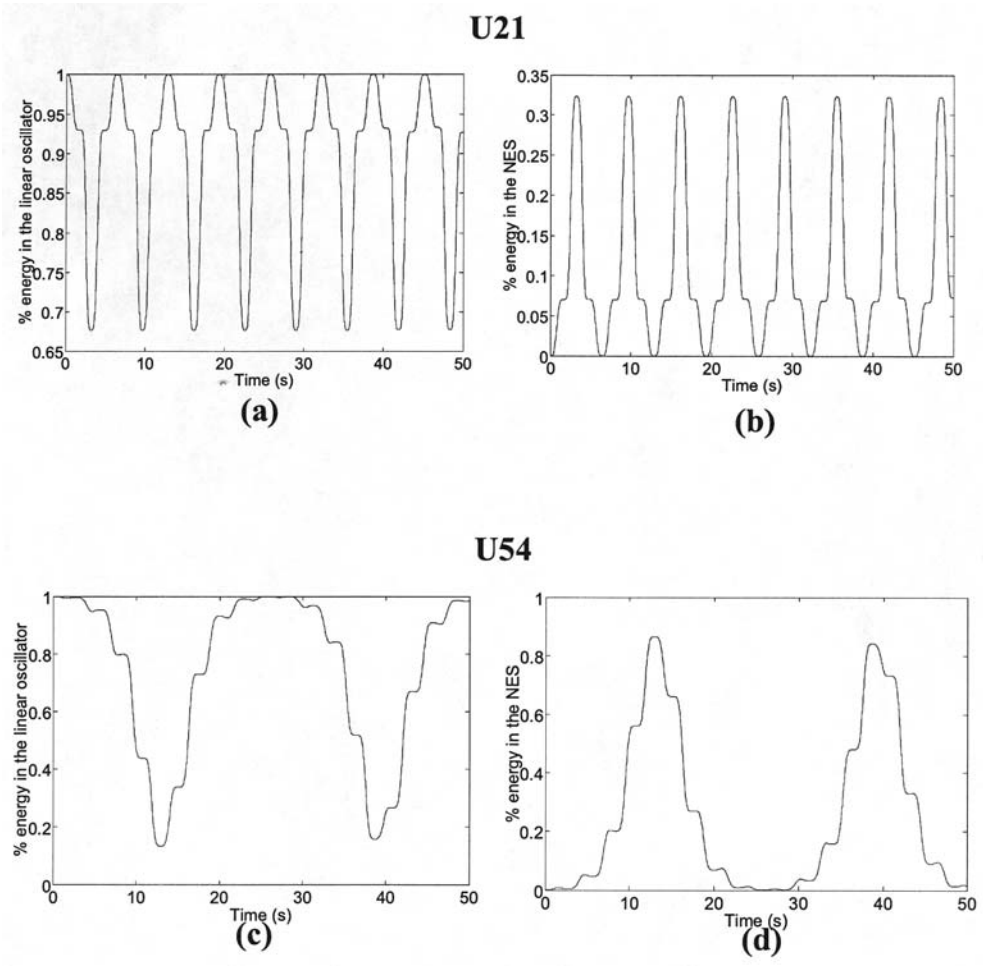


FIG. 8. *Exchanges of energy during nonlinear beat phenomena corresponding to special orbits on* (a), (b) $U21$, *and* (c), (d) $U54$.

$$A = \frac{\varepsilon\omega^2}{\omega_0^2 - \omega^2} B \quad \text{and} \quad D = \frac{4\varepsilon\omega^2}{\omega_0^2 - 4\omega^2} G,$$

$$B = \pm\sqrt{\frac{4\varepsilon\omega^4(Z_2 - 8Z_1)}{9CZ_1^3 Z_2}} \quad \text{and} \quad G = \pm\sqrt{\frac{32\varepsilon\omega^4(2Z_1 - Z_2)}{9CZ_2^3 Z_1}},$$

$$Z_1 = \frac{\varepsilon\omega^2}{\omega_0^2 - \omega^2} - 1 \quad \text{and} \quad Z_2 = \frac{4\varepsilon\omega^2}{\omega_0^2 - 4\omega^2} - 1,$$

$$Y_1 = \frac{A}{\omega}, \quad V_1 = \frac{B}{\omega}, \quad Y_2 = \frac{D}{2\omega}, \quad V_2 = \frac{G}{2\omega}.$$

Hence, a two-frequency approximation is satisfactory for this family of periodic motions. The frequency ω_{SO} at which the special orbit appears is computed by imposing the initial conditions $y(0) = v(0) = \dot{v}(0) = 0$, which leads to the relation

$$B = -2G \qquad \text{(special orbit)}.$$

The instantaneous fraction of total energy in the linear oscillator during the nonlinear beat phenomenon is estimated to be

$$E_{linear}(t) = \frac{\left[(\omega_0^2 - 4\omega_{SO}^2)\sin\omega_{SO}t - 2(\omega_0^2 - \omega_{SO}^2)\sin 2\omega_{SO}t\right]^2}{9\omega_{SO}^2\omega_0^2} + \frac{\left[(\omega_0^2 - 4\omega_{SO}^2)\cos\omega_{SO}t - 4(\omega_0^2 - \omega_{SO}^2)\cos 2\omega_{SO}t\right]^2}{9\omega_0^4}. \tag{3}$$

The nonlinear coefficient C has no influence on the fraction of total energy transferred to the NES during the nonlinear beat; this means that, during the beat, the instantaneous energies of the linear oscillator and the NES are directly proportional to the nonlinear coefficient. Moreover, as the mass of the NES tends to zero, the frequency where the special orbit is realized tends to the limit $\omega_{SO} \to \omega_0$, and, as a result, $E_{linear}(t) \to 1$, and the energy transferred to the NES during the beat tends to zero. However, we note that this is a result satisfied only asymptotically, since, as indicated by the results depicted in Figure 8, even for very small mass ratios, i.e., $\varepsilon = 0.05$, as much as 86% of the total energy can be transferred to the NES during a cycle of the special orbit of branch $U54$.

Considering now the damped system, we will show that following an initial nonlinear beat phenomenon, either one of the main (fundamental or subharmonic) energy pumping mechanisms can be activated through a nonlinear transition (jump) in the dynamics. It was previously mentioned that the two main energy pumping mechanisms are qualitatively different from the third mechanism, which is based on the excitation of a nonlinear beat phenomenon (special orbit); indeed, damping is a prerequisite for the realization of the two main mechanisms, leading to an irreversible energy transfer from the linear oscillator to the NES, whereas a special orbit is capable of transferring energy without dissipation, though this transfer is not irreversible but periodic. This justifies our earlier assertion that the third mechanism does not represent an independent mechanism for energy pumping, but rather triggers it, and through a nonlinear transition activates either of the two main mechanisms. This will become apparent in the following numerical simulations.

The following simulations concern the transient dynamics of the damped system (1) with parameters $\varepsilon = 0.05$, $\omega_0^2 = 1$, $C = 1$, $\lambda_1 = \lambda_2 = 0.0015$ and an impulse of magnitude Y applied to the linear oscillator (corresponding to initial conditions $y(0+) = v(0+) = \dot{v}(0+) = 0$, $\dot{y}(0+) = Y$). By varying the magnitude of the impulse we study the different nonlinear transitions that take place in the dynamics and their effects on energy pumping. The responses of the system to the relatively strong impulse $Y = 0.25$ are depicted in Figure 9. Inspection of the WTs of the responses (cf. Figures 9(c)–(d)), and of the portion of total instantaneous energy captured by the NES (cf. Figure 9(e)), reveals that at the initial stage of the motion (until approximately $t = 120$ s) the (stable) b-special orbit on branch $U32$ is excited (since the NES response possesses two main frequency components at 1 and 3/2 rad/s), and a nonlinear beat phenomenon takes place. (Note the continuous exchange of energy between the two subsystems—reversibility in this initial stage of the motion.) For $t > 120$ s, the dynamics undergoes a transition (jump) to branch $S11+$, and fundamental energy pumping to the NES occurs on a prolonged 1:1 resonance capture (cf. Figures 9(c)–(d)); eventually, 84% of the input energy is dissipated by the damper of the NES (cf. Figure 9(f)).

Lowering the magnitude of the impulse to $Y = 0.11$ gives rise to a different set of nonlinear transitions, as the simulations of Figure 10 indicate. In this case the (stable)

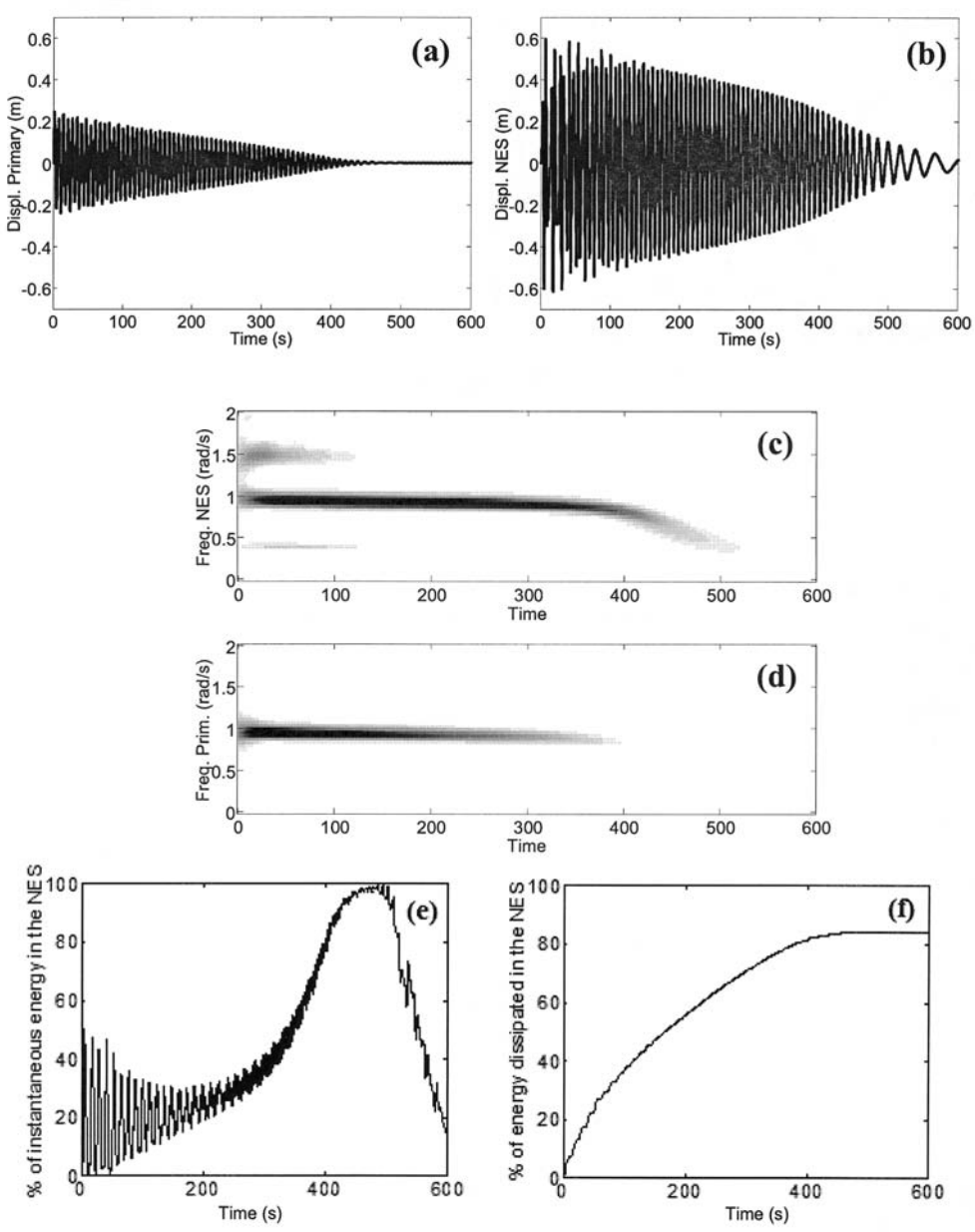


FIG. 9. *Energy pumping by nonlinear beat, transition to* $S11+$. *Shown are transient responses of* (a) *the linear oscillator and* (b) *NES; WTs of the motion of* (c) *NES and* (d) *the linear oscillator;* (e) *percentage of instantaneous total energy in the NES;* (f) *percentage of total input energy dissipated by the NES.*

b-special orbit of branch $U43$ is initially excited, which then activates subharmonic energy pumping through a nonlinear transition to the tongue $S13-$. In other words, the lower tongue appears to act as "bait" and activates energy pumping through 1:3 resonance capture, i.e., by capturing locally the transient dynamics in its domain of attraction. Figure 10(e) reveals that a nonlinear beat phenomenon occurs until

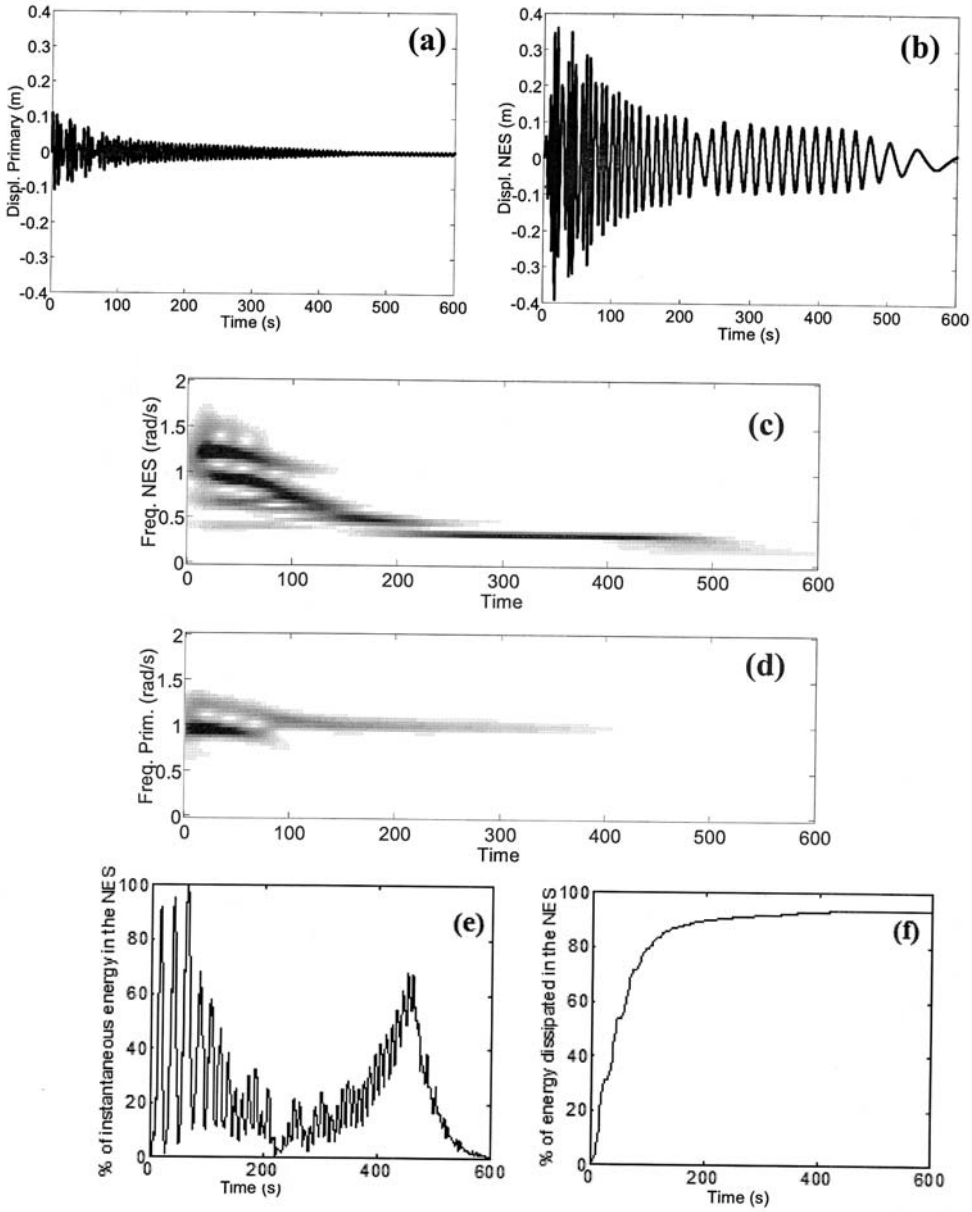


FIG. 10. *Energy pumping by nonlinear beat, transition to* $S13-$. *Shown are transient responses of* (a) *the linear oscillator and* (b) *NES; WTs of the motion of* (c) *NES and* (d) *the linear oscillator;* (e) *percentage of instantaneous total energy in the NES;* (f) *percentage of total input energy dissipated by the NES.*

$t = 150$ s approximately, and in turn activates 1:3 subharmonic energy pumping to the NES (cf. Figures 10(c)–(d)); eventually, 94% of the input energy is dissipated by the NES (cf. Figure 10(f)).

Comparing the two simulations, we conclude that excitation of the *b*-special orbit of $U43$ leads to more effective energy pumping compared to the *b*-special orbit of $U32$. The reason rests with the localization properties of the special orbits, i.e.,

their capacity to transfer a larger fraction of the total energy to the NES during a cycle of the oscillation. Indeed, the localization properties of the b-special orbits of branches $U(p+1)p$ are enhanced as the order p increases, i.e., as their second frequency component, $\omega = (p+1)/p$, approaches that of the first, $\omega = 1$, and the special orbits topologically approach the branch $S11-$ (Lee et al. (2005)). In that context, the b-special orbit of $U43$ is capable of transferring a larger fraction of the input than that being transferred by the b-special orbit of $U32$ (compare Figures 9(e) and 10(e)), and hence the enhanced energy pumping results of the second simulation. We note at this point that the (unstable) a-special orbits of these branches are localized in the linear oscillator and do not affect energy pumping.

We now test the previous theoretical finding that, for sufficiently small impulse magnitudes, no energy pumping can occur, i.e., none of the three aforementioned energy pumping mechanisms can be activated. The simulations for $Y = 0.08$ are depicted in Figure 11. The WT of the NES response of Figure 11(c) shows the presence of a frequency component below $\omega = 1$ at the initial stage of the motion, which indicates excitation of a low-frequency special orbit in the lower part ($\omega < 1$) of the frequency-energy plot (i.e., $U12$). As explained previously, those orbits are localized to the linear oscillator and, as a result, cannot transfer a sufficient amount of energy to the NES in the initial stage of the motion. Accordingly, neither the fundamental nor the subharmonic energy pumping mechanism is eventually activated, leading to a much smaller amount of energy dissipated by the NES (around 45% in this case). This result confirms our previous assertion that energy pumping through nonlinear beat can be activated only above a critical energy threshold (cf. Figure 6).

To demonstrate more clearly the effect of the b-special orbits on energy pumping, in Figure 12 we depict the percentage of input energy eventually dissipated at the NES for varying magnitudes of the impulse for the system with parameters $\varepsilon = 0.05$, $\omega_0^2 = 1$, $C = 1$, $\lambda_1 = \lambda_2 = 0.01$. In the same plot we depict the positions of the special orbits of the undamped system and the critical threshold predicted in Figure 6. We conclude that strong energy pumping is associated with the excitation of b-special orbits of the branches $U(p+1)p$ in the neighborhood above the critical threshold, whereas excitation of a-special orbits below the critical threshold does not lead to rigorous energy pumping. As mentioned previously, in the neighborhood of the critical threshold the b-special orbits are strongly localized to the NES, whereas a-orbits are nonlocalized. We also note from Figure 12 the deterioration of energy pumping as we increase the magnitude of the impulse well above the critical threshold, where high-frequency special orbits are excited; this is a consequence of the fact that well above the critical threshold the special orbits are weakly localized to the NES.

Extending the previous result, in Figure 13 we depict the contours of energy eventually dissipated at the NES, but now for the case of two impulses of magnitudes $\dot{y}(0)$ and $\varepsilon\dot{v}(0)$ applied to both the linear oscillator and the NES, respectively. The system parameters used were identical to those of the previous simulation of Figure 12. Superimposed on contours of energy dissipated at the NES are certain high- and low-frequency U- and S-branches of the undamped system together with their special orbits, in order to confirm for this case the essential role of the high-frequency special orbits in energy pumping. Indeed, high levels of energy dissipation are encountered in neighborhoods of contours of high-frequency U-branches, whereas low values are noted in the vicinity of low-frequency branches. These results agree qualitatively with our earlier theoretical and numerical findings and enable us to assess and establish the robustness of energy pumping when the NES is not initially at rest.

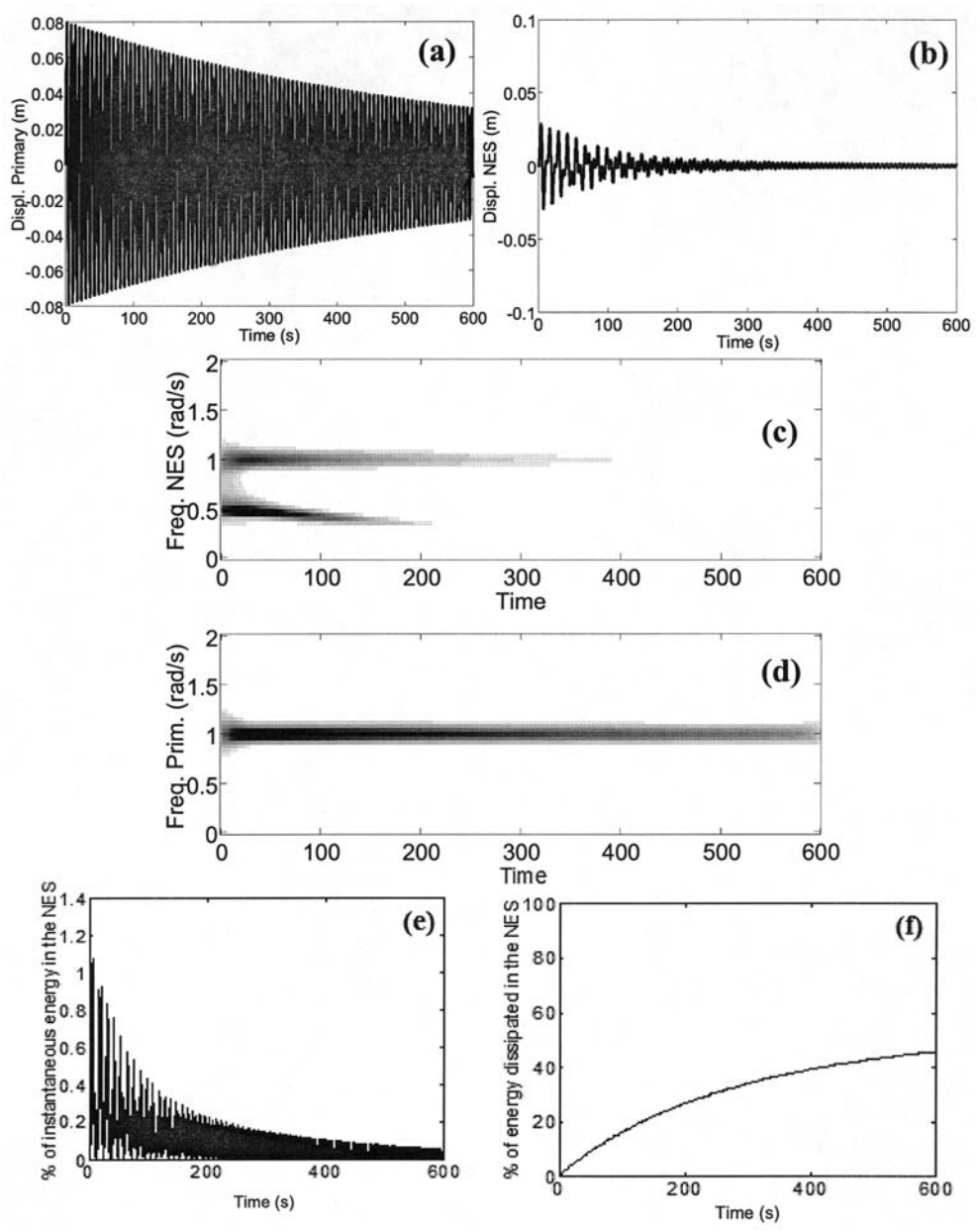


FIG. 11. *Absence of energy pumping for low excitation. Shown are transient responses of* (a) *the linear oscillator and* (b) *NES; WTs of the motion of* (c) *NES and* (d) *the linear oscillator;* (e) *percentage of instantaneous total energy in the NES;* (f) *percentage of total input energy dissipated by the NES.*

In the next section we provide analytical studies of the fundamental and subharmonic energy pumping mechanisms encountered in the damped system; since excitation of nonlinear beats is merely a means for activating the main two energy pumping mechanisms, it will not be analyzed below. We show that in each case we can reduce the governing dynamics of energy pumping to low-order slow-flow dynamical systems.

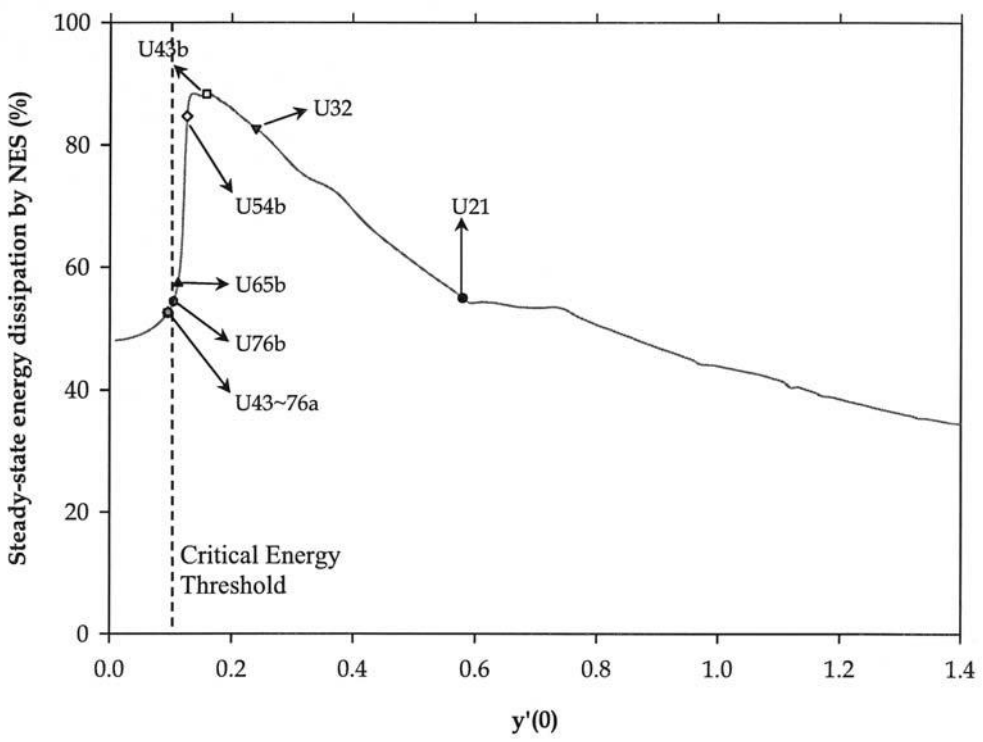


FIG. 12. *Percentage of input energy eventually dissipated at the NES for varying magnitudes of the impulse (the positions of certain special orbits are indicated).*

4. Slow-flow analysis. We now focus on the resonance capture dynamics that governs energy pumping in the damped system. This can be studied by performing a reduction of the dynamics, an approximate partition between slow and fast dynamics, and considering the evolution of the slow-flow dynamics when energy pumping takes place. We show that even though the system of coupled oscillators possesses essential (nonlinearizable) stiffness nonlinearities, analytical modeling of its dynamics at certain motion regimes can still be performed.

The theory of resonance processes for multifrequency systems was developed by Neishtadt (1997), (1999), where capture into and scattering on the resonance were discussed by considering them as random events and computing probabilities of capture and probabilistic distributions of the scattering amplitudes. By assuming small perturbations (e.g., weak nonlinearities), action-angle formulations and the averaging theorem were applied to provide analytical asymptotic validity of the approximations. Also, by introducing a mapping (called the *in-out function*) from a state of resonance capture to that of escape, glued averaging approximation was utilized to analytically describe motions when they are away from, captured into, and escaped from the resonance manifold.

Similar formulations were considered in Vakakis and Gendelman (2001), where the slow-flow equations were established also by the complexification/averaging technique. This method does not necessarily require the perturbations to be small, although it is similar to the (classical) averaging method; once the proper ansatz regarding the frequency content of the response is included, it is numerically verified that the

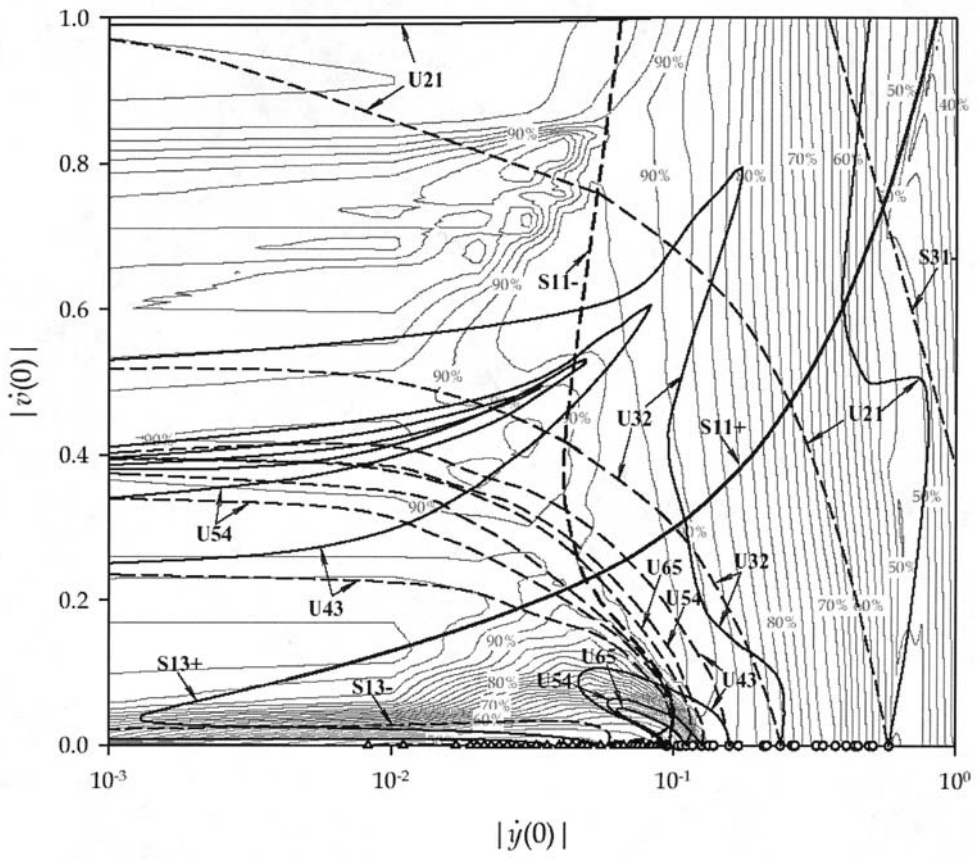


FIG. 13. *Contours of percentages of input energy eventually dissipated at the NES for the case when both oscillators are excited by impulses: superimposed are contours of high- and low-frequency branches of the undamped system (solid line: in phase, dashed line: out of phase branches); special curves in high- and low-frequency branches are denoted by* (◯) *and* (Δ)*, respectively.*

slow-flow model provides good approximation of the original dynamics and the entire resonance processes as well.

In this study, we show that even though the system of coupled oscillators possesses essential (nonlinearizable) stiffness nonlinearities, analytical modeling of its dynamics at certain motion regimes can still be performed by means of the multifrequency complexification/averaging method.

Focusing first on the fundamental energy pumping mechanism, we again consider system (1) and introduce the new complex variables

$$\begin{aligned} \psi_1(t) &= \dot{v}(t) + jv(t) \equiv \varphi_1(t)\, e^{jt}, \\ \psi_2(t) &= \dot{y}(t) + jy(t) \equiv \varphi_2(t)\, e^{jt}, \end{aligned} \tag{4}$$

where $\varphi_i(t)$, $i = 1, 2$, represent slowly varying complex amplitudes and $j = (-1)^{1/2}$. By writing (4) we introduce a partition of the dynamics into slow and fast components and seek slowly modulated fast oscillations at frequency $\omega = \omega_0 = 1$. As discussed previously, fundamental energy pumping is associated with this type of motion in the neighborhood of branch $S11+$ in the frequency-energy plot of the undamped

dynamics. Expressing the system responses in terms of the new complex variables, $y = (\psi_2 - \psi_2^*)/2j$, $v = (\psi_1 - \psi_1^*)/2j$ (where $(*)$ denotes complex conjugate); substituting into (1) with $P(t) = 0$; and averaging over the fast frequency, we derive a set of approximate slow modulation equations that govern the evolutions of the complex amplitudes,

$$\begin{aligned}
&\dot{\varphi}_1 + \left(\frac{j}{2}\right)\varphi_1 + \left(\frac{\lambda}{2}\right)(\varphi_1 - \varphi_2) - \left(\frac{3jC}{8\varepsilon}\right)|\varphi_1 - \varphi_2|^2(\varphi_1 - \varphi_2) = 0,\\
&\dot{\varphi}_2 - \left(\frac{\varepsilon\lambda}{2}\right)(\varphi_1 - \varphi_2) - \left(\frac{3jC}{8}\right)|\varphi_2 - \varphi_1|^2(\varphi_2 - \varphi_1) + \left(\frac{\varepsilon\lambda}{2}\right)\varphi_2 = 0.
\end{aligned} \tag{5}$$

For the sake of simplicity, we have assumed that $\lambda_1 = \lambda_2 = \lambda$ in (1). To derive a set of real modulation equations, we express the complex amplitudes in polar form, $\varphi_i(t) = a_i(t)\, e^{j\beta_i(t)}$, $i = 1, 2$, substitute into (5), and separately set equal to zero the real and imaginary parts. We then reduce (5) to an autonomous set of equations that govern the slow evolution of the two amplitudes $a_1(t)$ and $a_2(t)$ and the phase difference $\phi(t) = \beta_2(t) - \beta_1(t)$:

$$\begin{aligned}
&\dot{a}_1 + \left(\frac{\lambda}{2}\right)a_1 - \left(\frac{\lambda}{2}\right)a_2\cos\phi - \left(\frac{3C}{8\varepsilon}\right)\left(a_1^2 + a_2^2 - 2a_1a_2\cos\phi\right)a_2\sin\phi = 0,\\
&\dot{a}_2 - \left(\frac{\varepsilon\lambda}{2}\right)a_1\cos\phi + \varepsilon\lambda a_2 + \left(\frac{3C}{8}\right)\left(a_1^2 + a_2^2 - 2a_1a_2\cos\phi\right)a_1\sin\phi = 0,\\
&\dot{\phi} + \left(\frac{\lambda}{2}\right)\left[\left(\frac{\varepsilon a_1}{a_2}\right) + \left(\frac{a_2}{a_1}\right)\right]\sin\phi - \left(\frac{1}{2}\right)\\
&+ \left(\frac{3C}{8}\right)\left(a_1^2 + a_2^2 - 2a_1a_2\cos\phi\right)\left\{\left(\frac{1}{\varepsilon}\right)\left[1 - \left(\frac{a_2}{a_1}\right)\cos\phi\right] - \left[1 - \left(\frac{a_1}{a_2}\right)\cos\phi\right]\right\} = 0.
\end{aligned} \tag{6}$$

This reduced dynamical system governs the slow-flow dynamics of fundamental energy pumping. In particular, 1:1 *resonance capture (the underlying dynamical mechanism of fundamental energy pumping) is associated with non–time-like behavior of the phase variable ϕ or, equivalently, failure of the averaging theorem in the slow-flow* (6). Indeed, when ϕ exhibits time-like, nonoscillatory behavior, one can apply the averaging theorem over ϕ and prove that the amplitudes a_1 and a_2 decay exponentially with time and that no significant energy exchanges (energy pumping) can take place. In Figure 14(a) we depict 1:1 resonance capture in the slow-flow phase plane $(\dot{\phi}, \phi)$ for system (6) with $\varepsilon = 0.05$, $\lambda = 0.01$, $C = 1$, $\omega_0 = 1$ and initial conditions $a_1(0) = 0.01$, $a_2(0) = 0.24$, $\phi(0) = 0$. The oscillatory behavior of the phase variable in the neighborhood of the in-phase limit $\phi = 0+$ indicates 1:1 resonance capture (on branch $S11+$ of the frequency-energy plot of Figure 1) and leads to energy pumping from the linear oscillator to the NES, as evidenced by the build-up of amplitude a_1 (cf. Figure 14(b)). Escape from resonance capture is associated with time-like behavior of ϕ and rapid decrease of the amplitudes a_1 and a_2 (as predicted by averaging in (6)). A comparison of the analytical approximation (4)–(6) and direct numerical simulation for the previous initial conditions confirms the accuracy of the analysis.

Considering subharmonic energy pumping, we will focus on energy pumping in the neighborhood of tongue $S13-$ (similar analysis can be applied for other orders of subharmonic resonance captures). Due to the fact that motions in the neighborhood of $S13-$ possess two main frequency components, at frequencies 1 and 1/3, we express

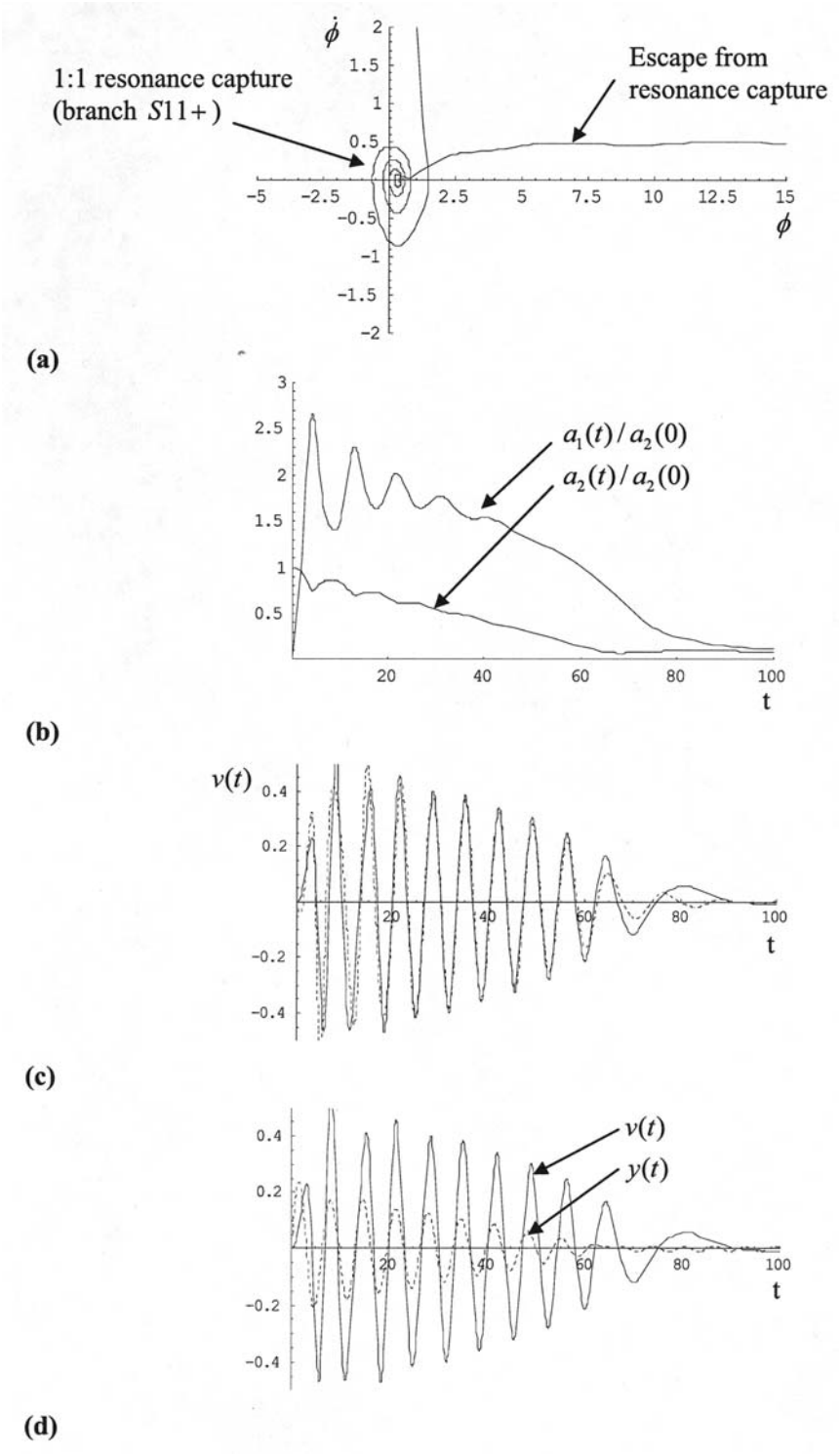


FIG. 14. *Fundamental energy pumping:* (a) 1:1 *resonance capture in the slow flow,* (b) *amplitude modulations,* (c) *comparison between analytical approximation (dashed line) and direct numerical simulation (solid line) for* $v(t)$, (d) *transient responses of the system.*

the responses of system (1) as

$$y(t) = y_1(t) + y_{1/3}(t), \qquad v(t) = v_1(t) + v_{1/3}(t), \tag{7}$$

where the indices represent the frequency of each term. As in the previous case, we introduce new complex variables,

$$\begin{aligned}
&\psi_1(t) = \dot{y}_1(t) + j\omega y_1(t) \equiv \varphi_1(t)\, e^{j\omega t}, \quad \psi_3(t) = \dot{y}_{1/3}(t) + j(\omega/3) y_{1/3}(t) \equiv \varphi_3(t)\, e^{j(\omega/3)t},\\
&\psi_2(t) = \dot{v}_1(t) + j\omega v_1(t) \equiv \varphi_2(t)\, e^{j\omega t}, \quad \psi_4(t) = \dot{v}_{1/3}(t) + j(\omega/3) v_{1/3}(t) \equiv \varphi_4(t)\, e^{j(\omega/3)t},
\end{aligned} \tag{8}$$

where $\varphi_i(t)$ represent slowly varying modulations of fast oscillations of frequencies 1 or 1/3. Expressing the system responses in terms of the new complex variables,

$$y = \frac{\psi_1 - \psi_1^*}{2j\omega} + \frac{3\,(\psi_3 - \psi_3^*)}{2j\omega}, \qquad v = \frac{\psi_2 - \psi_2^*}{2j\omega} + \frac{3\,(\psi_4 - \psi_4^*)}{2j\omega}, \tag{9}$$

substituting into (1) with $P(t) = 0$, and averaging over each of the two fast frequencies, we derive the slow modulation equations that govern the evolutions of the complex amplitudes,

$$\begin{aligned}
&\dot{\varphi}_1 + \left(\frac{j\omega}{2} - \frac{j}{2\omega}\right)\varphi_1 + \left(\frac{\varepsilon\lambda}{2}\right)(2\varphi_1 - \varphi_2) + \left(\frac{jC}{8\omega^3}\right)\Big\{3\Big[9\varphi_3^3 - 27\varphi_3^2\varphi_4 - 9\varphi_4^3\\
&\quad - (\varphi_1 - \varphi_2)\,|\varphi_1 - \varphi_2|^2 + 27\varphi_3\varphi_4^2 - 18\,(\varphi_1 - \varphi_2)\,|\varphi_3 - \varphi_4|^2\Big]\Big\} = 0,\\
&\dot{\varphi}_3 + \left(\frac{j\omega}{6} - \frac{3j}{2\omega}\right)\varphi_3 + \left(\frac{\varepsilon\lambda}{2}\right)(2\varphi_3 - \varphi_4)\\
&\quad + \left(\frac{jC}{8\omega^3}\right)\Big\{-9\Big[\varphi_1\left(2\,(\varphi_3 - \varphi_4)\,(\varphi_1^* - \varphi_2) - 3\,(\varphi_3^* - \varphi_4^*)^2\right)\\
&\quad + \varphi_2\left(2\,(\varphi_4 - \varphi_3)\,(\varphi_1^* - \varphi_2) + 3\,(\varphi_3^* - \varphi_4^*)^2\right) + 9\,(\varphi_3 - \varphi_4)\,|\varphi_3 - \varphi_4|^2\Big]\Big\} = 0,\\
&\dot{\varphi}_2 + \left(\frac{j\omega}{2}\right)\varphi_2 + \left(\frac{\lambda}{2}\right)(\varphi_2 - \varphi_1) - \left(\frac{jC}{\varepsilon 8\omega^3}\right)\Big\{3\Big[9\varphi_3^3 - 27\varphi_3^2\varphi_4 - 9\varphi_4^3\\
&\quad - (\varphi_1 - \varphi_2)\,|\varphi_1 - \varphi_2|^2 + 27\varphi_3\varphi_4^2 - 18\,(\varphi_1 - \varphi_2)\,|\varphi_3 - \varphi_4|^2\Big]\Big\} = 0,\\
&\dot{\varphi}_4 + \left(\frac{j\omega}{6}\right)\varphi_4 + \left(\frac{\lambda}{2}\right)(\varphi_4 - \varphi_3)\\
&\quad - \left(\frac{jC}{\varepsilon 8\omega^3}\right)\Big\{-9\Big[\varphi_1\left(2\,(\varphi_3 - \varphi_4)\,(\varphi_1^* - \varphi_2) - 3\,(\varphi_3^* - \varphi_4^*)^2\right)\\
&\quad + \varphi_2\left(2\,(\varphi_4 - \varphi_3)\,(\varphi_1^* - \varphi_2) + 3\,(\varphi_3^* - \varphi_4^*)^2\right) + 9\,(\varphi_3 - \varphi_4)\,|\varphi_3 - \varphi_4|^2\Big]\Big\} = 0,
\end{aligned} \tag{10}$$

where again it was assumed that $\lambda_1 = \lambda_2 = \lambda$ in (1). To derive a set of real modulation equations, we express the complex amplitudes in polar forms $\varphi_i(t) = a_i(t)\, e^{j\beta_i(t)}$ and derive an autonomous set of seven slow-flow modulation equations that govern the amplitudes $a_i = |\varphi_i|$, $i = 1, \ldots, 4$, and the phase differences $\phi_{12} = \beta_1 - \beta_2$, $\phi_{13} = \beta_1 - 3\beta_3$, and $\phi_{14} = \beta_1 - 3\beta_4$.

The equations of the autonomous slow flow will not be reproduced here, but it suffices to state that they are of the form

$$
\begin{aligned}
&\dot{a}_1 + \left(\frac{\varepsilon\lambda}{2}\right)(2a_1 - a_2) + g_1(a,\phi) = 0,\\
&\dot{a}_3 + \left(\frac{\varepsilon\lambda}{2}\right)(2a_3 - a_4) + g_3(a,\phi) = 0,\\
&\dot{a}_2 + \left(\frac{\lambda}{2}\right)(a_2 - a_1) + \frac{g_2(a,\phi)}{\varepsilon} = 0,\\
&\dot{a}_4 + \left(\frac{\lambda}{2}\right)(a_4 - a_3) + \frac{g_4(a,\phi)}{\varepsilon} = 0,\\
&\dot{\phi}_{12} + f_{12}(a) + g_{12}(a,\phi;\varepsilon) = 0,\\
&\dot{\phi}_{13} + f_{13}(a) + g_{13}(a,\phi) = 0,\\
&\dot{\phi}_{14} + f_{14}(a) + g_{14}(a,\phi;\varepsilon) = 0,
\end{aligned}
\tag{11}
$$

where the functions g_i and g_{ij} are 2π-periodic in terms of the phase angles $\phi = (\phi_{12}\ \phi_{13}\ \phi_{14})^T$ and by a we denote the (4×1) vector of the amplitudes a_i. In this case (as for the fundamental energy pumping mechanism), strong energy transfer between the linear and nonlinear oscillators can occur only when a subset of phase angles ϕ_{ij} does not exhibit time-like behavior; that is, when some phase angles possess oscillatory (nonmonotonic) behavior with respect to time. This can be seen from the structure of the slow flow (11) where, if the phase angles exhibit time-like behavior and the functions g_i are small, averaging over these phase angles can be performed to show that the amplitudes a_i decrease monotonically with time; in that case no significant energy exchanges between the linear and nonlinear components of the system can take place. It follows that *subharmonic energy pumping is associated with non–time-like behavior of (at least) a subset of the slow phase angles* ϕ_{ij} *in* (11).

In Figure 15 we present the results of the numerical integration of the slow-flow (10)–(11) for the system with parameters $\varepsilon = 0.05$, $\lambda = 0.03$, $C = 1$, $\omega_0 = 1$. The motion is initiated on branch $S13-$ with initial conditions $v(0) = y(0) = 0$ and $\dot{v}(0) = 0.01499$, $\dot{y}(0) = -0.059443$ (it corresponds exactly to the simulation of Figure 5). The corresponding initial conditions and the value of the frequency ω of the reduced slow-flow model were computed by minimizing the difference between the analytical and numerical responses of the system in the interval $t \in [0, 100]$: $\varphi_1(0) = -0.0577$, $\varphi_2(0) = 0.0016$, $\varphi_3(0) = -0.0017$, $\varphi_4(0) = 0.0134$, and $\omega = 1.0073$. This result indicates that, initially, nearly all energy is stored in the fundamental frequency component of the linear oscillator, with the remainder confined to the subharmonic frequency component of the NES. In Figures 15(a)–(b) we depict the temporal evolution of the amplitudes a_i, from which we conclude that subharmonic energy pumping in the system is mainly realized through energy transfer from the (fundamental) component at frequency ω of the linear oscillator, to the (subharmonic) component at frequency $\omega/3$ of the NES (as judged from the build-up of the amplitude a_3 and the diminishing of a_1). A smaller amount of energy is transferred from the fundamental frequency component of the linear oscillator to the corresponding fundamental component of the NES (as judged by the evolution of the amplitude a_2).

These conclusions are supported by the plots of Figures 15(c)–(e), where the temporal evolutions of the phase differences $\phi_{12} = \beta_1 - \beta_2$, $\phi_{13} = \beta_1 - 3\beta_3$, and $\phi_{14} = \beta_1 - 3\beta_4$ are shown. Absence of strong energy exchange between the funda-

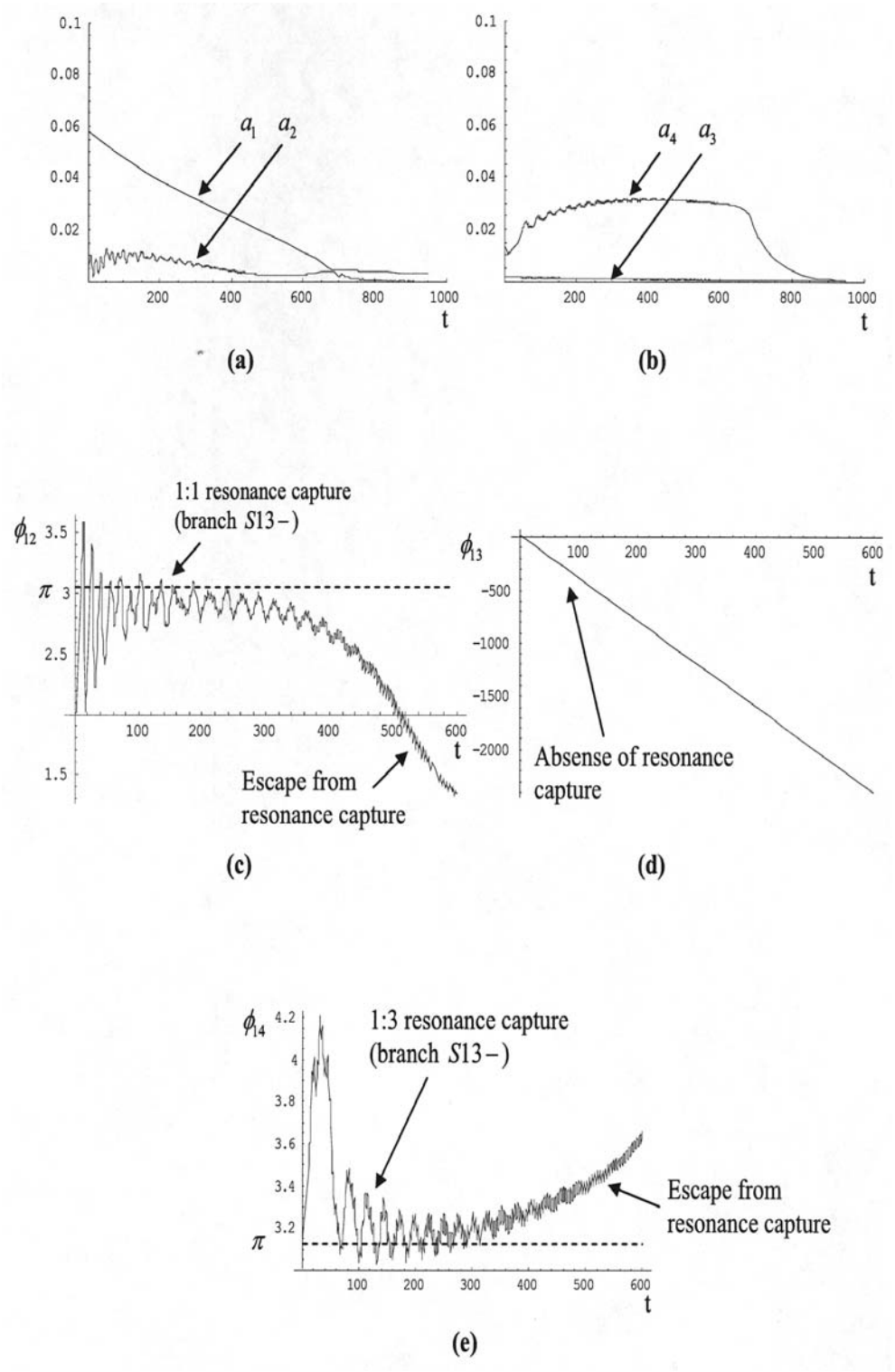


FIG. 15. *Subharmonic energy pumping:* (a), (b) *amplitude modulations;* (c), (d), (e) *phase modulations.*

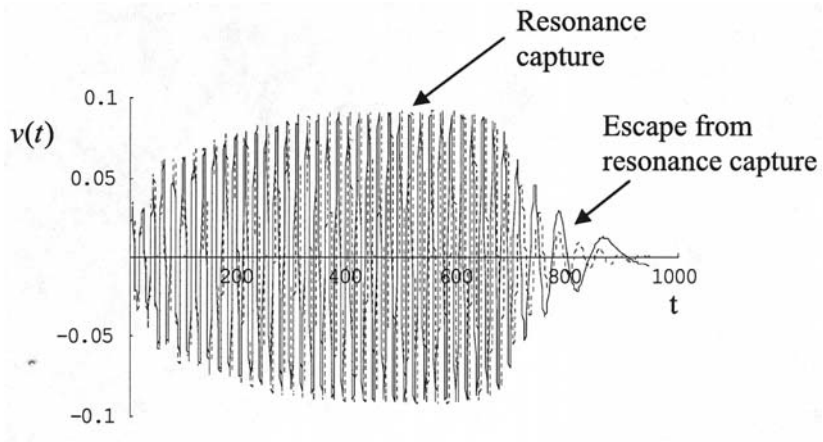


FIG. 16. *Transient response of NES for* 1:3 *subharmonic energy pumping: comparison between analytical approximation (dashed line) and direct numerical simulation (solid line).*

mental and subharmonic frequency components of the linear oscillator is associated with the time-like behavior of the phase difference ϕ_{13}, whereas energy pumping from the fundamental component of the linear oscillator to the two frequency components of the NES is associated with oscillatory early-time behavior of the phase differences ϕ_{12} and ϕ_{14}. Oscillatory responses of ϕ_{12} and ϕ_{14} correspond to 1:1 and 1:3 resonance captures, respectively, between the corresponding frequency components of the linear oscillator and the NES; as time increases, time-like responses of the phase variables are associated with escapes from the corresponding regimes of resonance capture. In addition, we note that the oscillations of the angles ϕ_{12} and ϕ_{14} take place in the neighborhood of π, which confirms that, in this particular example, subharmonic energy pumping is activated by the excitation of an antiphase branch of periodic solutions (such as $S13-$). The analytical results are in full agreement with the wavelet transforms depicted in Figures 5(c)–(d), where the response of the linear oscillator possesses a strong frequency component at the fundamental frequency $\omega_0 = 1$, whereas the NES oscillates mainly at frequency $\omega_0/3$.

The accuracy of the analytical model (10)–(11) in capturing the dynamics of subharmonic energy pumping is confirmed by the plot depicted in Figure 16, where the analytical response of the NES is found to be in satisfactory agreement with the numerical response obtained by the direct simulation of (1). It is interesting to note that the reduced analytical model is capable of accurately modeling the strongly nonlinear, damped, transient response of the NES in the resonance capture region. The analytical model fails, however, during the escape from resonance capture since the ansatz (7)–(8) is not valid in that regime of the motion. Indeed, after escape from resonance capture, the motion approximately evolves along the backbone curve of the frequency-energy plot; eventually $S15$ is reached, the motion of which cannot be described by the ansatz (7)–(8), thereby leading to the failure of the analytical model.

The results presented so far provide a measure of the complicated dynamics encountered in the two-DOF system under consideration. It is logical to assume that by increasing the degrees of freedom of the system the dynamics will be even more complex. That this is indeed the case is evidenced by the numerical simulations presented in the next section, where resonance capture cascades are reported in multi-degree-

of-freedom (MDOF) linear systems with essentially nonlinear end attachments. By resonance capture cascades we denote complicated sudden transitions between different branches of solutions (modes), which are accompanied by sudden changes in the frequency content of the system responses. As shown in previous works (Vakakis et al. (2003)), such multifrequency transitions can drastically enhance energy pumping from the linear system to the essentially nonlinear attachment.

5. Increasing the DOF of the linear system: Resonance capture cascades. To provide an indication of the complex multifrequency transitions that can take place in coupled oscillators with essentially nonlinear local attachments, we now increase the number of DOF of the linear subsystem to two and examine the system

$$\begin{gathered} \ddot{y}_2 + \omega_0^2 y_2 + \lambda_2 \dot{y}_2 + d(y_2 - y_1) = 0, \\ \ddot{y}_1 + \omega_0^2 y_1 + \lambda_1 \dot{y}_1 + \lambda_3(\dot{y}_1 - \dot{v}) + d(y_1 - y_2) + C(y_1 - v)^3 = 0, \\ \varepsilon \ddot{v} + \lambda_3(\dot{v} - \dot{y}_1) + C(v - y_1)^3 = 0. \end{gathered} \tag{12}$$

The system parameters are chosen as $\omega_0^2 = 136.9$, $\lambda_1 = \lambda_2 = 0.155$, $\lambda_3 = 0.544$, $d = 1.2 \times 10^3$, $\varepsilon = 1.8$, and $C = 1.63 \times 10^7$, with linear natural frequencies $\omega_1 = 11.68$ and $\omega_2 = 50.14$.

In Figure 17(a) we depict the relative response $v(t) - y_1(t)$ of the system for initial displacements $y_1(0) = 0.01$, $y_2(0) = v(0) = -0.01$ and zero initial velocities. The multifrequency content of the transient response is evident and is quantified in Figure 17(b), where the instantaneous frequency of the time series is computed by applying the numerical Hilbert transform (Huang et al. (1998)).

As energy decreases due to damping dissipation, a series of eight *resonance capture cascades* is observed, i.e., of transient resonances of the NES with a number of nonlinear modes of the system. The complexity of the nonlinear dynamics of the system is evidenced by the fact that of these eight captures only two (labeled IV and VII in Figure 17(b)) involve the linearized in-phase and out-of-phase modes of the linear oscillator, with the remaining involving essentially nonlinear interactions of the NES with different low- and high-frequency nonlinear modes of the system. During each resonance capture the NES passively absorbs energy from the nonlinear mode involved, before escape from resonance capture occurs and the NES transiently resonates with the next mode in the series. In essence, *the NES acts as a passive, broadband boundary controller*, absorbing, confining, and eliminating vibration energy from the linear oscillator. Similar types of resonance capture cascades were reported in previous works where grounded NESs, weakly coupled to the linear structure, were examined (Vakakis et al. (2003)). The capacity of the NES to resonantly interact with linear and nonlinear modes in different frequency ranges is due to its essential nonlinearity (i.e., the absence of a linear term in the nonlinear stiffness characteristic), which precludes any preferential resonant frequency.

Finally, we note that the complex nonlinear transitions between modes depicted in Figure 17 can be interpreted and understood by studying the topology and bifurcations of periodic orbits of the corresponding undamped system. As shown in previous sections, the weakly damped, forced dynamics is expected to depend on the periodic dynamics of the underlying undamped system.

6. Concluding remarks. Even though the systems considered in this work possess rather simple configurations and small numbers of DOF, they exhibit interesting passive energy transfer properties. Indeed, under rather general conditions, it is possible to transfer passively, irreversibly, and robustly a significant portion of the energy

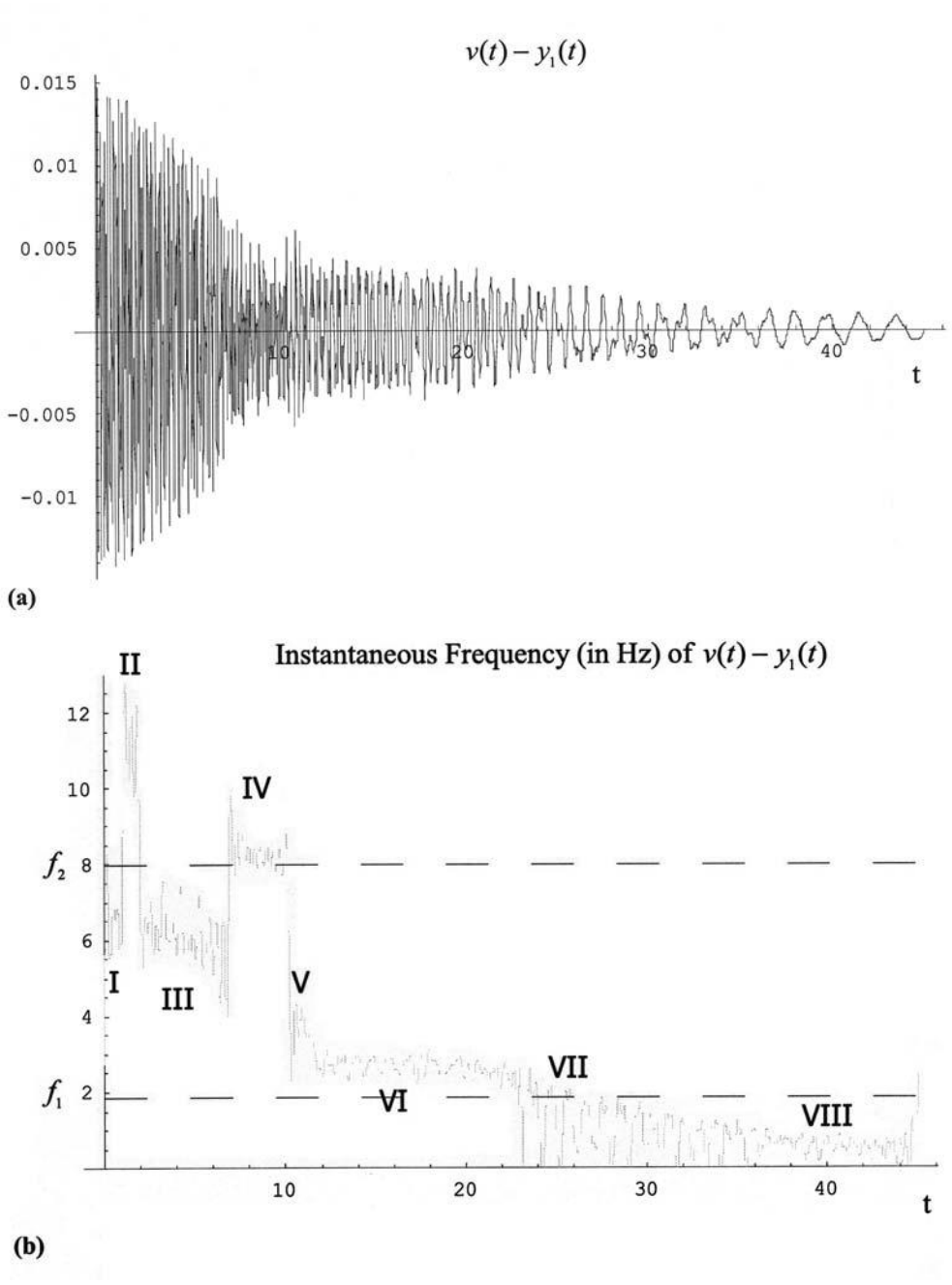


FIG. 17. *Resonance capture cascades in the* 2*-DOF system with nonlinear end attachment:* (a) *relative transient response* $v(t) - y_1(t)$*;* (b) *instantaneous frequency (resonance captures indicated).*

of the linear oscillator to the nonlinear attachment; confine it; and passively dissipate it locally without "radiating back" the transferred energy to the primary system. Moreover, this nonlinear energy pumping occurs over low- as well as high-frequency ranges, and involves broadband disturbances. This last feature clearly distinguishes

the present configuration from previous classical vibration absorber designs, where energy absorption was limited to narrowband disturbances, and the absorbers were effective only in the vicinity of a single frequency.

Three mechanisms for energy pumping were discussed in this work. Two of them rely on resonance capture of the damped dynamics on either fundamental or subharmonic resonant manifolds in phase space. Viewed from a different perspective, in these cases irreversible energy transfer from the linear oscillator to the nonlinear attachment takes place when the dynamics is restricted to a damped nonlinear normal mode invariant manifold, whose mode shape becomes strongly localized to the nonlinear attachment as the energy deceases due to damping dissipation. A third mechanism relies on nonlinear beats to initiate (but not cause) strong energy pumping; these beats act as "bridging orbits" (or "catalysts") for facilitating energy transfer by activating either one of the previously mentioned mechanisms. It is interesting that all these phenomena occur despite the lightness of the nonlinear attachment compared to the linear oscillator and the complete absence of any active (energy source) element in the system.

The considered nonlinear attachment holds promise as an efficient, robust, and modular passive absorbing device for eliminating undesired broadband disturbances of small- or large-scale structures. As such it can find application in diverse problems in engineering and physics, including vibration and shock isolation of machines and structures, seismic mitigation, packaging, and instability (such as limit cycle oscillation or flutter) suppression.

REFERENCES

S. Aubry, S. Kopidakis, A. M. Morgante, and G. P. Tsironis (2001), *Analytic conditions for targeted energy transfer between nonlinear oscillators or discrete breathers*, Phys. B, 296, pp. 222–236.

C. W. Cai, H. C. Chan, and Y. K. Cheung (2000), *Localized modes in a two-degree-coupled periodic system with a nonlinear disordered subsystem*, Chaos Solitons Fractals, 11, pp. 1481–1492.

M. F. Golnaraghi (1991), *Vibration suppression of flexible structures using internal resonance*, Mech. Res. Comm., 18, pp. 135–143.

N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu (1998), *The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 454, pp. 903–995.

K. R. Khusnutdinova and D. E. Pelinovsky (2003), *On the exchange of energy in coupled Klein–Gordon oscillators*, Wave Motion, 38, pp. 1–10.

G. Kopidakis, S. Aubry, and G. P. Tsironis (2001), *Targeted energy transfer through discrete breathers in nonlinear systems*, Phys. Rev. Lett., 87, paper 165501-1.

Y. S. Lee, G. Kerschen, A. F. Vakakis, P. Panagopoulos, L. A. Bergman, and D. M. McFarland (2005), *Complicated dynamics of a linear oscillator with an essentially nonlinear local attachment*, Phys. D, 204, pp. 41–69.

P. Malatkar and A. H. Nayfeh (2003), *On the transfer of energy between widely spaced modes in structures*, Nonlinear Dynam., 31, pp. 225–242.

L. I. Manevitch (1999), *Complex representation of dynamics of coupled oscillators*, in Mathematical Models of Nonlinear Excitations, Transfer Dynamics and Control in Condensed Systems, Kluwer Academic Publishers/Plenum, New York, pp. 269–300.

P. Maniadis, G. Kopidakis, and S. Aubry (2004), *Classical and quantum targeted energy transfer between nonlinear oscillators*, Phys. D, 188, pp. 153–177.

A. M. Morgante, M. Johansson, S. Aubry, and G. Kopidakis (2002), *Breather-phonon resonances in finite-size lattices: "Phantom breathers,"* J. Phys. A, 35, pp. 4999–5021.

A. I. Neishtadt (1997), *Scattering by resonances*, Celest. Mech. Dynam. Astronom., 65, pp. 1–20.

A. I. Neishtadt (1999), *On adiabatic invariance in two-frequency systems*, in Hamiltonian Systems with Three or More Degrees of Freedom, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 533, Kluwer Academic Publishers, Norwell, MA, pp. 193–213.

P. Salemi, M. F. Golnaraghi, and G. R. Heppler (1997), *Active control of forced and unforced structural vibration*, J. Sound Vibration, 208, pp. 15–32.

D. L. Vainchtein, E. V. Rovinsky, L. M. Zelenyi, and A. I. Neishtadt (2004), *Resonances and particle stochastization in nonhomogeneous electromagnetic fields*, J. Nonlinear Sci., 14, pp. 173–205.

A. F. Vakakis and O. Gendelman (2001), *Energy pumping in nonlinear mechanical oscillators* II: *Resonance capture*, J. Appl. Mech., 68, pp. 42–48.

A. F. Vakakis, L. I. Manevitch, Yu. V. Mikhlin, V. N. Pilipchuk, and A. A. Zevin (1996), *Normal Modes and Localization in Nonlinear Systems*, Wiley Interscience, New York.

A. F. Vakakis, L. I. Manevitch, O. Gendelman, and L. A. Bergman (2003), *Dynamics of linear discrete systems connected to local essentially nonlinear attachments*, J. Sound Vibration, 264, pp. 559–577.

S. J. Zhu, Y. F. Zheng, and Y. M. Fu (2004), *Analysis of nonlinear dynamics of a two-degree-of-freedom vibration system with nonlinear damping and nonlinear spring*, J. Sound Vibration, 271, pp. 15–24.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 680–699

SOLUTIONS TO A MODEL WITH NONUNIFORMLY PARABOLIC TERMS FOR PHASE EVOLUTION DRIVEN BY CONFIGURATIONAL FORCES*

HANS-DIETER ALBER† AND PEICHENG ZHU†

Abstract. We prove the existence of solutions global in time to an initial-boundary value problem for a system of partial differential equations, which consists of the equations of linear elasticity and a nonlinear nonuniformly parabolic equation of second order. The problem models the behavior in time of materials with martensitic phase transformations. This model with diffusive phase interfaces was derived from a model with sharp interfaces, whose evolution is driven by configurational forces, and can be considered to be a regularization of that model. Our existence proof, which contributes to the verification of the model, is only valid in one space dimension.

Key words. nonlinear degenerate parabolic equation, existence of solutions, evolution of phase boundaries, martensitic transformations, configurational forces

AMS subject classifications. 35K65, 74N20

DOI. 10.1137/050629951

1. Introduction.

Two main types of phase transformations in solid materials can be distinguished, diffusion dominated and diffusionless transformations. In this article we study a model for the behavior in time of materials with diffusionless transformations. The model has diffusive interfaces and consists of the partial differential equations of linear elasticity coupled to a quasilinear nonuniformly parabolic equation of second order. It is derived in [2, 3] from a sharp interface model for diffusionless phase transitions and can be considered to be a regularization of that model. The physical background, the sharp interface model, and the derivation of the new model are sketched in the appendix. To verify the validity of the new model investigations are necessary, in which not only must simulations be carried out but also the analytical properties of the model must be determined. We contribute to the verification by showing that in the case of one space dimension an initial-boundary value problem to this model has solutions global in time. We first formulate this initial-boundary value problem in the three-dimensional case, reduce it to the one-dimensional case, and conclude the introduction by stating our main result.

Let $\Omega \subset \mathbb{R}^3$ be an open set. It represents the material points of a solid body. The different phases are characterized by the order parameter $S(t,x) \in \mathbb{R}$. A value of $S(t,x)$ near to zero indicates that the material is in the matrix phase at the point $x \in \Omega$ at time t; a value near to one indicates that the material is in the second phase. The other unknowns are the displacement $u(t,x) \in \mathbb{R}^3$ of the material point x at time t and the Cauchy stress tensor $T(t,x) \in \mathcal{S}^3$, where $\mathcal{S}^3$ denotes the set of symmetric 3×3-matrices. The unknowns must satisfy the following quasi-static equations:

*Received by the editors April 25, 2005; accepted for publication (in revised form) September 2, 2005; published electronically January 6, 2006. This work was supported by the Deutsche Forschungsgemeinschaft under grant AL 333/3-2.

http://www.siam.org/journals/siap/66-2/62995.html

†Department of Mathematics, Darmstadt University of Technology, Schlossgartenstraße 7, 64289 Darmstadt, Germany (alber@mathematik.tu-darmstadt.de, zhu@mathematik.tu-darmstadt.de).

$$-\operatorname{div}_x T(t,x) = b(t,x), \tag{1.1}$$

$$T(t,x) = D\big(\varepsilon(\nabla_x u(t,x)) - \bar{\varepsilon} S(t,x)\big), \tag{1.2}$$

$$S_t(t,x) = -c\big(\psi_S(\varepsilon(\nabla_x u(t,x)), S(t,x)) - \nu\Delta_x S(t,x)\big)|\nabla_x S(t,x)| \tag{1.3}$$

for $(t,x) \in (0,\infty)\times\Omega$. The boundary and initial conditions are

$$u(t,x) = \gamma(t,x), \quad S(t,x) = 0, \qquad (t,x) \in [0,\infty)\times\partial\Omega, \tag{1.4}$$

$$S(0,x) = S_0(x), \quad x \in \Omega. \tag{1.5}$$

Here $\nabla_x u$ denotes the 3×3-matrix of first order derivatives of u, the deformation gradient; $(\nabla_x u)^T$ denotes the transposed matrix; and

$$\varepsilon(\nabla_x u) = \frac{1}{2}\left(\nabla_x u + (\nabla_x u)^T\right)$$

is the strain tensor. $\bar{\varepsilon} \in \mathcal{S}^3$ is a given matrix, the misfit strain, and $D : \mathcal{S}^3 \to \mathcal{S}^3$ is a linear, symmetric, positive definite mapping, the elasticity tensor. In the free energy

$$\psi(\varepsilon, S) = \frac{1}{2}\big(D(\varepsilon - \bar{\varepsilon}S)\big)\cdot(\varepsilon - \bar{\varepsilon}S) + \hat{\psi}(S) \tag{1.6}$$

we choose for $\hat{\psi} \in C^2(\mathbb{R}, [0,\infty))$ a double well potential with minima at $S = 0$ and $S = 1$. The scalar product of two matrices is $A\cdot B = \sum a_{ij}b_{ij}$. Also, ψ_S is the partial derivative, $c > 0$ is a constant, and ν is a small positive constant. Given are the volume force $b : [0,\infty)\times\Omega \to \mathbb{R}^3$ and the data $\gamma : [0,\infty)\times\partial\Omega \to \mathbb{R}^3$, $S_0 : \Omega \to \mathbb{R}$.

This completes the formulation of the initial-boundary value problem. The equations (1.1) and (1.2) differ from the system of linear elasticity only by the term $\bar{\varepsilon}S$. The evolution equation (1.3) for the order parameter S is nonuniformly parabolic because of the term $\nu\Delta S|\nabla_x S|$. Since this initial-boundary value problem is derived from a sharp interface model, to verify that it is indeed a diffusive interface model regularizing the sharp interface model, it must be shown that (1.1)–(1.5) with positive ν have global in time solutions, and that these solutions tend to solutions of the sharp interface model for $\nu \to 0$. This would also be a method to prove existence of solutions to the original sharp interface model.

In this article we contribute to only the first part of this program and show that in one space dimension the initial-boundary value problem has solutions. Whether solutions in three space dimensions exist and whether these solutions converge to a solution of the sharp interface model for $\nu \to 0$ is an open problem not investigated here. The model and therefore the existence result is of interest not only in three dimensions but also in one space dimension. However, we believe that this one-dimensional existence result can also be a basic building block in an existence proof for higher space dimensions, which is sketched at the end of this introduction.

Related to our investigations is the model for diffusion dominated phase transformations obtained by coupling the elasticity equations (1.1), (1.2) with the Cahn–Hilliard equation, which is mentioned in the appendix. This model has recently been studied in [5, 7, 8].

Statement of the main result. We now assume that all functions depend only on the variables x_1 and t, and, to simplify the notation, denote x_1 by x. The set $\Omega = (a,d)$ is a bounded open interval with constants $a < d$. We write $Q_{T_e} := (0,T_e)\times\Omega$, where T_e is a positive constant, and define

$$(v,\varphi)_Z = \int_Z v(y)\varphi(y)\,dy$$

for $Z = \Omega$ or $Z = Q_{T_e}$. If v is a function defined on Q_{T_e}, we denote the mapping $x \to v(t,x)$ by $v(t)$. If no confusion is possible, we sometimes drop the argument t and write $v = v(t)$. We still allow that the material points can be displaced in three directions; hence $u(t,x) \in \mathbb{R}^3$, $T(t,x) \in \mathcal{S}^3$, and $S \in \mathbb{R}$. If we denote the first column of the matrix $T(t,x)$ by $T_1(t,x)$ and set

$$\varepsilon(u_x) = \frac{1}{2}((u_x,0,0) + (u_x,0,0)^T) \in \mathcal{S}^3,$$

then with these definitions (1.1)–(1.3) in the case of one space dimension can be written in the form

$$-T_{1x} = b, \tag{1.7}$$

$$T = D(\varepsilon(u_x) - \bar{\varepsilon}S), \tag{1.8}$$

$$S_t = c\left(T \cdot \bar{\varepsilon} - \hat{\psi}'(S) + \nu S_{xx}\right)|S_x|, \tag{1.9}$$

which must be satisfied in Q_{T_e}. Here we have inserted $\psi_S(\varepsilon, S) = -T \cdot \bar{\varepsilon} + \hat{\psi}'(S)$. Since (1.7), (1.8) are linear, the inhomogeneous Dirichlet boundary condition for u can be reduced in the standard way to the homogeneous condition. For simplicity we thus assume that $\gamma = 0$. The initial and boundary conditions therefore are

$$u(t,x) = 0, \qquad (t,x) \in (0,T_e) \times \partial\Omega, \tag{1.10}$$

$$S(t,x) = 0, \qquad (t,x) \in (0,T_e) \times \partial\Omega, \tag{1.11}$$

$$S(0,x) = S_0(x), \quad x \in \Omega. \tag{1.12}$$

To define weak solutions of this initial-boundary value problem we note that because of $\frac{1}{2}(|y|y)' = |y|$, equation (1.9) is equivalent to

$$S_t - c\nu\frac{1}{2}(|S_x|S_x)_x - c\left(T \cdot \bar{\varepsilon} - \hat{\psi}'(S)\right)|S_x| = 0. \tag{1.13}$$

DEFINITION 1.1. *Let $b \in L^\infty(0,T_e,L^2(\Omega))$, $S_0 \in L^\infty(\Omega)$. A function (u,T,S) with*

$$u \in L^\infty(0,T_e;W_0^{1,\infty}(\Omega)), \tag{1.14}$$

$$T \in L^\infty(Q_{T_e}), \tag{1.15}$$

$$S \in L^\infty(Q_{T_e}) \cap L^\infty(0,T_e;H_0^1(\Omega)) \tag{1.16}$$

is a weak solution to the problem (1.7)–(1.12) *if* (1.7), (1.8), (1.10) *are satisfied weakly and if, for all $\varphi \in C_0^\infty((-\infty,T_e) \times \Omega)$,*

$$(S,\varphi_t)_{Q_{T_e}} - c\nu\frac{1}{2}(|S_x|S_x,\,\varphi_x)_{Q_{T_e}} + c\big(\big(T \cdot \bar{\varepsilon} - \hat{\psi}'(S)\big)|S_x|,\,\varphi\big)_{Q_{T_e}} + (S_0,\varphi(0))_\Omega = 0. \tag{1.17}$$

The main result of this article is the following.

THEOREM 1.1. *For all $S_0 \in H_0^1(\Omega)$ and $b \in C(\overline{Q}_{T_e})$ with $b_t \in C(\overline{Q}_{T_e})$ there exists a weak solution (u,T,S) of the problem* (1.7)–(1.12)*, which in addition to* (1.14)–(1.17) *satisfies*

$$S_t \in L^{\frac{4}{3}}(Q_{T_e}), \quad S_x \in L^{\frac{8}{3}}(0,T_e;L^q(\Omega)) \quad \text{for any } 1 < q < \infty \tag{1.18}$$

and

$$(|S_x|S_x)_x \in L^{\frac{4}{3}}(Q_{T_e}), \qquad S_{xt} \in L^{\frac{4}{3}}(0,T_e;W^{-1,\frac{4}{3}}(\Omega)). \tag{1.19}$$

The remaining sections are devoted to the proof of this theorem. The main difficulty in the proof stems from the fact that the coefficient $\nu|S_x|$ of the highest order derivative S_{xx} in (1.9) is not bounded away from zero and that it is not differentiable with respect to S_x.

To prove Theorem 1.1 we therefore consider in section 2 a modified initial-boundary value problem, which consists of (1.7), (1.8), (1.10)–(1.12) and

$$S_t - (c\nu|S_x|_\kappa + \kappa)S_{xx} - c\big(T\cdot\bar{\varepsilon} - \hat{\psi}'(S)\big)|S_x|_\kappa = 0, \quad x\in\Omega,\ t>0, \tag{1.20}$$

with a constant $\kappa > 0$. Here we use the notation

$$|p|_\kappa := \frac{|p|^2}{\sqrt{\kappa^2+|p|^2}}. \tag{1.21}$$

Since (1.20) is a uniformly parabolic equation, we can use a standard theorem to conclude that the modified initial-boundary value problem has a sufficiently smooth solution $(u^\kappa, T^\kappa, S^\kappa)$. For this solution we derive in section 3 a priori estimates independent of κ.

To select a subsequence converging to a solution for $\kappa \to 0$ we need a compactness result. However, our a priori estimates are not strong enough to show that the sequence S_x^κ is compact; instead, we can show only that the sequence $|S_x^\kappa|S_x^\kappa$, or more precisely an approximation to this sequence, has bounded derivatives and thus is compact. It turns out that this is enough to prove existence of a solution. For the compactness proof in section 4 we use the Aubin–Lions lemma; since one of our a priori estimates for derivatives of the approximating sequence is valid only in $L^1(0,T_e;H^{-2}(\Omega))$, we must use the generalized form of this lemma given by Roubíček [14], which is valid in L^1.

The method of proof is limited to one space dimension, since for the a priori estimates it is crucial that the term $|S_x|S_{xx}$ in (1.9) can be written in the form $\frac{1}{2}(|S_x|S_x)_x$. In the higher-dimensional case the corresponding term $|\nabla_x S|\Delta_x S$ cannot be rewritten in this way. Yet, if in the case of two space dimensions new coordinates (r,χ) are chosen such that r is constant on the level curves of the function $(x_1,x_2)\mapsto S(t,x_1,x_2)$ and such that $\nabla_x S(t,x)$ is a tangential vector to the curve $\chi = \text{const}$, then the evolution equation (1.3) can be written in the form

$$S_t = -c\big(\psi_S(\varepsilon(\nabla_x u),S) - \nu S_{rr}\big)|S_r| + c\nu|S_r|S_r K, \tag{1.22}$$

where K is the curvature of the curve $r = \text{const}$. Applying this formula, it might be possible to prove existence in two space dimensions by iteration: Choose the coordinate system (r,χ) to a known approximate solution $S^{(n)}$ and determine the next iterate $S^{(n+1)}$ by solving the initial-boundary value problem to (1.1), (1.2), (1.22), using the methods of our one-dimensional existence proof and considering $g_n = c\nu|S_r^{(n)}|S_r^{(n)}K$ to be a known right-hand side. A solution would be obtained if convergence of this procedure can be shown. The same idea could also be used in three and higher space dimensions.

2. Existence of solutions to the modified problem. In this section, we study the modified initial-boundary value problem and show that it has a Hölder continuous classical solution. To formulate this problem, let $\chi \in C_0^\infty(\mathbb{R}, [0,\infty))$ satisfy $\int_{-\infty}^{\infty} \chi(t)dt = 1$. For $\kappa > 0$, we set

$$\chi_\kappa(t) := \frac{1}{\kappa}\chi\left(\frac{t}{\kappa}\right),$$

and for $S \in L^\infty(Q_{T_e}, \mathbb{R})$ we define

$$(\chi_\kappa * S)(t,x) = \int_0^{T_e} \chi_\kappa(t-s)S(s,x)ds. \tag{2.1}$$

The modified initial-boundary value problem consists of the equations

$$-T_{1x} = b, \tag{2.2}$$

$$T = D(\varepsilon(u_x) - \bar{\varepsilon}\chi_\kappa * S), \tag{2.3}$$

$$S_t = (c\nu|S_x|_\kappa + \kappa)S_{xx} + c(T\cdot\bar{\varepsilon} - \hat{\psi}'(S))|S_x|_\kappa\,, \tag{2.4}$$

which must hold in Q_{T_e}, and of the boundary and initial conditions

$$u(t,x) = 0, \quad (t,x) \in (0,T_e)\times\partial\Omega, \tag{2.5}$$

$$S(t,x) = 0, \quad (t,x) \in (0,T_e)\times\partial\Omega, \tag{2.6}$$

$$S(0,x) = S_0(x), \quad x \in \Omega. \tag{2.7}$$

To formulate an existence theorem for this problem we need some function spaces: For nonnegative integers m, n and a real number $\alpha \in (0,1)$ we denote by $C^{m+\alpha}(\overline{\Omega})$ the space of m-times differentiable functions on $\overline{\Omega}$, whose mth derivative is Hölder continuous with exponent α. The space $C^{\alpha,\alpha/2}(\overline{Q}_{T_e})$ consists of all functions on $\overline{Q}_{T_e}$, which are Hölder continuous in the parabolic distance

$$\mathrm{d}((t,x),(s,y)) := \sqrt{|t-s| + |x-y|^2}.$$

$C^{m,n}(\overline{Q}_{T_e})$ and $C^{m+\alpha,n+\alpha/2}(\overline{Q}_{T_e})$, respectively, are the spaces of functions, whose x-derivatives up to order m and t-derivatives up to order n belong to $C(\overline{Q}_{T_e})$ or to $C^{\alpha,\alpha/2}(\overline{Q}_{T_e})$, respectively.

THEOREM 2.1. *Let $\nu, \kappa > 0$, $T_e > 0$; suppose that the function $b \in C(\overline{Q}_{T_e})$ has the derivative $b_t \in C(\overline{Q}_{T_e})$ and that the initial data $S_0 \in C^{2+\alpha}(\overline{\Omega})$ satisfy $S_0|_{\partial\Omega} = S_{0,x}|_{\partial\Omega} = S_{0,xx}|_{\partial\Omega} = 0$. Then there is a solution*

$$(u,T,S) \in C^{2,1}(\overline{Q}_{T_e}) \times C^{1,1}(\overline{Q}_{T_e}) \times C^{2+\alpha,1+\alpha/2}(\overline{Q}_{T_e})$$

of the modified initial-boundary value problem (2.2)–(2.7). *This solution satisfies $S_{tx} \in L^2(Q_{T_e})$ and*

$$\max_{\overline{Q}_{T_e}} |S| \le \max_{\overline{\Omega}} |S_0|. \tag{2.8}$$

Proof. Note first that if S is given, then for every t, (2.2), (2.3), (2.5) form a linear elliptic boundary value problem for the unknown function $x \mapsto (u(t,x), T(t,x))$. In

[3] it is shown that the unique solution is given by

$$u(t,x) = u^*\left(\int_a^x (\chi_\kappa * S)(t,y)dy - \frac{x-a}{d-a}\int_a^d (\chi_\kappa * S)(t,y)dy\right) + w(t,x), \tag{2.9}$$

$$T(t,x) = D(\varepsilon^* - \bar\varepsilon)(\chi_\kappa * S)(t,x) - \frac{D\varepsilon^*}{d-a}\int_a^d (\chi_\kappa * S)(t,y)dy + \sigma(t,x), \tag{2.10}$$

where $u^* \in \mathbb{R}^3, \varepsilon^* \in \mathcal{S}^3$ are suitable constants depending only on $\bar\varepsilon$ and D, and where for every $t \in [0,T_e]$ the function $(w(t),\sigma(t)) : \Omega \to \mathbb{R}^3 \times \mathcal{S}^3$ is the solution to the boundary value problem

$$\begin{aligned} -\sigma_{1x}(t) &= b(t),\\ \sigma(t) &= D\varepsilon(w_x(t)),\\ w(t)_{|\partial\Omega} &= 0. \end{aligned}$$

Since by assumption b and b_t belong to $C(\bar Q_{T_e})$, it follows that $(w,\sigma) \in C^{2,1}(\bar Q_{T_e}) \times C^{1,1}(\bar Q_{T_e})$. We insert (2.10) into (2.4) and obtain the equation

$$S_t = a_1(S_x)S_{xx} + a_2\left(t,x,S,S_x,\chi_\kappa * S, \frac{1}{d-a}\int_a^d (\chi_\kappa * S)(t,y)dy\right) \tag{2.11}$$

in Q_{T_e}, where

$$a_1(p) = c\nu|p|_\kappa + \kappa$$

and

$$a_2(t,x,S,p,r,s) = c\big(\bar\varepsilon \cdot D(\varepsilon^* - \bar\varepsilon)r - \bar\varepsilon \cdot D\varepsilon^* s + \bar\varepsilon \cdot \sigma(t,x) - \hat\psi'(S)\big)|p|_\kappa .$$

The equations (2.11), (2.6), and (2.7) form an initial-boundary value problem with nonlocal terms, which is equivalent to the problem (2.2)–(2.7). To prove Theorem 2.1 it therefore suffices to show that this initial-boundary value problem is solvable. This follows from the following claim.

THEOREM 2.2. *Let $T_e > 0$, $M > 0$, and suppose that the coefficient functions $a_1 \in C^1(\mathbb{R},[0,\infty))$ and $a_2 \in C^1(\overline{Q}_{T_e} \times [-M,M] \times \mathbb{R} \times [-M,M]^2, \mathbb{R})$ satisfy the equations and inequalities*

$$a_2(t,x,S,0,r,s) = 0, \tag{2.12}$$

$$\mu_1(1+|p|)^{m-2} \le a_1(p) \le \mu_2(1+|p|)^{m-2}, \tag{2.13}$$

$$\left|\frac{\partial a_1}{\partial p}\right|(1+|p|)^3 + \left|\frac{\partial a_2}{\partial p}\right|(1+|p|) + |a_2| \le \mu_3(1+|p|)^m, \tag{2.14}$$

$$\left|\frac{\partial a_2}{\partial x}\right| \le (\mu_4 + P(|p|))(1+|p|)^{m+1}, \tag{2.15}$$

$$\max\left(\frac{\partial a_2}{\partial S}, \frac{\partial a_2}{\partial r}, \frac{\partial a_2}{\partial s}\right) \le (\mu_4 + P(|p|))(1+|p|)^m, \tag{2.16}$$

where $P(\rho)$ is a nonnegative continuous function that tends to zero for $\rho \to \infty$, $\mu_1, \dots, \mu_4$ are positive constants, and m is an arbitrary number.

If the number μ_4 is sufficiently small, depending on the numbers M, μ_1, μ_2, μ_3, and $\hat{P} = \max_{\rho \geq 0} P(\rho)$, and if the initial data $S_0 \in C^{2+\alpha}(\bar{\Omega}, [-M, M])$ satisfy the compatibility conditions $S_0|_{\partial\Omega} = 0$ and

$$a_1(S_{0,x}(x))S_{0,xx}(x) + a_2(0, x, S_0(x), S_{0,x}(x), r, s) = 0 \tag{2.17}$$

for all $x \in \partial\Omega$ and for all $-M \leq r, s \leq M$, then a solution $S \in C^{2+\alpha,1+\alpha/2}(\overline{Q}_{T_e})$ of the problem (2.11), (2.6), *and* (2.7) *exists. This solution has derivatives $S_{tx} \in L^2(Q_{T_e})$.*

A proof of Theorem 2.2 is obtained by modification of the proof of the analogous Theorem 5.2 in [11, p. 564], which is valid for the quasi-linear parabolic initial boundary value problem

$$\begin{aligned} S_t &= a_1(S_x)S_{xx} + a_2(t, x, S, S_x), \\ S(t, x) &= 0, \qquad (t, x) \in (0, T_e) \times \partial\Omega, \\ S(0, x) &= S_0(x), \qquad x \in \Omega, \end{aligned}$$

which does not contain nonlocal terms. The theorem in [11] states that if the coefficient functions satisfy the conditions (2.12)–(2.16) and if the initial data satisfy compatibility and regularity conditions analogous the ones given above, then this initial-boundary value problem has a solution S with the regularity stated in Theorem 2.2. Actually, in [11] more general coefficient functions are considered. The proof is based on the Leray–Schauder fixed point theorem. We leave the modification, which is technical, to the reader.

End of the proof of Theorem 2.1. It is immediately seen that the coefficients a_1 and a_2 in (2.11) satisfy the relations (2.12)–(2.16) with $m = 3$. In particular, we can choose μ_1, μ_2, μ_3, $\hat{P}$ such that the inequalities (2.15), (2.16) hold for every $\mu_4 > 0$, with a suitable function P depending on μ_4. Moreover, from the assumption $S_0|_{\partial\Omega} = S_{0,x}|_{\partial\Omega} = S_{0,xx}|_{\partial\Omega} = 0$ together with (2.12) it follows that the compatibility condition (2.17) holds. Thus, Theorem 2.2 asserts that a solution $S \in C^{2+\alpha,1+\alpha/2}(\overline{Q}_{T_e})$ of (2.11), (2.6), (2.7) exists with $S_{xt} \in L^2(Q_{T_e})$. The functions T and u with the regularity stated in Theorem 2.1 are obtained from (2.10), (2.9). Finally, since $a_1(0) = \kappa > 0$ and $a_2(t, x, S, 0, r, s) = 0$, we can apply [11, Theorem 2.9, p. 23] to (2.11) and conclude that the estimate (2.8) holds. □

3. A priori estimates. In this section we establish a priori estimates for solutions of the modified problem, which are uniform with respect to κ. We remark that the estimates in Lemma 3.1 and Corollary 1, though stated in the one-dimensional case, can be generalized to higher space dimensions.

In what follows we assume that

$$0 < \kappa \leq 1, \tag{3.1}$$

since we consider the limit $\kappa \to 0$. The $L^2(\Omega)$-norm is denoted by $\|\cdot\|$, and the letter C stands for varies positive constants independent of κ. Supplementing (1.21), we also use the notation

$$[p]_\kappa := \frac{p|p|}{\sqrt{\kappa^2 + p^2}}. \tag{3.2}$$

We start by constructing a family of approximate solutions to the modified problem. To this end let T_e be a fixed positive number, and choose for every κ a function $S_0^\kappa \in C_0^\infty(\Omega)$ such that

$$\|S_0^\kappa - S_0\|_{H_0^1(\Omega)} \to 0, \quad \kappa \to 0, \tag{3.3}$$

where $S_0 \in H_0^1(\Omega)$ are the initial data given in Theorem 1.1. We insert for S_0 in (2.7) the function S_0^κ and choose for b in (2.2) the function given in Theorem 1.1. These functions satisfy the assumptions of Theorem 2.1, and hence there is a solution $(u^\kappa, T^\kappa, S^\kappa)$ of the modified problem (2.2)–(2.7), which exists in Q_{T_e}. The inequality (2.8) and Sobolev's embedding theorem yield for this solution

$$\sup_{0<\kappa\le 1} \|S^\kappa\|_{L^\infty(Q_{T_e})} \le \sup_{0<\kappa\le 1} \|S_0^\kappa\|_{L^\infty(\Omega)} \le C. \tag{3.4}$$

Remembering that σ in (2.10) belongs to $C^{1,1}(\bar{Q}_{T_e})$, we conclude from (3.4) that also

$$\max_{\overline{Q}_{T_e}} |c(T^\kappa \cdot \bar{\varepsilon} - \hat{\psi}'(S^\kappa))| \le C. \tag{3.5}$$

LEMMA 3.1. *There holds for any $t \in [0, T_e]$*

$$\|S_x^\kappa(t)\|^2 + \int_0^t \int_\Omega (\nu |S_x^\kappa|_\kappa + 2\kappa)\, |S_{xx}^\kappa|^2 dx d\tau \le C. \tag{3.6}$$

Proof. Observe first that $S_{tx}^\kappa \in L^2(Q_{T_e})$, by Theorem 2.1, which yields that for almost all t

$$\frac{1}{2}\frac{d}{dt}\|S_x^\kappa(t)\|^2 = \int_\Omega S_x^\kappa(t) S_{xt}^\kappa(t) dx.$$

Using this relation and (3.5), we obtain, by multiplication of (2.4) by $-S_{xx}^\kappa$ and integration by parts with respect to x, where we take the boundary condition (2.6) into account, that for almost all t

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|S_x^\kappa\|^2 + \int_\Omega (\nu |S_x^\kappa|_\kappa + \kappa)\, |S_{xx}^\kappa|^2 dx = \int_\Omega c(\hat{\psi}'(S^\kappa) - T^\kappa \cdot \bar{\varepsilon}) |S_x^\kappa|_\kappa S_{xx}^\kappa dx \\
&\le C \int_\Omega |S_x^\kappa|_\kappa |S_{xx}^\kappa| dx = C \int_\Omega |S_x^\kappa|^{\frac{1}{2}}_\kappa |S_x^\kappa|^{\frac{1}{2}}_\kappa |S_{xx}^\kappa| dx \\
&\le \frac{\nu}{2} \int_\Omega |S_x^\kappa|_\kappa |S_{xx}^\kappa|^2 dx + \frac{2C^2}{\nu} \int_\Omega (|S_x^\kappa|_\kappa)^2 dx. \qquad (3.7)
\end{aligned}$$

We subtract the term $\frac{\nu}{2}\int_\Omega |S_x^\kappa|_\kappa |S_{xx}^\kappa|^2 dx$ on both sides of this inequality and use Gronwall's lemma to derive (3.6) from the resulting estimate, noting also (3.3). ∎

COROLLARY 1. *There holds for any $t \in [0, T_e]$*

$$\int_0^t \int_\Omega (|S_x^\kappa|_\kappa |S_{xx}^\kappa|)^{\frac{4}{3}}\, dx d\tau \le C. \tag{3.8}$$

Proof. By Hölder's inequality, we have for some $2 > p \ge 1$, $q = \frac{2}{p}$, and $\frac{1}{q} + \frac{1}{q'} = 1$

that

$$\begin{aligned}
&\int_0^t \int_\Omega (|S_x^\kappa|_\kappa |S_{xx}^\kappa|)^p \, dx d\tau \\
&= \int_0^t \int_\Omega (|S_x^\kappa|_\kappa)^{\frac{p}{2}} \left((|S_x^\kappa|_\kappa)^{\frac{p}{2}} |S_{xx}^\kappa|^p \right) dx d\tau \\
&\leq \left(\int_0^t \int_\Omega (|S_x^\kappa|_\kappa)^{\frac{pq'}{2}} \, dx d\tau \right)^{\frac{1}{q'}} \left(\int_0^t \int_\Omega (|S_x^\kappa|_\kappa)^{\frac{pq}{2}} |S_{xx}^\kappa|^{pq} dx d\tau \right)^{\frac{1}{q}} \\
&\leq \left(\int_0^t \int_\Omega (|S_x^\kappa|_\kappa)^{\frac{p}{2-p}} \, dx d\tau \right)^{\frac{2-p}{2}} \left(\int_0^t \int_\Omega |S_x^\kappa|_\kappa |S_{xx}^\kappa|^2 dx d\tau \right)^{\frac{p}{2}}. \qquad (3.9)
\end{aligned}$$

Inequality (3.6) implies for $\frac{p}{2-p} \leq 2$, i.e., $p \leq \frac{4}{3}$, that the right-hand side of (3.9) is bounded. □

LEMMA 3.2. *There hold*

$$\int_0^t \int_\Omega \left| ([S_x^\kappa]_\kappa |S_x^\kappa|_\kappa)_x \right|^{\frac{4}{3}} dx d\tau \leq 2^{\frac{8}{3}} \int_0^t \int_\Omega ||S_x^\kappa|_\kappa S_{xx}^\kappa|^{\frac{4}{3}} \, dx d\tau \leq C, \qquad (3.10)$$

$$\int_0^t \| |S_x^\kappa|_\kappa \|_{L^\infty(\Omega)}^{\frac{8}{3}} \, d\tau = \int_0^t \| [S_x^\kappa]_\kappa |S_x^\kappa|_\kappa \|_{L^\infty(\Omega)}^{\frac{4}{3}} \, d\tau \leq C. \qquad (3.11)$$

Proof. We first show that (3.11) is a consequence of (3.10). Equation (3.10) and the Poincaré inequality imply

$$\begin{aligned}
\int_0^t \| \, [S_x^\kappa]_\kappa |S_x^\kappa|_\kappa - \overline{[S_x^\kappa]_\kappa |S_x^\kappa|_\kappa} \|_{L^{\frac{4}{3}}}^{\frac{4}{3}} d\tau &\leq C \int_0^t \int_\Omega \left| ([S_x^\kappa]_\kappa |S_x^\kappa|_\kappa)_x \right|^{\frac{4}{3}} dx d\tau \\
&\leq C, \qquad (3.12)
\end{aligned}$$

where for the function $f = f(t,x)$ we have used the notation

$$\bar{f}(t) = \frac{1}{|\Omega|} \int_\Omega f(t,x) dx.$$

$|\Omega|$ is the volume of the domain Ω. Equations (1.21) and (3.2) imply $|S_x^\kappa|_\kappa \leq |S_x^\kappa|$ and $|\, [S_x^\kappa]_\kappa \,| \leq |S_x^\kappa|$. From Lemma 3.1 we thus conclude

$$\begin{aligned}
\int_0^t \int_\Omega \left| \overline{[S_x^\kappa]_\kappa |S_x^\kappa|_\kappa} \right|^{\frac{4}{3}} dx d\tau &\leq \frac{1}{|\Omega|^{\frac{4}{3}}} \int_0^t \int_\Omega \left(\int_\Omega |S_x^\kappa|^2 dx \right)^{\frac{4}{3}} dx d\tau \\
&= \frac{1}{|\Omega|^{\frac{1}{3}}} \int_0^t \|S_x^\kappa\|^{\frac{8}{3}} d\tau \leq \int_0^t C d\tau \leq Ct. \qquad (3.13)
\end{aligned}$$

Combination of the above two inequalities yields

$$\int_0^t \| [S_x^\kappa]_\kappa |S_x^\kappa|_\kappa \|_{L^{\frac{4}{3}}(\Omega)}^{\frac{4}{3}} \, d\tau \leq C. \qquad (3.14)$$

Invoking (3.10), we assert that

$$[S_x^\kappa]_\kappa |S_x^\kappa|_\kappa \in L^{\frac{4}{3}}(0, T_e; W^{1,\frac{4}{3}}(\Omega)),$$

whence by the Sobolev imbedding theorem we obtain

$$\int_0^t \| [S_x^\kappa]_\kappa |S_x^\kappa|_\kappa \|_{L^\infty(\Omega)}^{\frac{4}{3}} d\tau \le C \int_0^t \| [S_x^\kappa]_\kappa |S_x^\kappa|_\kappa \|_{W^{1,\frac{4}{3}}(\Omega)}^{\frac{4}{3}} d\tau \le C. \tag{3.15}$$

Thus (3.11) is proved, and it remains to verify (3.10).

To simplify the notation in the following computation we write $y = S_x^\kappa$. Using that

$$\begin{aligned} ([y]_\kappa |y|_\kappa)_x &= \left(\frac{y^3|y|}{\kappa^2+y^2} \right)_x = \frac{2|y|^3(2\kappa^2+y^2)}{(\kappa^2+y^2)^2} y_x \\ &= |y|_\kappa \frac{2|y|(2\kappa^2+y^2)}{(\kappa^2+y^2)^{\frac{3}{2}}} y_x, \end{aligned} \tag{3.16}$$

we obtain from Young's inequality that

$$\begin{aligned} \left|([y]_\kappa |y|_\kappa)_x\right| &= |y|_\kappa \frac{2|y|(\kappa^2+y^2)}{(\kappa^2+y^2)^{\frac{3}{2}}} |y_x| \\ &\le 2|y|_\kappa \frac{\frac{1}{3}|y|^3 + \frac{2}{3}(\kappa^2+y^2)^{\frac{3}{2}}}{(\kappa^2+y^2)^{\frac{3}{2}}} |y_x| \\ &= 2|y|_\kappa \frac{\frac{1}{3}|y|^{2\,\frac{3}{2}} + \frac{2}{3}(\kappa^2+y^2)^{\frac{3}{2}}}{(\kappa^2+y^2)^{\frac{3}{2}}} |y_x| \\ &\le 2|\,|y|_\kappa\, y_x| = 2|\,|S_x^\kappa|_\kappa\, S_{xx}^\kappa|. \end{aligned} \tag{3.17}$$

Therefore, from (3.8) we have

$$\int_0^t \int_\Omega \left|([S_x^\kappa]_\kappa |S_x^\kappa|_\kappa)_x\right|^{\frac{4}{3}} dx d\tau \le \int_0^t \int_\Omega |2\,|S_x^\kappa|_\kappa\, S_{xx}^\kappa|^{\frac{4}{3}} dx d\tau \le 2^{\frac{4}{3}} C, \tag{3.18}$$

which is (3.10) and completes the proof of this lemma. □

LEMMA 3.3. *The function S_t^κ belongs to $L^{\frac{4}{3}}(Q_{T_e})$, and we have the estimates*

$$\|S_t^\kappa\|_{L^{4/3}(Q_{T_e})} \le C, \tag{3.19}$$

$$\|S_x^\kappa S_{xt}^\kappa\|_{L^1(0,T_e;H^{-2}(\Omega))} \le C, \tag{3.20}$$

$$\| ([S_x^\kappa]_\kappa |S_x^\kappa|_\kappa)_t \|_{L^1(0,T_e;H^{-2}(\Omega))} \le C. \tag{3.21}$$

Proof. From (2.4) and the estimates (3.6), (3.5), and (3.8) we immediately see that $S_t^\kappa \in L^{\frac{4}{3}}(Q_{T_e})$ and that (3.19) holds. Therefore we need to prove only the remaining two estimates.

To prove the first one we show that there is a constant C, which is independent of κ, such that

$$\left|(S_x^\kappa S_{xt}^\kappa, \varphi)_{Q_{T_e}}\right| \le C \|\varphi\|_{L^\infty(0,T_e;H_0^2(\Omega))} \tag{3.22}$$

for all $\varphi \in L^\infty(0,T_e;H_0^2(\Omega))$. This estimate implies (3.20), since $L^1(0,T_e;H^{-2}(\Omega))$ is isometrically imbedded into the dual space of $L^\infty(0,T_e;H_0^2(\Omega))$.

For the proof of (3.22) recall first that $S_{xt}^\kappa \in L^2(Q_{T_e})$, which implies that the right-hand side is well defined. We integrate by parts to get

$$(S_x^\kappa S_{xt}^\kappa, \varphi)_{Q_{T_e}} = (S_t^\kappa, -S_{xx}^\kappa \varphi)_{Q_{T_e}} + (S_t^\kappa, -S_x^\kappa \varphi_x)_{Q_{T_e}} =: I_1 + I_2. \tag{3.23}$$

To estimate I_1 we apply (2.4) and obtain

$$(S_t^\kappa, -S_{xx}^\kappa\varphi)_{Q_{T_e}} = \big((c\nu|S_x^\kappa|_\kappa+\kappa)S_{xx}^\kappa + c\big(T\cdot\bar\varepsilon' - \hat\psi'(S^\kappa)\big)|S_x^\kappa|_\kappa, -S_{xx}^\kappa\varphi\big)_{Q_{T_e}}. \tag{3.24}$$

We estimate the right-hand side of this equation term by term. For the first term we obtain from Lemma 3.1

$$\begin{aligned}\Big|((c\nu|S_x^\kappa|_\kappa+\kappa)S_{xx}^\kappa, -S_{xx}^\kappa\varphi)_{Q_{T_e}}\Big| &\le \|\varphi\|_{L^\infty(Q_{T_e})}\int_{Q_{T_e}}(c\nu|S_x^\kappa|_\kappa+\kappa)|S_{xx}^\kappa|^2 d(x,\tau)\\ &\le C\|\varphi\|_{L^\infty(Q_{T_e})} \le C\|\varphi\|_{L^\infty(0,T_e;H_0^2(\Omega))}.\end{aligned} \tag{3.25}$$

For the second term it follows from (3.5) and (3.8) that

$$\begin{aligned}\big|\big(c\big(T\cdot\bar\varepsilon' - \hat\psi'(S^\kappa)\big)|S_x^\kappa|_\kappa, -S_{xx}^\kappa\varphi\big)_{Q_{T_e}}\big| &\le C\|\varphi\|_{L^\infty(Q_{T_e})}\int_0^{T_e}\| |S_x^\kappa|_\kappa S_{xx}^\kappa\|_{L^1}d\tau\\ &\le C\|\varphi\|_{L^\infty(0,T_e;H_0^2(\Omega))}.\end{aligned} \tag{3.26}$$

The estimates (3.25) and (3.26) together yield

$$|I_1| \le C\|\varphi\|_{L^\infty(0,T_e;H_0^2(\Omega))}. \tag{3.27}$$

Now we estimate I_2. From (2.4) and (3.5) we have

$$\begin{aligned} I_2 &= \big|(S_t^\kappa, S_x^\kappa\varphi_x)_{Q_{T_e}}\big|\\ &= \big|\big((c\nu|S_x^\kappa|_\kappa+\kappa)S_{xx}^\kappa + c\big(T\cdot\bar\varepsilon' - \hat\psi'(S^\kappa)\big)|S_x^\kappa|_\kappa, -S_x^\kappa\varphi_x\big)_{Q_{T_e}}\big|\\ &\le C\int_{Q_{T_e}}\big(|S_x^\kappa|_\kappa|S_{xx}^\kappa| + \kappa|S_{xx}^\kappa| + |S_x^\kappa|_\kappa\big)|S_x^\kappa\varphi_x|d(x,t)\\ &=: C(I_{2,1}+I_{2,2}+I_{2,3}).\end{aligned} \tag{3.28}$$

We are now going to deal with $I_{2,1}, I_{2,2}$, and $I_{2,3}$. Using the Cauchy–Schwarz inequality and invoking the estimates (3.6), (3.8), and (3.11), we arrive at

$$\begin{aligned} I_{2,1} &= \int_{Q_{T_e}}|S_x^\kappa|_\kappa|S_{xx}^\kappa S_x^\kappa\varphi_x|d(x,t)\\ &\le C\int_0^{T_e}\||S_x^\kappa|_\kappa\|_{L^\infty(\Omega)}^{\frac12}\|\varphi_x\|_{L^\infty(\Omega)}\int_\Omega(|S_x^\kappa|_\kappa)^{\frac12}\,|S_{xx}^\kappa S_x^\kappa|dxd\tau\\ &\le C\int_0^{T_e}\||S_x^\kappa|_\kappa\|_{L^\infty(\Omega)}^{\frac12}\,\|\varphi_x\|_{L^\infty(\Omega)}\left(\int_\Omega|S_x^\kappa|^2dx\right)^{\frac12}\left(\int_\Omega\big(|S_x^\kappa|_\kappa|S_{xx}^\kappa|^2\big)\,dx\right)^{\frac12}d\tau\\ &\le C\|\varphi\|_{L^\infty(0,T;H_0^2(\Omega))}\int_0^{T_e}\||S_x^\kappa|_\kappa\|_{L^\infty(\Omega)}^{\frac12}\left\|(|S_x^\kappa|_\kappa)^{\frac12}S_{xx}^\kappa\right\|_{L^2}d\tau\\ &\le C\|\varphi\|_{L^\infty(0,T;H_0^2(\Omega))}\left(\int_0^{T_e}\||S_x^\kappa|_\kappa\|_{L^\infty(\Omega)}d\tau\right)^{\frac12}\left(\int_0^{T_e}\left\|(|S_x^\kappa|_\kappa)^{\frac12}\,S_{xx}^\kappa\right\|_{L^2(\Omega)}^2d\tau\right)^{\frac12}\\ &\le C\|\varphi\|_{L^\infty(0,T;H_0^2(\Omega))}.\end{aligned} \tag{3.29}$$

The other terms are easier to handle. It follows from the estimate (3.6) and the assumption $0 < \kappa \le 1$ that

$$\begin{aligned} I_{2,2} &= \int_{Q_{T_e}} \kappa |S^\kappa_{xx} S^\kappa_x \varphi_x| d(\tau, x) \\ &\le C\kappa^{\frac{1}{2}} \|\varphi_x\|_{L^\infty(Q_T)} \int_{Q_{T_e}} \kappa^{\frac{1}{2}} |S^\kappa_{xx}| |S^\kappa_x| d(\tau, x) \\ &\le C\kappa^{\frac{1}{2}} \|\varphi_x\|_{L^\infty(Q_T)} \left(\int_{Q_{T_e}} \kappa |S^\kappa_{xx}|^2 dx \right)^{\frac{1}{2}} \left(\int_{Q_{T_e}} |S^\kappa_x|^2 d(\tau, x) \right)^{\frac{1}{2}} \\ &\le C \|\varphi\|_{L^\infty(0,T;H^2_0(\Omega))}. \end{aligned} \tag{3.30}$$

Finally, (3.6) and the fact that $|S^\kappa_x|_\kappa \le |S^\kappa_x|$ imply

$$\begin{aligned} I_{2,3} &= \int_{Q_{T_e}} |S^\kappa_x|_\kappa |S^\kappa_x \varphi_x| d(t, x) \\ &\le C \|\varphi_x\|_{L^\infty(Q_{T_e})} \int_{Q_{T_e}} |S^\kappa_x|^2 d(t, x) \\ &\le C \|\varphi\|_{L^\infty(0,T_e;H^2_0(\Omega))}. \end{aligned} \tag{3.31}$$

The estimates (3.28)–(3.31) yield

$$|I_2| \le C \|\varphi\|_{L^\infty(0,T;H^2_0(\Omega))}.$$

This inequality and (3.23), (3.27) together yield the desired estimate (3.22); hence (3.20) follows.

To prove the third statement of the lemma we define

$$\mathcal{R}_\kappa := (c\nu |S^\kappa_x|_\kappa + \kappa) S^\kappa_{xx} + c\big(T \cdot \bar{\varepsilon}' - \hat{\psi}'(S^\kappa)\big) |S^\kappa_x|_\kappa \tag{3.32}$$

and set $y = S^\kappa_x$. Remembering that $S^\kappa_{xt} \in L^2(Q_{T_e})$, we obtain as in (3.16) that

$$([y]_\kappa |y|_\kappa)_t = \frac{2|y|^3(2\kappa^2 + y^2)}{(\kappa^2 + y^2)^2} y_t. \tag{3.33}$$

Multiply (2.4) by

$$\left(S^\kappa_x \varphi \frac{y|y|(2\kappa^2 + y^2)}{(\kappa^2 + y^2)^2} \right)_x = \left(\varphi \frac{|y|^3(2\kappa^2 + y^2)}{(\kappa^2 + y^2)^2} \right)_x,$$

integrate the resulting equation with respect to (x, t) over Q_{T_e}, and note (3.33) to obtain

$$\begin{aligned} 0 &= \left(S^\kappa_t - \mathcal{R}_\kappa, \left(S^\kappa_x \varphi \frac{y|y|(2\kappa^2 + y^2)}{(\kappa^2 + y^2)^2} \right)_x \right)_{Q_{T_e}} \\ &= -\left(S^\kappa_{xt}, S^\kappa_x \varphi \frac{y|y|(2\kappa^2 + y^2)}{(\kappa^2 + y^2)^2} \right)_{Q_{T_e}} - \left(\mathcal{R}_\kappa, \left(S^\kappa_x \varphi \frac{y|y|(2\kappa^2 + y^2)}{(\kappa^2 + y^2)^2} \right)_x \right)_{Q_{T_e}} \\ &= -\frac{1}{2} \left(([S^\kappa_x]_\kappa |S^\kappa_x|_\kappa)_t, \varphi \right)_{Q_{T_e}} - \left(\mathcal{R}_\kappa, \left(S^\kappa_x \varphi \frac{y|y|(2\kappa^2 + y^2)}{(\kappa^2 + y^2)^2} \right)_x \right)_{Q_{T_e}} \\ &= -\frac{1}{2} \left(([S^\kappa_x]_\kappa |S^\kappa_x|_\kappa)_t, \varphi \right)_{Q_{T_e}} - \left(\mathcal{R}_\kappa, (S^\kappa_{xx} \varphi + S^\kappa_x \varphi_x) \frac{y|y|(2\kappa^2 + y^2)}{(\kappa^2 + y^2)^2} \right)_{Q_{T_e}} \\ &\quad - \left(\mathcal{R}_\kappa, S^\kappa_x \varphi \left(\frac{y|y|(2\kappa^2 + y^2)}{(\kappa^2 + y^2)^2} \right)_x \right)_{Q_{T_e}}. \end{aligned} \tag{3.34}$$

To estimate the last two terms on the right-hand side of this inequality we note that

$$\left(\frac{y|y|(2\kappa^2+y^2)}{(\kappa^2+y^2)^2}\right)_x = \frac{4|y|\kappa^4}{(\kappa^2+y^2)^3}y_x.$$

Thus we have the inequalities

$$\left|\frac{y|y|(2\kappa^2+y^2)}{(y^2+\kappa^2)^2}\right| \le \frac{(y^2+\kappa^2)^2}{(\kappa^2+y^2)^2} = 1$$

and

$$\left|y\left(\frac{y|y|(2\kappa^2+y^2)}{(\kappa^2+y^2)^2}\right)_x\right| = \frac{4y^2\kappa^4}{(\kappa^2+y^2)^3}|y_x| \le \frac{4}{3}\frac{(y^2+\kappa^2)^3}{(\kappa^2+y^2)^3}|y_x| = \frac{4}{3}|S^\kappa_{xx}|,$$

which yield the estimates

$$\begin{aligned}&\left|\left(\mathcal{R}_\kappa, (S^\kappa_{xx}\varphi + S^\kappa_x\varphi_x)\frac{y|y|(2\kappa^2+y^2)}{(\kappa^2+y^2)^2}\right)_{Q_{T_e}}\right| \\ &\le C\int_{Q_{T_e}} |\mathcal{R}_\kappa|\left(|S^\kappa_{xx}\varphi| + |S^\kappa_x\varphi_x|\right)d(\tau,x) \qquad (3.35)\end{aligned}$$

and

$$\left|\left(\mathcal{R}_\kappa, S^\kappa_x\varphi\left(\frac{y|y|(2\kappa^2+y^2)}{(\kappa^2+y^2)^2}\right)_x\right)_{Q_{T_e}}\right| \le C\int_{Q_{T_e}} |\mathcal{R}_\kappa S^\kappa_{xx}\varphi| d(\tau,x). \qquad (3.36)$$

The term $\int_{Q_{T_e}} |\mathcal{R}_\kappa S^\kappa_{xx}\varphi| d(\tau,x)$ coincides with the right-hand side of (3.24), which was estimated in (3.25)–(3.27) by $C\|\varphi\|_{L^\infty(0,T;H^2_0(\Omega))}$. The term

$$\int_{Q_{T_e}} |\mathcal{R}_\kappa S^\kappa_x\varphi_x| d(\tau,x) = \int_{Q_{T_e}} |S^\kappa_t S^\kappa_x\varphi_x| d(\tau,x)$$

was estimated in (3.28)–(3.31) by $C\|\varphi\|_{L^\infty(0,T;H^2_0(\Omega))}$. These results and (3.34)–(3.36) yield

$$\left|\left(([S^\kappa_x]_\kappa|S^\kappa_x|_\kappa)_t\,, \varphi\right)_{Q_{T_e}}\right| \le C\|\varphi\|_{L^\infty(0,T_e;H^2_0(\Omega))},$$

which implies (3.21). □

4. Existence of solutions to the phase field model. In this section we use the a priori estimates established in the previous section to study the convergence of $(u^\kappa, T^\kappa, S^\kappa)$ as $\kappa \to 0$. We shall show that there is a subsequence, which converges to a weak solution of the initial-boundary value problem (1.7)–(1.12), thereby proving Theorem 1.1.

Note first that the estimates (3.6), (3.19), the fact that $S^\kappa(t,x) = 0$ for all $(t,x) \in [0,T_e]\times\partial\Omega$, and Poincaré's inequality imply

$$\|S^\kappa\|_{W^{1,4/3}(Q_{T_e})} \le C\,, \qquad (4.1)$$

for a constant C independent of κ. Hence, we can select a sequence $\kappa_n \to 0$ and a function $S \in W^{1,4/3}(Q_{T_e})$ such that the sequence S^{κ_n}, which we again denote by S^κ, satisfies

$$\|S^\kappa - S\|_{L^{4/3}(Q_{T_e})} \to 0, \qquad S^\kappa_x \rightharpoonup S_x\,, \qquad S^\kappa_t \rightharpoonup S_t\,, \qquad (4.2)$$

where the weak convergence is in $L^{4/3}(Q_{T_e})$.

As usual, since (1.9) is nonlinear, the weak convergence of S_x^κ is not enough to prove that the limit function solves this equation. In the following lemma we therefore show that S_x^κ converges pointwise almost everywhere.

LEMMA 4.1. *There exists a subsequence of S_x^κ (we still denote it by S_x^κ) such that*

$$S_x^\kappa \to S_x, \qquad \text{a.e. in } Q_{T_e}, \tag{4.3}$$

$$[S_x^\kappa]_\kappa \to S_x, \quad |S_x^\kappa|_\kappa \to |S_x|, \qquad \text{a.e. in } Q_{T_e}, \tag{4.4}$$

$$|S_x^\kappa|_\kappa \rightharpoonup |S_x|, \quad [S_x^\kappa]_\kappa \rightharpoonup S_x, \qquad \text{weakly in } L^{\frac{4}{3}}(Q_{T_e}), \tag{4.5}$$

$$[S_x^\kappa]_\kappa |S_x^\kappa|_\kappa \to S_x|S_x|, \qquad \text{strongly in } L^{\frac{4}{3}}(0,T_e;L^2(\Omega)), \tag{4.6}$$

as $\kappa \to 0$.

The proof is based on the following two results.

THEOREM 4.1. *Let B_0 be a normed linear space imbedded compactly into another normed linear space B, which is continuously imbedded into a Hausdorff locally convex space B_1, and $1 \le p < +\infty$. If $v, v_i \in L^p(0,T_e;B_0), i \in \mathbb{N}$, the sequence $\{v_i\}_{i\in\mathbb{N}}$ converges weakly to v in $L^p(0,T_e;B_0)$, and $\{\frac{\partial v_i}{\partial t}\}_{i\in\mathbb{N}}$ is bounded in $L^1(0,T_e;B_1)$, then v_i converges to v strongly in $L^p(0,T_e;B)$.*

LEMMA 4.2. *Let $(0,T_e)\times\Omega$ be an open set in $\mathbb{R}^+\times\mathbb{R}^n$. Suppose functions g_n, g are in $L^q((0,T_e)\times\Omega)$ for any given $1 < q < \infty$, which satisfy*

$$\|g_n\|_{L^q((0,T_e)\times\Omega)} \le C, \qquad g_n \to g \text{ a.e. in } (0,T_e)\times\Omega.$$

Then g_n converges to g weakly in $L^q((0,T_e)\times\Omega)$.

Theorem 4.1 is a general version of the Aubin–Lions lemma valid under the weak assumption $\partial_t v_i \in L^1(0,T_e;B_1)$. This version, which we need here, is proved in [14]. A proof of Lemma 4.2 can be found in [12, p. 12].

Proof of Lemma 4.1. We choose $p = \frac{4}{3}$ and

$$B_0 = W^{1,\frac{4}{3}}(\Omega), \quad B = L^2(\Omega), \quad B_1 = H^{-2}(\Omega).$$

These spaces satisfy the assumptions of the theorem. Since the estimates (3.10) and (3.21) imply that the sequence $([S_x^\kappa]_\kappa|S_x^\kappa|_\kappa)$ is uniformly bounded in $L^p(0,T_e;B_0)$ for $\kappa \to 0$ and $([S_x^\kappa]_\kappa|S_x^\kappa|_\kappa)_t$ is uniformly bounded in $L^1(0,T_e;B_1)$, it follows from Theorem 4.1 that there is a subsequence, still denoted by $([S_x^\kappa]_\kappa|S_x^\kappa|_\kappa)$, which converges strongly in $L^p(0,T_e;B) = L^{\frac{4}{3}}(0,T_e;L^2(\Omega))$ to a limit function $G \in L^{\frac{4}{3}}(0,T_e;L^2(\Omega))$. Consequently, from this sequence we can select another subsequence, denoted in the same way, which converges almost everywhere in Q_{T_e}. Using that the mapping $y \mapsto f(y) := y|y|$ has a continuous inverse $f^{-1} : \mathbb{R} \to \mathbb{R}$, we infer that also the sequence $[S_x^\kappa]_\kappa = f^{-1}([S_x^\kappa]_\kappa|S_x^\kappa|_\kappa)$ converges pointwise almost everywhere in Q_{T_e}.

From this we deduce also that the sequence S_x^κ converges pointwise almost everywhere. To see this, let $y_\kappa = S_x^\kappa$, $v_\kappa = [S_x^\kappa]_\kappa$ and $v = \lim_{\kappa\to 0} v_\kappa$. From

$$y_\kappa^4 = v_\kappa^2(\kappa^2 + y_\kappa^2) = v_\kappa^2\kappa^2 + v_\kappa^2 y_\kappa^2$$

we conclude

$$y_\kappa^4 - v_\kappa^2\kappa^2 - v_\kappa^2 y_\kappa^2 = 0, \tag{4.7}$$

and hence

$$y_\kappa^2 = \frac{v_\kappa^2 + \sqrt{v_\kappa^4 + 4v_\kappa^2\kappa^2}}{2},$$

since the second solution of (4.7) is negative. Therefore, for $\kappa \to 0$,

$$y_\kappa^2 = \frac{v_\kappa^2 + \sqrt{v_\kappa^4 + 4v_\kappa^2\kappa^2}}{2} \to \frac{v^2 + \sqrt{v^4}}{2} = v^2.$$

From the fact that $\mathrm{sign}(v_\kappa) = \mathrm{sign}(y_\kappa)$ we thus obtain

$$\begin{aligned} |y_\kappa - v_\kappa|^2 &= y_\kappa^2 - 2y_\kappa v_\kappa + v_\kappa^2 \\ &= y_\kappa^2 - 2|y_\kappa||v_\kappa| + v_\kappa^2 \to v^2 - 2|v||v| + |v|^2 = 0, \end{aligned} \tag{4.8}$$

and hence

$$\lim_{\kappa\to 0} S_x^\kappa = \lim_{\kappa\to 0} y_\kappa = \lim_{\kappa\to 0} v_\kappa = v = \lim_{\kappa\to 0} [S_x^\kappa]_\kappa.$$

Therefore S_x^κ converges pointwise almost everywhere in Q_{T_e}. Since $S_x^\kappa \rightharpoonup S_x$ weakly in $L^{\frac{4}{3}}(Q_{T_e})$, we conclude from Lemma 4.2 that $S_x^\kappa \to S_x$ and $[S_x]^\kappa \to S_x$ almost everywhere in Q_{T_e}. This proves (4.3) and (4.4). Relation (4.4) yields $[S_x^\kappa]_\kappa |S_x^\kappa|_\kappa \to S_x|S_x|$ almost everywhere in Q_{T_e}, which implies that the limit function G of $[S_x^\kappa]_\kappa |S_x^\kappa|_\kappa$ is equal to $S_x|S_x|$. This proves (4.6).

To prove (4.5) we note that the estimate $|[S_x^\kappa]_\kappa| = |S_x^\kappa|_\kappa \le |S_x^\kappa|$ and the inequality (4.1) together imply that the sequences $[S_x^\kappa]_\kappa$ and $|S_x^\kappa|_\kappa$ are uniformly bounded in $L^{\frac{4}{3}}(Q_{T_e})$. Thus, (4.5) is a consequence of (4.4) and Lemma 4.2. ∎

Proof of Theorem 1.1. Define the functions u and T by

$$\begin{aligned} u(t,x) &= u^* \left(\int_a^x S(t,y)dy - \frac{x-a}{d-a} \int_a^d S(t,y)dy \right) + w(t,x), \\ T(t,x) &= D(\varepsilon^* - \bar\varepsilon)S - D\varepsilon^* \frac{1}{d-a} \int_a^d S(t,y)dy + \sigma(t,x), \end{aligned} \tag{4.9}$$

where for S we insert the limit function of the sequence S^κ given in (4.2), and where $u^* \in \mathbb{R}^3$, $\varepsilon^* \in \mathcal{S}^3$, and (w,σ) are the same constants and functions as in (2.9) and (2.10). We prove that (u,T,S) is a weak solution of problem (1.7)–(1.12).

To this end note that (3.4) and (4.2) imply $S \in L^\infty(Q_{T_e})$. From this relation, from the above definition of u and T, and from $(w,\sigma) \in C^{2,1}(\bar Q_{T_e}) \times C^{1,1}(\bar Q_{T_e})$, we immediately see that u and T satisfy (1.14) and (1.15). Observe next that $\|S^\kappa\|_{L^\infty(0,T_e;H_0^1(\Omega))} \le C$, by (3.6). This implies $S \in L^\infty(0,T_e;H_0^1(\Omega))$, since we can select a subsequence of S^κ which converges weakly to S in this space. Thus, S satisfies (1.16).

It is shown in [3] that the functions u and T defined in this way satisfy (1.7), (1.8), and (1.11). (We remarked this previously.) It therefore suffices to show that (1.9) and (1.12) are fulfilled in the weak sense. By definition, these equations are satisfied in the weak sense if the relation (1.17) holds. To verify (1.17) we use that, by construction, (T^κ, S^κ) solves (2.4), (2.6), and (2.7). If we multiply (2.4) by a test function $\varphi \in C_0^\infty((-\infty,T_e)\times\Omega)$ and integrate the resulting equation over Q_{T_e}, we obtain

$$\begin{aligned} 0 &= (S_t^\kappa, \varphi)_{Q_{T_e}} + \big(-(c\nu|S_x^\kappa|_\kappa + \kappa)S_{xx}^\kappa - c\big(T^\kappa\cdot\bar\varepsilon - \hat\psi'(S^\kappa)\big)|S_x^\kappa|_\kappa, \varphi\big)_{Q_T} \\ &= -(S_0^\kappa, \varphi(0))_\Omega - (S^\kappa, \varphi_t)_{Q_{T_e}} + \big(c\nu \int_0^{S_x^\kappa} |y|_\kappa dy + \kappa S_x^\kappa, \varphi_x\big)_{Q_{T_e}} \\ &\quad + \big(c\big(T^\kappa\cdot\bar\varepsilon - \hat\psi'(S^\kappa)\big)|S_x^\kappa|_\kappa, \varphi\big)_{Q_{T_e}}. \end{aligned}$$

Equation (1.17) follows from this relation if we show that

$$(S_0^\kappa, \varphi(0))_\Omega \to (S_0, \varphi(0))_\Omega, \tag{4.10}$$

$$(S^\kappa, \varphi_t)_{Q_{T_e}} \to (S, \varphi_t)_{Q_{T_e}}, \tag{4.11}$$

$$\left(\int_0^{S_x^\kappa} |y|_\kappa dy, \varphi_x\right)_{Q_{T_e}} \to \left(\frac{1}{2}|S_x|S_x, \varphi_x\right)_{Q_{T_e}}, \tag{4.12}$$

$$\left(\left(T^\kappa \cdot \bar\varepsilon - \hat\psi'(S^\kappa)\right)|S_x^\kappa|_\kappa, \varphi\right)_{Q_{T_e}} \to \left(\left(T \cdot \bar\varepsilon - \hat\psi'(S)\right)|S_x|, \varphi\right)_{Q_{T_e}}, \tag{4.13}$$

$$(\kappa S_x^\kappa, \varphi_x)_{Q_{T_e}} \to 0, \tag{4.14}$$

for $\kappa \to 0$. Now, the relation (4.10) follows from (3.3), the relation (4.11) is a consequence of (4.2), and the relation (4.14) is obtained from (4.1). To prove (4.12) we use that

$$\begin{aligned}\int_0^{S_x^\kappa} |y|_\kappa dy - \frac{1}{2}S_x|S_x| &= \left(\int_0^{S_x^\kappa} |y|_\kappa dy - \frac{1}{2}[S_x]_\kappa |S_x|_\kappa\right) + \frac{1}{2}\left([S_x]_\kappa |S_x|_\kappa - S_x|S_x|\right)\\ &=: I_1 + I_2.\end{aligned} \tag{4.15}$$

The relation (4.6) implies

$$\|I_2\|_{L^{\frac{4}{3}}(0,T_e;L^2(\Omega))} \to 0 \tag{4.16}$$

for $\kappa \to 0$. Moreover,

$$\begin{aligned}|I_1| &= \left|\int_0^{S_x^\kappa} |y|_\kappa dy - \int_0^{S_x^\kappa} |y| dy\right| = \left|\int_0^{S_x^\kappa} \left(\frac{y^2}{\sqrt{\kappa^2+y^2}} - |y|\right) dy\right|\\ &\le \int_0^{|S_x^\kappa|} \frac{|y|}{\sqrt{\kappa^2+y^2}} \left|\sqrt{\kappa^2+y^2} - |y|\right| dy \le \int_0^{|S_x^\kappa|} \kappa dy = \kappa |S_x^\kappa|,\end{aligned}$$

whence (3.6) implies

$$\|I_1\|_{L^{\frac{4}{3}}(0,T_e;L^2(\Omega))} \le C\|I_1\|_{L^2(Q_{T_e})} \le C\kappa \to 0$$

for $\kappa \to 0$. From this relation and from (4.15), (4.16) we obtain

$$\left\|\int_0^{S_x^\kappa} |y|_\kappa dy - \frac{1}{2}S_x|S_x|\right\|_{L^{\frac{4}{3}}(0,T_e;L^2(\Omega))} \to 0,$$

which implies (4.12). To verify (4.13) we note that (2.10) and (4.9) yield

$$\begin{aligned}&T^\kappa(t,x) - T(t,x)\\ &= D(\varepsilon^* - \bar\varepsilon)(\chi_\kappa * S^\kappa - S)(t,x) - \frac{D\varepsilon^*}{d-a}\int_a^d (\chi_\kappa * S^\kappa - S)(t,y)dy.\end{aligned} \tag{4.17}$$

From (2.1) and (4.2) we conclude that

$$\begin{aligned}\|\chi_\kappa * S^\kappa - S\|_{L^{\frac{4}{3}}(Q_{T_e})} &\le \|\chi_\kappa * (S^\kappa - S)\|_{L^{\frac{4}{3}}(Q_{T_e})} + \|(S - \chi_\kappa * S)\|_{L^{\frac{4}{3}}(Q_{T_e})} \\ &\le \|(S - \chi_\kappa * S)\|_{L^{\frac{4}{3}}(Q_{T_e})} + \|S^\kappa - S\|_{L^{\frac{4}{3}}(Q_{T_e})} \to 0\end{aligned}$$

for $\kappa \to 0$. Since ε^* and $\bar{\varepsilon}$ are constants, we infer from this relation and from (4.17) that

$$\|T - T^\kappa\|_{L^{\frac{4}{3}}(Q_{T_e})} \to 0$$

for $\kappa \to 0$. Thus, after selecting a subsequence we have $T^\kappa \to T$ almost everywhere in Q_{T_e}. Putting this together with (4.3) and (4.4), we see that $(T^\kappa \cdot \bar{\varepsilon} - \hat{\psi}'(S^\kappa))|S_x^\kappa|_\kappa$ tends to $(T \cdot \bar{\varepsilon} - \hat{\psi}'(S))|S_x|$ almost everywhere in Q_{T_e}. Since (3.6) and (3.5) imply that $(T^\kappa \cdot \bar{\varepsilon} - \hat{\psi}'(S^\kappa))|S_x^\kappa|_\kappa$ is uniformly bounded in $L^2(Q_{T_e})$, we deduce from Lemma 4.2 that

$$(T^\kappa \cdot \bar{\varepsilon} - \hat{\psi}'(S^\kappa))|S_x^\kappa|_\kappa \rightharpoonup (T \cdot \bar{\varepsilon} - \hat{\psi}'(S))|S_x|,$$

weakly in $L^2(Q_{T_e})$, which implies (4.13). Consequently (1.17) holds.

It remains to prove that the solution has the regularity properties stated in (1.18) and (1.19). The relation $S_t \in L^{\frac{4}{3}}(Q_{T_e})$ is implied by (4.2). To verify the second assertion in (1.18), we use estimate (3.11) to get

$$\int_0^{T_e} \|[S_x^\kappa]_\kappa\|_{L^q(\Omega)}^{\frac{8}{3}} dt \le C$$

for any $1 < q < \infty$, since Ω is bounded. Using this estimate and (4.4), we infer from Lemma 4.2 that $[S_x^\kappa]_\kappa \rightharpoonup S_x$ in $L^{\frac{8}{3}}(0, T_e; L^q(\Omega))$, whence $S_x \in L^{\frac{8}{3}}(0, T_e; L^q(\Omega))$ follows.

To prove (1.19), we recall that $[S_x^\kappa]_\kappa |S_x^\kappa|_\kappa$ converges to $|S_x|S_x$ strongly in the space $L^{\frac{4}{3}}(0, T_e; L^2(\Omega)) \subset L^{\frac{4}{3}}(Q_{T_e})$ and that $([S_x^\kappa]_\kappa |S_x^\kappa|_\kappa)_x$ is uniformly bounded in $L^{\frac{4}{3}}(Q_{T_e})$ for $\kappa \to 0$, by (3.10). This taken together implies that $(|S_x|S_x)_x \in L^{\frac{4}{3}}(Q_{T_e})$. Finally, to prove the second assertion of (1.19) we choose a test function $\varphi \in L^4(0, T_e; W_0^{1,4}(\Omega))$, multiply (2.4) by $-\varphi_x$, and integrate the resulting equation over Q_{T_e} to obtain

$$0 = (S_t^\kappa - \mathcal{R}_\kappa, -\varphi_x)_{Q_{T_e}} = (S_{xt}^\kappa, \varphi)_{Q_{T_e}} + (\mathcal{R}_\kappa, \varphi_x)_{Q_{T_e}}, \tag{4.18}$$

with $\mathcal{R}_\kappa$ defined in (3.32). Invoking the estimates (3.6), (3.5), and (3.8), we deduce that

$$\|\mathcal{R}_\kappa\|_{L^{\frac{4}{3}}(Q_{T_e})} \le C,$$

and hence (4.18) yields

$$(S_{xt}^\kappa, \varphi)_{Q_{T_e}} \le \|\mathcal{R}_\kappa\|_{L^{\frac{4}{3}}(Q_{T_e})} \|\varphi_x\|_{L^4(Q_{T_e})} \le C\|\varphi\|_{L^4(0,T_e;W_0^{1,4}(\Omega))},$$

which means that S_{xt}^κ is uniformly bounded in $L^{\frac{4}{3}}(0, T_e; W^{-1,\frac{4}{3}}(\Omega))$. From this estimate and from $S_t^\kappa \rightharpoonup S_t$ in $L^{\frac{4}{3}}(Q_{T_e})$ we deduce easily that S_{xt} belongs to the dual space of $L^4(0, T_e; W_0^{1,4}(\Omega))$, which is $L^{\frac{4}{3}}(0, T_e; W^{-1,\frac{4}{3}}(\Omega))$. ∎

Appendix. Background of the model. We briefly sketch the physical background, the sharp interface model, and the derivation of the diffusive interface model (1.1)–(1.5) from this sharp interface model.

Material phases are characterized by the structure of the crystal lattice, in which the atoms are arranged. An interface between different material phases moves if the crystal lattice in front of the interface is transformed from one structure to the other. Often phase transformations are triggered by diffusion processes. A well-known model for diffusion dominated transformations is the Cahn–Hilliard equation. Here we consider a sharp interface model for diffusionless transformations, also called martensitic transformations; cf. [9, p. 162]. This sharp interface model is an initial-boundary value problem for the unknown functions u, T and for the unknown interface $\Gamma(t) \subseteq \Omega$ between two material phases, which is a free boundary. It consists of (1.1), (1.2); of the interface conditions

$$s(t,x)[S](t,x) = c\left(-\langle T\rangle(t,x)\cdot\bar{\varepsilon}\,[S](t,x) + [\hat{\psi}(S)](t,x)\right), \tag{A.1}$$

$$[u](t,x) = 0, \qquad [T](t,x)n(t,x) = 0, \tag{A.2}$$

which must hold for $x \in \Gamma(t)$; of a Dirichlet boundary condition for u; and of the initial condition (1.5). We use the notation $[f] = f^+ - f^-$ and $\langle f\rangle = \frac{1}{2}(f^+ + f^-)$, where f^+, f^- are the limit values of the function f on both sides of $\Gamma(t)$. Moreover, $s(t,x) \in \mathbb{R}^3$ denotes the normal speed of the interface $\Gamma(t)$, which is measured as positive in the direction for which $[S](t,x)$ is positive. Here c is a positive constant.

Equation (A.1), a constitutive equation, determines the normal speed s of the phase interface as a function of the term $-\langle T\rangle\cdot\bar{\varepsilon}[S] + [\hat{\psi}(S)]$. Some computations show that this term is equal to the expression $n\cdot[C]n$ with the Eshelby tensor C and the normal vector n to $\Gamma(t)$ (cf. [3]) and thus is a configurational force. We assume that s depends linearly on the configurational force, which is the most simple constitutive assumption. Thus, in this model the evolution of the phase interface is driven by the configurational force along the interface, an assumption appropriate for martensitic transformations.

Though configurational forces were introduced in the first half of the last century, it was clearly stated for the first time in [1] that (1.1), (1.2), (A.1), (A.2) form a closed initial-boundary value problem. Applications of this model can be found, for example, in [6, 13, 15], where equilibrium configurations for materials with phase transitions are determined, and in [10], where the evolution of phase interfaces in ferroelectric materials is modeled. In a sense, this free initial-boundary value problem from solid mechanics is comparable to the Stefan problem in fluid mechanics.

The initial-boundary value problem (1.1)–(1.5) can be considered to be a regularization of this sharp interface model, which could be used to prove existence of solutions of the sharp interface model, and it can also be considered to be a diffusive interface model for martensitic phase transitions, which is useful by itself and avoids some disadvantages of the model with sharp interfaces. We are interested in both aspects.

The derivation of (1.1)–(1.5) given in [2, 3] uses a rigorous method. To make the model plausible, we derive the model here in a different, short, but formal way. To this end we replace the phase interface $\Gamma(t)$, across which the order parameter jumps from 0 to 1, by finitely many interfaces parallel to the original interface, and consider a new order parameter, again denoted by S, with small jumps across these interfaces, such that the sum of the jumps is equal to 1. We assume that the new order

parameter satisfies (A.1) and (A.2) along all interfaces. If we increase the number of interfaces and decrease the jump height, the new order parameter will converge to a continuous or even differentiable order parameter, for which the normal speed of the level manifolds is equal to the limit of the normal speed of the interfaces. For this limit speed we obtain from (A.1)

$$s(t,x) = c \lim_{[S]\to 0} (-\langle T\rangle \cdot \bar{\varepsilon} + \hat{\psi}'(S^*)) = c(-T\cdot\bar{\varepsilon} + \hat{\psi}'(S)) = c\,\psi_S(\varepsilon(\nabla_x u), S).$$

The limit order parameter thus satisfies the Hamilton–Jacobi transport equation

$$S_t = -c\,\psi_S(\varepsilon(\nabla_x u), S)\,|\nabla_x S|, \tag{A.3}$$

since the level manifolds of solutions of this equation have this normal speed.

The idea suggests itself to approximate the solution of the sharp interface model by smooth solutions (u, T, S) of the system (1.1), (1.2), (A.3). Yet, examples in one space dimension show that in general the function S in such a smooth solution develops a jump after finite time. The reason for this is that the function $\hat{\psi}'$ appearing in ψ_S is not monotone, since $\hat{\psi}$ is a double well potential. After S has developed a jump, (A.3) can no longer be used to govern the evolution of S. To avoid this problem and to force solutions to stay smooth, (A.3) has been replaced by (1.3), which contains the regularizing term $\nu|\nabla_x S|\Delta_x S$ with the small positive parameter ν. This yields the model (1.1)–(1.5).

The choice of this special regularizing term follows from the second law of thermodynamics, which every model must satisfy. This law requires that there exist a free energy ψ^* and a flux q such that $\frac{\partial}{\partial t}\,\psi^* + \operatorname{div}_x q \le b\cdot u_t$ holds; cf. [4]. If we choose

$$\psi^*(\varepsilon, S, \nabla_x S) = \psi(\varepsilon, S) + \frac{\nu}{2}|\nabla_x S|^2, \quad q(\varepsilon, u_t, S, \nabla_x S, S_t) = -(Tu_t) - \nu\,(S_t \nabla_x S),$$

it follows by a short computation for solutions (u, T, S) of (1.1), (1.2) that

$$\frac{\partial}{\partial t}\psi^*(\varepsilon, S, \nabla_x S) - \operatorname{div}_x(Tu_t) - \nu\operatorname{div}_x(S_t\nabla_x S) - b\cdot u_t = (\psi_S(\varepsilon, S) - \nu\Delta_x S)S_t.$$

Inserting (1.3) into this equation shows that the right-hand side is nonpositive, whence the second law is fulfilled. This would not be true for the standard regularization $S_t = -c\psi_S(\varepsilon(\nabla_x u), S)\,|\nabla_x S| + \nu\Delta S$ of (A.3).

REFERENCES

[1] R. Abeyaratne and J. K. Knowles, *On the driving traction acting on a surface of strain discontinuity in a continuum*, J. Mech. Phys. Solids, 38 (1990), pp. 345–360.

[2] H.-D. Alber, *Evolving microstructure and homogenization*, Continum. Mech. Thermodyn., 12 (2000), pp. 235–286.

[3] H.-D. Alber and P. Zhu, *Evolution of phase boundaries by configurational forces*, Arch. Ration. Mech. Anal., submitted; available online at wwwbib.mathematik.tu-darmstadt.de/Math-Net/Preprints/.

[4] H. W. Alt and I. Pawlow, *On the entropy principle of phase transition models with a conserved order parameter*, Adv. Math. Sci. Appl., 6 (1996), pp. 291–376.

[5] E. Bonetti, P. Colli, W. Dreyer, G. Gilardi, G. Schimpanera, and J. Sprekels, *On a model for phase separation in binary alloys driven by mechanical effects*, Phys. D, 165 (2002), pp. 48–65.

[6] G. Buratti, Y. Huo, and I. Müller, *Eshelby tensor as a tensor of free enthalpy,* J. Elasticity, 72 (2003), pp. 31–42.

[7] M. Carrive, A. Miranville, and A. Pierus, *The Cahn–Hilliard equation for deformable elastic continua,* Adv. Math. Sci. Appl., 10 (2000), pp. 539–569.
[8] H. Garcke, *On Cahn–Hilliard systems with elasticity,* Proc. Roy. Soc. Edinburgh, Sect. A, 133 (2003), pp. 307–331.
[9] E. Hornbogen and H. Warlimont, *Metallkunde*, 4th ed., Springer-Verlag, 2001.
[10] R. James, *Configurational forces in magnetism with application to the dynamics of a small-scale ferromagnetic shape memory cantilever*, Contin. Mech. Thermodyn., 14 (2002), pp. 55–86.
[11] O. Ladyzenskaya, V. Solonnikov, and N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, Trans. Math. Monogr. 23, AMS, Providence, RI, 1968.
[12] J. Lions, *Quelques Methodes de Resolution des Problemes aux Limites Non Lineaires*, Dunod Gauthier-Villars, Paris, 1969.
[13] R. Müller and D. Gross, 3*D simulation of equilibrium morphologies of precipitates*, Computational Materials Sci., 11 (1998), pp. 35–44.
[14] T. Roubíček, *A generalization of the Lions–Temam compact imbedding theorem*, Casopis Pest. Mat., 115 (1990), pp. 338–342.
[15] S. Socrate and D. Parks, *Numerical determination of the elastic driving force for directional coarsening in Ni-superalloys*, Acta Metall. Mater., 40 (1993), pp. 2185–2209.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 700–720

PERIODIC STATIONARY PATTERNS GOVERNED BY A CONVECTIVE CAHN–HILLIARD EQUATION*

MICHAEL A. ZAKS†, ALLA PODOLNY‡, ALEXANDER A. NEPOMNYASHCHY§, AND ALEXANDER A. GOLOVIN¶

Abstract. We investigate bifurcations of stationary periodic solutions of a convective Cahn–Hilliard equation, $u_t+Duu_x+(u-u^3+u_{xx})_{xx}=0$, describing phase separation in driven systems, and study the stability of the main family of these solutions. For the driving parameter $D<D_0=\sqrt{2}/3$, the periodic stationary solutions are unstable. For $D>D_0$, the periodic stationary solutions are stable if their wavelength belongs to a certain stability interval. It is therefore shown that in a driven phase-separating system that undergoes spinodal decomposition the coarsening can be stopped by the driving force, and formation of stable periodic structures is possible. The modes that destroy the stability at the boundaries of the stability interval are also found.

Key words. Cahn–Hilliard equation, pattern formation, stability

AMS subject classifications. 35B10, 35B32, 35K35, 70K44, 74N20

DOI. 10.1137/040615766

1. Introduction.

In recent decades, the spontaneous formation of spatially inhomogeneous patterns has been an object of extensive investigations. Many phenomena have been understood by using the *complex Ginzburg–Landau equation* [1] as a basic model. However, this equation is not valid when the growth or decay of a spatially homogeneous disturbance is forbidden by symmetry arguments or by a conservation law, and therefore such a disturbance is neutrally stable. The instability spectrum contains a neutrally stable mode, with the zero wavenumber corresponding to infinitesimal shift of the stationary solution [2]. In this case the situation is much more intricate, and there is a variety of different order parameter equations that describe the nonlinear evolution of long-wave disturbances, depending on the symmetry of the problem [3].

In the case of a monotonic instability, the nonlinear evolution of long-wave disturbances is typically described by the *Kuramoto–Sivashinsky* (KS) *equation*

$$u_t + u_{xxxx} + u_{xx} - \kappa_1 (u^2)_x = 0, \tag{1.1}$$

where $u(x,t)$ is the order parameter and κ_1 is a constant. This equation has been derived in a large number of physical contexts; e.g., it describes the instability of oscillations in reaction-diffusion systems [4], the instability of a flame front [5], film flow instability [6, 7], and instability of solidification fronts [8, 9, 10]. Though for

*Received by the editors September 25, 2004; accepted for publication (in revised form) August 10, 2005; published electronically January 26, 2006. This research was supported by the Minerva Center for Nonlinear Physics of Complex Systems funded through the BMBF, the Binational US-Israel Science Foundation, and US Department of Energy grant DE-FG02-03ER46069.
http://www.siam.org/journals/siap/66-2/61576.html

†Institute of Physics, Humboldt University, D-12489 Berlin, Germany (zaks@physik.hu-berlin.de). This author acknowledges the support of SFB-555.

‡Department of Mathematics, Technion—Israel Institute of Technology, Haifa 32000, Israel (asspekto@techunix.technion.ac.il).

§Department of Mathematics and Minerva Center for Nonlinear Physics of Complex Systems, Technion—Israel Institute of Technology, Haifa 32000, Israel (nepom@math.technion.ac.il).

¶Corresponding author. Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208 (a-golovin@northwestern.edu).

the KS equation there exists an interval of locally stable spatially periodic stationary solutions [7, 11, 12], the most remarkable type of dynamics is spatio-temporal chaos [13, 14].

If the problem is invariant with respect to the reflection $x \to -x$, as well as to the transformation $u \to -u$, the appropriate order parameter equation is the *Cahn–Hilliard* (CH) *equation*,

$$u_t + u_{xxxx} + u_{xx} - \kappa_2 (u^3)_{xx} = 0. \tag{1.2}$$

This equation, originally proposed as a model of the phase-transition kinetics with a conserved order parameter [15], arises in many physical problems, including faceting transition by a temperature change [16, 17], secondary flows produced by the instability of the Kolmogorov flow [18], and zigzag instability of convection patterns [19]. Unlike the KS equation, the CH equation is potential; i.e., it has a Lyapunov functional that monotonically decreases with time. The characteristic feature of the dynamics governed by the CH equation is the formation of kinks, i.e., domain walls between domains with nearly constant values of the order parameter. Stationary solutions of the CH equation, which can be represented as equidistant sequences of kinks, are unstable with respect to disturbances that change the distances between kinks [20, 18] and lead to an exponentially slow *coarsening* of the structure [21, 22].

Recently, a nonpotential modification of the CH equation, the *convective* (or *driven*) *Cahn–Hilliard* (CCH) *equation*,

$$u_t + u_{xxxx} + u_{xx} - \kappa_1 (u^2)_x - \kappa_2 (u^3)_{xx} = 0, \tag{1.3}$$

has been attracting a great deal of attention. This equation, containing nonlinearities typical of both KS and CH equations, has been proposed to describe several physical processes, namely spinodal decomposition of phase separating systems in an external field [23, 24, 25], step instability on a crystal surface [26], faceting of growing thermodynamically unstable surfaces [27, 28, 29, 30, 31], as well as dewetting of a thin film flowing down an inclined plane [32]. A coordinate-invariant form of this equation has been derived in [33, 34]. Depending on the value of the parameter $D = 2\kappa_1/\kappa_2^{1/2}$ that characterizes the relative "strength" of the two nonlinearities, one can observe the coarsening of kinks (for small D) [25, 28, 29, 31], spatio-temporal chaos (for large D) [26, 30], as well as other dynamic regimes [26, 30]. A characteristic feature of the dynamics governed by the CCH equation at intermediate values of D is the appearance of wavy regimes with a simple *temporal* behavior (stationary and traveling waves) but with a complicated *spatial* structure [30].

The present paper is devoted to the investigation of stationary patterns governed by the CCH equation. The bifurcations of stationary solutions are discussed in section 2. The stability analysis is performed in section 3.

2. Stationary solutions.

2.1. Formulation of the problem. Let us rescale the variable u, $u(x) = U(x)/(\kappa_2)^{1/2}$, and define $D = 2\kappa_1/\kappa_2^{1/2}$. The CCH equation (1.3) is transformed into the form

$$U_t + (U_{xx} + U - U^3)_{xx} - \frac{D}{2}(U^2)_x = 0. \tag{2.1}$$

In the present paper we consider stationary solutions $U = U(x)$ of (2.1) in the infinite region $-\infty < x < \infty$. Only solutions bounded at infinity are relevant. Integrating the corresponding ordinary differential equation ((2.1) with $U_t = 0$), one

obtains the following problem:

$$U_{xxx} + (U - U^3)_x - \frac{D}{2}U^2 = -\frac{D}{2}A, \qquad -\infty < x < \infty, \tag{2.2}$$

$$x \to \pm\infty, \qquad |U| < \infty. \tag{2.3}$$

Recall that for $D = 0$, which corresponds to the CH equation, all stationary solutions of (2.2) can be found analytically [35]. In the opposite limit, $D \to \infty$ (the KS equation), the set of stationary solutions is much more complicated than in the case of the CH equation. Bifurcations of periodic stationary and traveling-wave solutions of the KS equation were studied in [36, 37]. Later, without loss of generality, we assume $D > 0$.

It is convenient to investigate the bifurcations of stationary states of (2.1) in terms of the dynamical system defined by (2.2) with parameters D and A. Notably, this dynamical system is measure-preserving. It is also reversible, i.e., invariant with respect to inversion, $x \to -x$, $U \to -U$, as well as invariant with respect to translation $x \to x + C$. Accordingly, all stationary solutions are either invariant with respect to these transformations, or (up to an arbitrary shift along x) exist in pairs: if $U(x+C)$ is a family of solutions, then $-U(C - x)$ is also a family of solutions. Both families may coincide.

2.2. Constant and heteroclinic solutions. For $A = 0$ the only bounded solution of the problem (2.2), (2.3) is the trivial equilibrium solution $U = 0$. The linearized problem possesses three eigenvalues on the imaginary axis: $\lambda_1 = 0$ and $\lambda_{2,3} = \pm i$; accordingly, the equilibrium state is structurally unstable. Increasing A removes the degeneracy: the equilibrium splits into two constant solutions, $U = U^{\pm} = \pm\sqrt{A}$; besides, the imaginary eigenvalues are responsible for the creation of the periodic (with respect to x) orbit with period $\approx 2\pi$ whose amplitude locally grows as $\sim \sqrt{A}$ (for small A); below, this orbit will be referred to as the "main family."

In fact, the equilibria and the closed orbit are merely the simplest solutions born in the course of unfolding the degeneracy: formation of the complicated invariant set was investigated for the KS equation, where it was called the "cocoon bifurcation" [38]. For our purposes, however, these simple stationary states suffice.

Let us start the analysis of the problem (2.2), (2.3) with the consideration of the equilibrium points $U = U^{\pm} = \pm\sqrt{A}$. Consider perturbations of the constant solution, $U = U^{\pm} + \tilde{U}(x)$, $\tilde{U}(x) \sim e^{\lambda x}$, and linearize (2.2) to obtain the following equation for the eigenvalues:

$$\lambda^3 + (1 - 3A)\lambda \mp D\sqrt{A} = 0. \tag{2.4}$$

All eigenvalues are real if

$$D^2 \le D_*^2 = \frac{4(3A-1)^3}{27A}, \qquad A > \frac{1}{3}. \tag{2.5}$$

Otherwise, two eigenvalues are complex conjugate, and the third one is real. For the equilibrium point $U = U^+ = \sqrt{A}$, one of the eigenvalues is positive and the other two eigenvalues are either negative or have negative real parts. The unstable manifold $W^u(U^+)$ is one-dimensional, while the stable manifold $W^s(U^+)$ is two-dimensional. For the symmetric counterpart of U^+, the solution $U^- = -\sqrt{A}$, the eigenvalues have

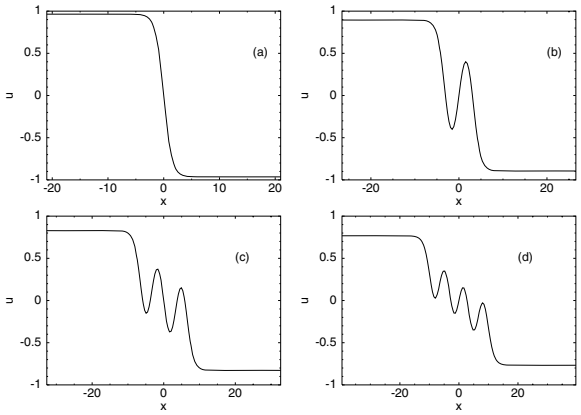


FIG. 1. *Different negative kinks at* $D = 0.1$: (a) $A = 0.929289$ *(solution* (2.7)*),* (b) $A = 0.799874$, (c) $A = 0.685489$, (d) $A = 0.588339$.

opposite signs; hence its stable manifold $W^s(U^-)$ is one-dimensional, and its unstable manifold $W^u(U^-)$ is two-dimensional.

Heteroclinic solutions ("kinks") joining the equilibrium points $U = U^+$ and $U = U^-$ correspond to trajectories on $M_+ = W^u(U^-) \cap W^s(U^+)$ ("positive kinks") and $M_- = W^u(U^+) \cap W^s(U^-)$ ("negative kinks"). As a matter of fact, there are *exact* solutions of this kind [23], one for a positive kink with $A = A_+ = 1 + D/\sqrt{2}$,

$$U = U_+(x) = U_+^0 \tanh \frac{U_+^0}{\sqrt{2}}(x - x_0), \tag{2.6}$$
$$U_+^0 = \sqrt{1 + D/\sqrt{2}}, \quad x_0 = const,$$

and the other for a negative kink with $A = A_- = 1 - D/\sqrt{2}$, $D < \sqrt{2}$,

$$U = U_-(x) = -U_-^0 \tanh \frac{U_-^0}{\sqrt{2}}(x - x_0), \tag{2.7}$$
$$U_-^0 = \sqrt{1 - D/\sqrt{2}}, \quad x_0 = const.$$

Since the manifolds $W^s(U^+)$ and $W^u(U^-)$ are two-dimensional for any $A > 0$, their intersection in the three-dimensional phase space is generic, and therefore the solution (2.6) is a representative of a *family* of positive kinks $U_+(x; A)$. A negative kink requires matching of two one-dimensional manifolds in the three-dimensional space, which is not generic, but a codimension-2 event. The symmetry reduces the codimension of this event to 1. Accordingly, we can expect that, for a given D, negative kinks exist only for isolated values of A. Solution (2.7) appears to have the largest value of A and the simplest spatial profile among the possible negative kinks. Several examples of negative kinks are shown in Figure 1.

Among the presented negative kinks only the solution (2.7) is monotonic; the others have additional humps. At very small values of D these humps almost reach the value U_+^0; as D grows, the humps become smaller, and for $D > 0.3$ the profiles of all negative kinks become practically monotonic. As seen in Figure 2, on the parameter plane of D and A the lines of existence of negative kinks issue from the singular point $D = 0$, $A = 1$. With growth of D the corresponding values of A rapidly decrease; in the

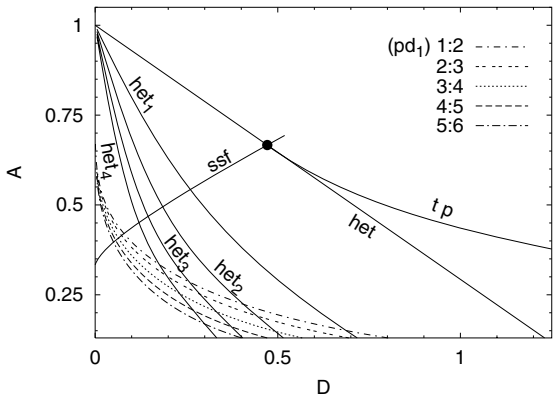


FIG. 2. *Parameter plane for small values of D. $het, het_1, \ldots, het_4$: curves of existence of heteroclinic contours; ssf: equilibria turn from saddle-foci into saddles; tp: ultimate saddle-center bifurcation of the main family; p:q-marked dash-dotted lines: branching of period-q-upled orbits from the main periodic solution, multipliers of the latter being* $\exp(2\pi ip/q)$*; filled circle: point* $(D_0, A = 1 - D_0/\sqrt{2})$.

domain of small A, where the equilibrium solutions approach each other (recall that $A = 0$ corresponds to the saddle-center bifurcation), these lines become involved in the intricate pattern of the "cocoon bifurcation" [38]. Bifurcation curves for periodic solutions were obtained with the help of the original code, which combines the Newton algorithm on a suitable Poincaré surface with polynomial continuation in the extended parameter space.

In the phase space, a negative kink provides a connection from $U = U^+ = U^0_-$ to $U = U^- = -U^0_-$; since the positive kink is generic, the "backward" link is always there, and hence the presence of the negative kink implies the existence of a heteroclinic contour that connects the points U^+ and U^-. Accordingly, in the parameter plane each point that represents a negative kink also corresponds to a heteroclinic contour.

The scenario of the breakup of a heteroclinic contour depends on the eigenvalues of participating points of the equilibrium. The lines of existence of all contours start in the domain where the inequality (2.5) holds, all eigenvalues are real, and the equilibria are saddle points. With the increase of D these lines cross the boundary (curve ssf in Figure 2) on which saddles turn into saddle-foci.

The heteroclinic contour that includes the solution (2.7) (located on the line het in Figure 2) plays the principal role in the evolution of stationary patterns. For this contour the eigenvalues of the points $U^\pm = \pm\sqrt{A}$ are

$$\lambda_1 = \frac{\mp\left(\sqrt{1 - D/\sqrt{2}} + \sqrt{1 - 3D/\sqrt{2}}\right)}{\sqrt{2}},$$

$$\lambda_2 = \frac{\mp\left(\sqrt{1 - D/\sqrt{2}} - \sqrt{1 - 3D/\sqrt{2}}\right)}{\sqrt{2}},$$

$$\lambda_3 = \pm\sqrt{1 - D/\sqrt{2}} \cdot \sqrt{2}.$$

These eigenvalues are real if $D \le D_0 = \sqrt{2}/3 \approx 0.47$; otherwise two of them are complex conjugate and the third one is real.

2.3. Bifurcations of periodic solutions. Let us now consider solutions of (2.2) satisfying the periodicity condition

$$U(x+l) = U(x), \quad l > 0. \tag{2.8}$$

Obviously, for this class of solutions

$$A = \langle U^2 \rangle = \frac{1}{l}\int_0^l U^2(x)dx. \tag{2.9}$$

Since the dynamical system (2.2) is volume-preserving, the product of Floquet multipliers (eigenvalues of the monodromy matrix) for every periodic solution equals 1. Accordingly, such solutions have either two complex-conjugate multipliers on the unit circle (periodic orbit of *elliptic* type) or two real eigenvalues, one of them outside of the unit circle (periodic orbit of *hyperbolic* type). For an elliptic orbit, variation of parameters lets multipliers $\mu_{1,2} = \exp(\pm 2\pi i\phi)$ wander along the circle; passage of ϕ through each rational number p/q is accompanied by the "period q-tupling": branching of a periodic solution whose period is q times larger than the period of the original orbit. The details of the branching depend on q [39, 40]. For $q = 2$ and $q \ge 5$ the bifurcation is "one-sided": in the parameter space the newborn periodic solutions exist on only one side of the branching point. For $q = 3$ and $q = 4$, on the contrary, the bifurcation is "two-sided" (transcritical): new branches exist on both sides of the critical parameter value. On one of these branches the periodic solutions are of elliptic type (and, therefore, give rise to secondary sequences of similar bifurcations); the other branch corresponds to solutions of hyperbolic type.

One can see that, due to the symmetry properties of the equation, if $U(x)$ is a periodic solution of (2.2) with a certain value of A and

$$\langle U \rangle = \frac{1}{l}\int_0^l U(x)dx = C,$$

then $\tilde{U}(x) = -U(2x_0 - x)$, $x_0 = const$, is also a solution of (2.2) for the same value of A; obviously, $\langle \tilde{U} \rangle = -C$. The solutions $U(x)$ and $\tilde{U}(x)$ can coincide. In this case the solution $U(x)$ is odd (antisymmetric to reflections) with respect to the point $x = x_0$, and $C = 0$.

In the present paper, we focus on the "main" family of odd stationary periodic solutions: the family which bifurcates from the trivial solution $U(x) = 0$ with the increase of A from zero. In the case of the CH equation, $D = 0$, this family exhausts the set of odd stationary solutions that can be calculated analytically (using elliptic Jacobi functions). The wavelength l grows monotonically with A. In the limit $A \to 1$, the wavelength tends to infinity, and the periodic solution is transformed into heteroclinic (kink) solutions $U(x) = \pm\tanh((x - x_0)/\sqrt{2})$.

For $D \ne 0$, the main family with finite values of A was calculated numerically. For small values of A the periodic orbit is elliptic; with the increase of A it undergoes a countable set of bifurcations that create new periodic orbits (see [39, 40]). Several curves corresponding to such bifurcations are shown by dashed and dash-dotted lines in Figure 2. The last of those bifurcations is the period-doubling, which occurs on the line pd_1 where the main solution has two multipliers equal to –1; immediately beyond this line the solution is hyperbolic.

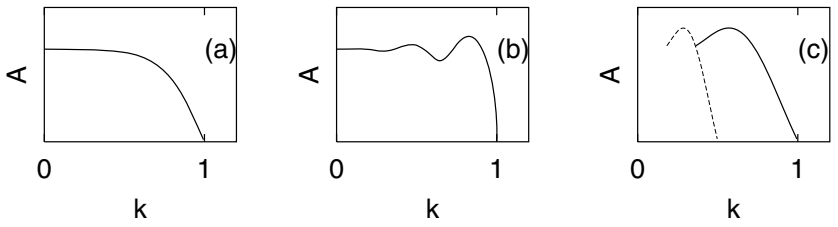


FIG. 3. *Sketch of the dependence $A = \langle U^2 \rangle$ on the wavenumber K of the main stationary solution.* (a) $D < D_0 = 0.4714\ldots$, (b) $D_0 < D < D_1 = 0.8254\ldots$, (c) $D > D_1$; *the dashed curve corresponds to the stationary solution with doubled period. The main family terminates in the "noose bifurcation."*

Depending on D, one can distinguish three different types of interrelation between A and the wavelength of the periodic solution (Figure 3).

In the region $0 < D < D_0 = \sqrt{2}/3$, the main family exists for $0 < A < A_- = 1 - D/\sqrt{2}$, i.e., just for all A below the value corresponding to the principal heteroclinic contour discussed above (see Figure 2, line *het*). The dependence of the wavelength l on the parameter A (or the dependence of A on the wavenumber $K = 2\pi/l$) is monotonic, and the wavelength tends to infinity as A approaches A_-.

In the region $D_0 < D < D_1 \approx 0.8254$, the solutions which belong to the main family exist for any $l > 2\pi$, but the dependence of A on l becomes nonmonotonic, and thus the inverse dependence of l on A is not univalued (see Figure 3(b)). There exists a countable set of values of l for which the function $A(l)$ has maxima and minima. At the extrema of $A(l)$ that can be called *turning points*, a pair of periodic solutions appears/disappears through a saddle-center bifurcation. Note that in the turning point both multipliers of the periodic orbit equal 1; accordingly, each turning point is a joint of two segments: one hyperbolic and the other elliptic. On the elliptic segment a new set of q-tupling bifurcations takes place. The absolute maximum of $A(l)$ is, in a sense, the "ultimate" saddle-center bifurcation: numerical evidence suggests that beyond this value of A no periodic solutions of (2.2) exist. In the bifurcation diagram of (2.2), displayed in Figure 2, the last saddle-center bifurcation is shown by the line *tp*, which (as well as the countable set of lines of other saddle-center bifurcations) starts in the point $D = D_0$, $A = 1 - D_0/\sqrt{2} = 2/3$, and proceeds to higher values of D.

In the case $D > D_1$, the configuration of the main branch of periodic solutions changes again (see Figure 3(c)): there remains only one turning point. Plotting, instead of $l(A)$, the dependence of a suitable geometrical characteristic of the orbit (we use for this purpose the value of du/dx at $u = 0$) on A, we observe in Figure 4 that the curve of the main family goes back and closes onto itself, forming a kind of a noose. This phenomenon was described in the context of the KS equation [36, 37, 41], where, for obvious reasons, it was named the *noose bifurcation*. The noose closes in the point of the period-doubling bifurcation (line pd_1 in bifurcation diagram of Figure 2). More precisely, the stationary periodic solution, antisymmetric with respect to the point $x = x_0$, can be expanded into a Fourier series

$$U(x; A) = \sum_{n=1}^{\infty} a_n(A) \sin \frac{2\pi n}{l}(x - x_0). \tag{2.10}$$

On the upper branch of the noose, when A decreases and approaches the bifurcation

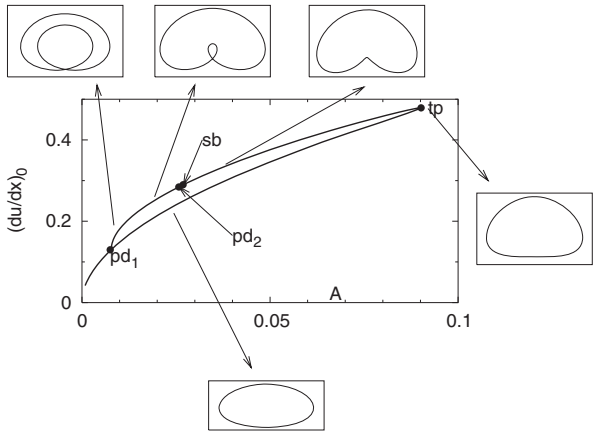


FIG. 4. *Transformations of the phase portrait along the noose branch; $D = 5$. Bifurcation points: $pd_{1,2}$ = period-doubling bifurcations, tp = turning point (saddle-center), sb = symmetry-breaking (pitchfork).*

value A_{pd_1}, all coefficients with odd values of n tend to zero. Thereby, the solution with the period l from the upper branch tends to the solution with the period $2l$ from the lower branch. Subplots of Figure 4 show different stages of continuous transformation of the phase portrait from the closed curve with two loops into the curve with one loop. An additional loop of the orbit, created by the period-doubling bifurcation, decreases in size during the motion along the upper branch of the noose and finally disappears in the small cusp.

Note that the period-doubling bifurcation pd_1 exists for all nonzero values of D, but at $D < D_1$ the orbit with the doubled period which branches off at pd_1 does not return to the main family. At $D = D_1$ a reconnection of branches by means of transcritical bifurcation takes place. Let us have a closer look at the details of this transition, since it is typical for the changes which transform the relatively simple set of stationary periodic solutions of the CH equation into the rich and diversified set of such solutions of the KS equation.

Along with the main family, we track several families of relatively simple secondary orbits which branch off the main family in the course of, respectively, period-doubling, period-tripling, and period-quadrupling. It is known that for reversible systems the period-tripling is a transcritical bifurcation: the 3-looped periodic orbit exists on both sides of the parameter value at which the multipliers of the main (1-looped) orbit equal $\exp(\pm 4\pi i/3)$.

For very small values of D the behavior of periodic states with the increase of A is reminiscent of the behavior typical of pure CH equations: their wavelength is a monotonically growing function of A, and solutions which bifurcate from the main family end up in respective heteroclinic contours. Due to pronounced additional humps of negative kinks, a contour formed by an m-humped negative kink and its positive counterpart is a closed curve with $m + 1$ loops. Accordingly, the periodic solution born via the period-doubling bifurcation on the line pd_1 of the parameter plane terminates on the line het_1, the solution born on the line $2 : 3$ ends up on the line het_2, etc. As D grows, the curves of existence of contours cross the line ssf (Figure 2), the saddle points in the contour are replaced by saddle-foci, and the periodic orbits approach such contours in a nonmonotonic way, developing turning

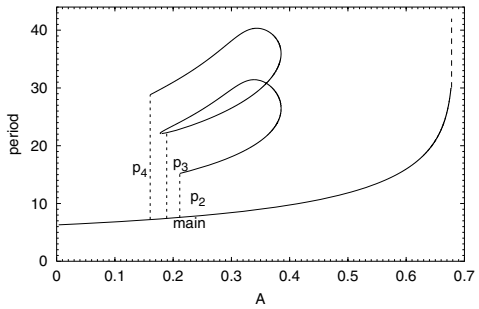


FIG. 5. *Dependence of period on A for periodic states at $D = 0.45$. Dashed lines are bifurcations of the main family: p_2 = period-doubling, p_3 = period-tripling, p_4 = period-quadrupling.*

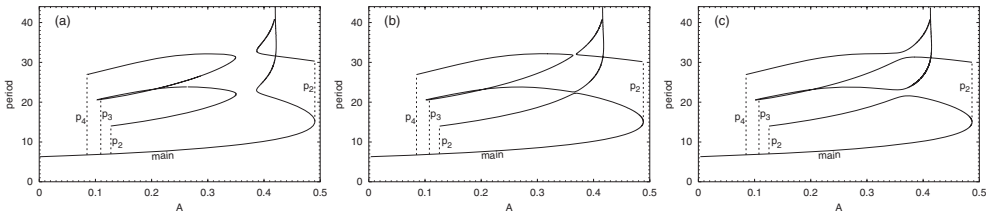


FIG. 6. *Recombinations of curves of periodic solutions. Notation as in Figure* 5. (a) $D = 0.82$, (b) $D = 0.8256$, (c) $D = 0.84$.

points and segments of ellipticity. As D is further increased, the families of periodic orbits detach from heteroclinic contours, and the respective curves on the "period-A" diagram recombine. The configuration of those curves, typical for values of D slightly below D_0, is presented in Figure 5.

It can be seen that the branch which is born from the main solution in the period-doubling bifurcation joins (by means of the turning point) one of the branches of the period-tripled family, whereas the second "tripled" branch matches with the period-quadrupled branch. Thereby, the doubled, tripled, and quadrupled orbits form, in fact, one family, in which the shape of the orbit continuously changes, acquiring and losing loops in the phase space. A similar effect for solutions of the KS equation was described in [42].

As discussed above, the turning point appears on the curve of the main family as soon as D exceeds the value D_0. The corresponding interval of ellipticity is bounded by the point in which both multipliers equal –1. This adds a new element to the picture shown in Figure 5: an additional period-doubling bifurcation of the main family which occurs only for $D > D_0$. In the parameter plane the corresponding bifurcation curve pd_3 is located slightly below the curve tp. (In Figure 2 they practically coincide graphically, and therefore pd_3 is not shown.) Like the main family, the newborn 2-looped solution has a (twice encircled) principal heteroclinic contour as an asymptote: as A is varied, its wavelength grows unboundedly.

This configuration of periodic curves exists until D reaches the value D_1; it is easily recognizable in Figure 6(a). When D is increased, the rightmost turning point of the main family approaches the turning point that joins the period-doubled and the period-tripled family. At $D = D_1 = 0.8254\ldots$ these turning points coalesce, and

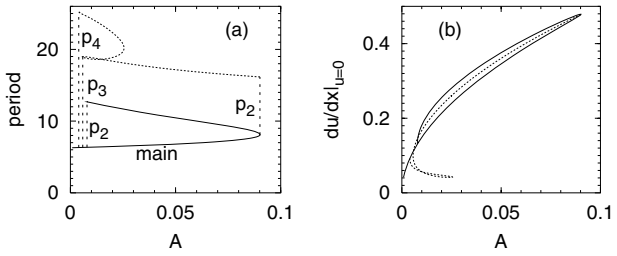


FIG. 7. *Periodic solutions at $D = 5$. Notation as in Figure 5. Solid line: main family; dotted line: secondary solutions.* (a) *Dependence of period on A;* (b) *geometric characteristics of orbit as a function of A.*

a transcritical bifurcation takes place. The segments of the curves recombine: for $D > D_1$ the main family joins the period-doubled branch (thereby forming a noose), whereas the period-tripled branch acquires the asymptotics of long periods and ends up in the principal heteroclinic contour.

Solution curves for $D = 0.8256$ (i.e., just above D_1) are shown in Figure 6(b). Besides the noose formed by the main family, one can notice that the system is close to the next transcritical bifurcation: the turning point which joins the second period-tripled and period-quadrupled branches gets close to the turning point on the curve of the period-doubled solution born at pd_3. This transcritical bifurcation takes place at $D = 0.8257\ldots$; results of recombination of solution curves are seen in Figure 6(c): now the period-tripled branch ends at the period-doubling bifurcation, whereas the period-quadrupled solution terminates in the heteroclinic contour.

In Figure 6(c) there are still two branches of periodic solutions connected to the principal heteroclinic contour. Since for $D > \sqrt{2}$ the contour is absent, those branches detach from it in the course of increase of D and interconnect. As a result, for moderate and high values of D one observers the picture typical of the KS equation [42] (Figure 7(a)). In terms of geometric characteristics of the orbit, the discussed families of secondary solutions merge into a single curve which has two turning points. It starts at the period-quadrupling bifurcation of the main family and terminates in the period-doubling bifurcation close to the turning point of the main family (Figure 7(b)).

Let us mention two more bifurcations of the main periodic solution, which occur in the range of moderate and large values of D and produce new families of stationary periodic solutions different from the main family. The *symmetry-breaking* bifurcation *sb* (see Figure 8) is a pitchfork bifurcation which generates two families of stationary periodic solutions $U_{up}(x)$, $U_{down}(x) = -U_{up}(2x_0 - x)$ with *nonzero* mean values $\langle U_{up}\rangle = -\langle U_{down}\rangle \neq 0$. These two types of solutions are shown in Figure 9(a),(b).

In this bifurcation point the periodic orbit has two multipliers equal to 1. The symmetry-breaking bifurcation, which happens only in presence of the noose, i.e., for $D > D_1$, is a consequence of the symmetry caused by reversibility of the dynamical system (2.2); for families of solutions which do not share the left-right symmetry (e.g., for traveling waves) such bifurcation is precluded. In the course of further evolution the wavelength of the solutions $U_{up}(x)$, $U_{down}(x)$ tends to infinity, and they terminate in the homoclinic orbits of equilibrium points $\pm\sqrt{A_h}$, respectively, where A_h corresponds to the line hom_1 in Figure 8. The other important bifurcation is also a period-doubling bifurcation (line pd_2 in Figure 8) that produces a new family of odd stationary periodic solutions with a zero mean value, which looks like a sequence of

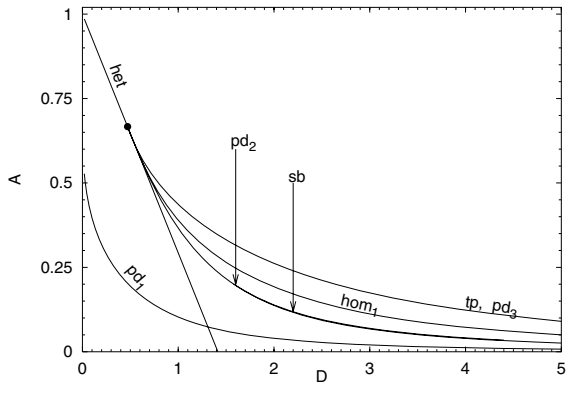


FIG. 8. *Basic bifurcation curves of the main family and its offspring. het: existence of the principal heteroclinic contour; tp: ultimate saddle-center bifurcation of the main family; pd_1, pd_2, pd_3: period-doubling bifurcations of the main family; sb: symmetry-breaking bifurcation (creation of "ups" and "downs"); hom_1: formation of homoclinic orbits to saddle-focus equilibria; filled circle: point* $(D_0,\, A = 1 - D_0/\sqrt{2})$.

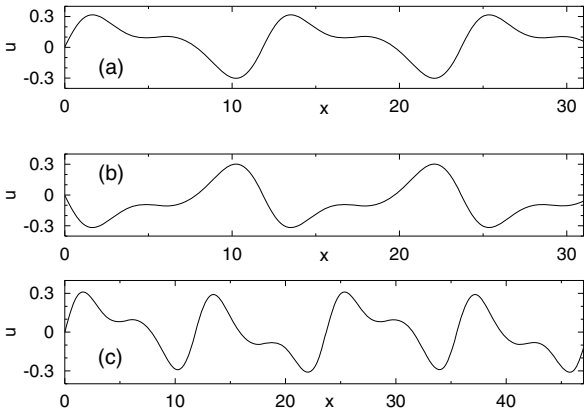


FIG. 9. *Solutions of* (2.2) *for* $D = 5$: (a) *the "up" solution,* (b) *the "down" solution,* (c) *the superposition of "up" and "down."*

alternating "up"- and "down"-elements as shown in Figure 9(c). The corresponding bifurcation value of A for the pd_2-bifurcation is slightly lower than that for the *sb*-bifurcation.

As seen in Figure 8, in the parameter plane the point $(D_0,\, A_0 = 1 - D_0/\sqrt{2})$ serves as a kind of an organizing center, origin of several bifurcation curves (in fact, there is also a countable set of curves of secondary homoclinics to a saddle-focus U^+, which starts from this point).

3. Stability of stationary periodic solutions. In this section we investigate the linear stability of the above-mentioned *main family* of spatially periodic stationary solutions of the convective CH equation (2.1) with respect to arbitrary infinitesimal perturbations in the infinite region. Stability of other families of stationary solutions of (2.1) is beyond the scope of the present paper.

3.1. Eigenvalue problem. Let $U(x)$ be a stationary periodic solution satisfying (2.1) with periodic boundary conditions $U(x) = U(x + l)$. Recall that for the main family of periodic solutions

$$\int_0^l U(x)dx = 0,$$

because $U(x)$ is odd.

Add a small perturbation $\widetilde{u}$ to the stationary solution, and linearize (2.1) to obtain the following problem for $\widetilde{u}$:

$$\widetilde{u}_t = -\widetilde{u}_{xx} - \widetilde{u}_{xxxx} + D(U\widetilde{u})_x + 3(U^2\widetilde{u})_{xx}, \tag{3.1}$$

$$|\widetilde{u}| < \infty \quad \text{for } |x| \to \infty.$$

Following the standard approach of the linear stability theory, consider normal-mode solutions of (3.1),

$$\widetilde{u}(x,t) = e^{t\sigma}V(x;\sigma). \tag{3.2}$$

Substitute (3.2) into (3.1) to obtain the equation

$$V(x)\sigma = -V_{xx} - V_{xxxx} + D(UV)_x + 3(U^2V)_{xx}. \tag{3.3}$$

Generally, (3.3) has four linearly independent solutions for any value of x, $V_j(x)$, $j = 1, 2, 3, 4$. Because of the periodicity of $u(x)$, the functions $V_j(x)$ and $V_j(x + L)$ are simultaneously the solutions of (3.3). Therefore,

$$V_j(x + L) = \sum_{k=1}^{4} B_{jk}(\sigma)V_k(x), \tag{3.4}$$

where $\hat{B}(\sigma) = \{B_{jk}(\sigma)\}$ is the monodromy matrix. We are interested only in bounded solutions of (3.3),

$$V(x) = \sum_{k=1}^{4} C_jV_j(x), \qquad |V(x)| < \infty \quad \text{for } x \to \pm\infty, \tag{3.5}$$

which correspond to the eigenvectors of the monodromy matrix,

$$\sum_{k=1}^{4} B_{jk}(\sigma)C_k = B(\sigma)C_j, \tag{3.6}$$

with the eigenvalues satisfying the additional condition,

$$|B(\sigma)| = 1. \tag{3.7}$$

Hence, one can represent the eigenfunctions $V(x)$ in the form

$$V(x) = v(x)e^{iqx}, \tag{3.8}$$

where $v(x)$ is a periodic function, $v(x+L) = v(x)$, and $q = \arg(B(\sigma))/L$ is a real number ("quasi wavenumber"). The function $V(x)$ is called the Floquet–Bloch function [43, 44].

The quasi wavenumber q can be taken from the interval

$$-\frac{\pi}{l} \le q \le \frac{\pi}{l} \tag{3.9}$$

(i.e., inside the "Brillouin zone" [45]). Substituting (3.8) into (3.3), one obtains the following eigenvalue problem:

$$\begin{aligned} \sigma(q)v(x) &= \mathcal{D}_q^2[(3U^2(x)-1)v(x)] - \mathcal{D}_q^4 v(x) + D\mathcal{D}_q[U(x)v(x)], \\ v(x+l) &= v(x), \end{aligned} \tag{3.10}$$

where $\mathcal{D}_q = d/dx + iq$. If there exists an eigenvalue $\sigma(q)$ with $\mathrm{Re}\sigma(q) > 0$, the solution $U(x)$ is unstable.

3.2. Long-wave asymptotics. For $q = 0$, (3.10) always has a solution $v(x) = U_x(x)$ corresponding to $\sigma(0) = 0$. Indeed, differentiate (2.2) twice with respect to x, to find

$$-v_{xxxx} - v_{xx} + D(Uv)_x + 3(U^2 v)_{xx} = 0. \tag{3.11}$$

The disturbance $v(x) = U_x(x)$ corresponds to an infinitesimal homogeneous translation of the stationary solution $U(x)$. One can expect that $\mathrm{Re}\sigma(q)$ can become positive at small q. In the present subsection we investigate the eigenvalue problem for long-wave disturbances, which correspond to nonhomogeneous translations applied to the stationary solution.

First, let us introduce the variable $X = Kx$, where $K = 2\pi/l$ is the wavenumber of the stationary regime, and rewrite the problem for the main-family stationary solutions as

$$\begin{aligned} -K^4 U'''' - K^2 U'' + DKUU' + K^2 (U^3)'' &= 0, \\ U(X+2\pi) &= U(X), \end{aligned} \tag{3.12}$$

where a prime means differentiation with respect to X. Note that

$$\int_0^{2\pi} U(X)dX = 0. \tag{3.13}$$

Later on, we shift the origin $X = 0$ in such a way that the stationary solution is an odd function of X, and therefore its Fourier series is

$$U = \sum_{n=1}^{N} a_n(K) \sin nX. \tag{3.14}$$

Equation (3.10) can be written as

$$\begin{aligned} \sigma(Q)v(X) &= K^2\mathcal{D}_Q^2[(3U^2(X)-1)v(X)] - K^4\mathcal{D}_Q^4 v(X) + DK\mathcal{D}_Q[U(X)v(X)], \\ v(X+2\pi) &= v(X), \end{aligned} \tag{3.15}$$

where $\mathcal{D}_Q = d/dX + iQ$, $Q \equiv q/K$. According to (3.9), one can choose $|Q| \le 1/2$.

Let Q be small. We are looking for a solution in the form

$$v = \sum_{n=0}^{\infty} v_n Q^n, \qquad \sigma = \sum_{n=0}^{\infty} \sigma^{(n)} Q^n, \tag{3.16}$$

where v_n are periodic functions of X with period 2π. Substituting (3.16) into (3.15), one obtains the sequence of problems in successive orders of Q.

In the zeroth order of Q one has

$$-K^4 v_0'''' + K^2[(3U^2-1)v_0]'' + DK(Uv_0)' - \sigma^{(0)} v_0 = 0, \tag{3.17}$$

$$v_0(X+2\pi) = v_0(X),$$

where prime denotes differentiation with respect to X. As shown above, the solution $v_0 = U'$, $\sigma^{(0)} = 0$ always exists. Later on, we assume that there are no other 2π-periodic solutions with $\sigma^{(0)} = 0$ for the chosen values of K and D. This assumption fails for the values of parameters corresponding to the symmetry-breaking bifurcation (line *sb* in Figure 8).

For every integer $i > 0$, the equation for v_i has the form

$$-k^4 v_i'''' + k^2[(3U^2-1)v_i]'' + Dk(Uv_i)' = R_i, \qquad v_i(X+2\pi) = v_i(X), \tag{3.18}$$

where R_i is some linear combination of the functions $v_j(X)$, $0 \le j < i$, and their derivatives. One can show that the left-hand side of (3.18) is a derivative of a periodic function. Therefore, the integral of the right-hand-side, R_i, over X from 0 to 2π must be zero.

In the first order of Q one obtains

$$\begin{aligned} &-K^4 v_1'''' + K^2[(3U^2-1)v_1]'' + DK(Uv_1)' \\ &= \sigma^{(1)} v_0 - iDUv_0 - 2iK^2[(3U^2-1)v_0]' + 4iK^4 v_0''', \end{aligned} \tag{3.19}$$

$$v_1(X+2\pi) = v_1(X).$$

In order to find the solution of (3.19), let us differentiate (3.12) with respect to the parameter K:

$$\begin{aligned} &-K^4[\partial_K U]'''' + K^2[(3U^2-1)\partial_K U]'' + DK[U\partial_K U]' \\ &\quad - 4K^3 U'''' - 2KU'' + DUU' + 6K[U^2U']' = 0. \end{aligned} \tag{3.20}$$

Comparing (3.19) with (3.20), one finds

$$v_1 = iK\partial_K U - \sigma^{(1)} f_1(X) + const\ U', \tag{3.21}$$

where $f_1(X)$ is the solution of the equation

$$-K^4 f_1'''' - K^2 f_1'' + DK(Uf_1)' + 3K^2(U^2 f_1)'' = U',$$

$$f_1(X+2\pi) = f_1(X).$$

Note that $f_1(X)$ is an even function.

In the second order of Q,

$$\begin{aligned} &-K^2 v_2'' - K^4 v_2'''' + DK(Uv_2)' + 3K^2(U^2 v_2)'' \\ &= (\sigma^{(1)} - iDKU)v_1 + 4iK^4 v_1''' + 2iK^2[(1-3U^2)v_1]' \\ &\quad + \sigma^{(2)} v_0 - 6K^4 v_0'' + K^2(3U^2-1)v_0, \end{aligned} \tag{3.22}$$

$$v_2(X + 2\pi) = v_2(X).$$

Integrating (3.22) over the period and substituting into (3.21), one finds

$$\left[\sigma^{(1)}\right]^2 = -\frac{K^2 \int_0^{2\pi} dX U \partial U/\partial K}{\int_0^{2\pi} dX f_1(X)} = -\frac{K^2}{2}\frac{dA(K)/dK}{\langle f_1 \rangle}, \tag{3.23}$$

where $A = \langle U^2 \rangle$.

If $[\sigma^{(1)}]^2 > 0$, two real roots with opposite signs exist. One of them is positive; therefore the stationary solution is unstable. If $[\sigma^{(1)}]^2 < 0$, there are two imaginary roots. In the latter case, the stability with respect to perturbations with small Q depends on the sign of $\sigma^{(2)}$, which can be obtained in a similar way from the equation in the third order of Q.

Thus, one can expect that there are boundaries between the regions of stability and instability of stationary solutions with respect to long-wave disturbances that coincide with the extrema of the function $A(K)$.

Note that a similar relation between the nonmonotonicity of the dependence between the amplitude and the wavenumber was established for the KS equation [7] and the reaction-diffusion equation [46].

3.3. Numerical method for finding eigenvalues. For arbitrary values of q the problem (3.10) is solved numerically. We use the *Fourier transform* of a function $v(x, q)$,

$$v(x,q) = \sum_{n=-\infty}^{\infty} \widehat{v}_n(q) e^{inkx}, \qquad k = \frac{2\pi}{l}. \tag{3.24}$$

Substituting (3.24) into (3.10), we get the following matrix equation for the eigenvalues and eigenvectors:

$$\sum_{n=-\infty}^{\infty} [M_{mn} - \sigma(q)\delta_{mn}]\widehat{v}_n = 0, \qquad -\infty < m < +\infty, \tag{3.25}$$

where

$$\begin{aligned} M_{mn} &= [(mk+q)^2 - (mk+q)^4]\delta_{mn} \\ &\quad + iD(mk+q)\widehat{U}_{m-n} - 3(mk+q)^2(\widehat{U^2})_{m-n}. \end{aligned} \tag{3.26}$$

The infinite system of equations (3.25) was replaced by a finite system,

$$\sum_{n=-N}^{N} [M_{mn} - \sigma(q)\delta_{mn}]\widehat{v}_n = 0, \qquad -N < m < N, \tag{3.27}$$

and solved numerically using Matlab. Numerical calculations were done with double precision. The results of the numerical calculations are described in the next subsection.

3.4. Results of the numerical analysis. First, we check that our approach confirms the known results concerning the stability of stationary periodic solutions for the KS equation (the case $D \to \infty$). Indeed, for the KS equation we have reproduced the results formerly obtained in [7, 8, 9, 10, 11, 12]. The stability interval is $0.766 < K < 0.838$. The left end of the stability interval is determined by the condition $\sigma^{(2)} = 0$ in the expansion (3.16). Here we observe an oscillatory instability for small quasi wavenumbers, $q \to 0$. At the right end of the interval, which coincides with the maximum of the function $A(K) = \langle U^2 \rangle(K)$, we obtain a monotonic instability, also for $q \to 0$.

For $D \le D_0$ the dependence $A(K)$ for stationary solutions is monotonic (see section 2), and it turns out that $[\sigma^{(1)}]^2 > 0$ for any K (see section 3.2). Therefore, there is no stability interval in this region.

TABLE 1

Boundaries of the stability intervals. Abbreviations: osc. is oscillatory instability, mon. is monotonic instability.

D	K_{left}	K_{right}
∞	0.766 osc.,$q \to 0$	0.838 mon.,$q \to 0$
5	0.677 osc.,$q \to 0$	0.775 mon.,$q \to 0$
2	0.537 osc., $q \to K/2$	0.640 mon.,$q \to 0$
1	0.376 osc.,$q \to K/2$	0.4752 mon.,$q \to 0$
0.8	0.314059 mon.,$q \to 0$	0.4065 mon.,$q \to 0$
0.5	0.111 mon.,$q \to K/2$	0.1767048 mon.,$q \to 0$

The investigation of the stability of main-family stationary solutions for $D > D_0$, where the dependence $A(K)$ is not monotonic, has been done numerically. The eigenvalue problem (3.27) determines $2N+1$ eigenvalues. Let $\sigma_1(q)$ be the eigenvalue with the maximum real part (see Figures 10–13). The analysis reveals the existence of the stability interval. The boundaries of the stability interval are presented in Table 1. Both boundaries move into the long-wave region with the decrease of D and tend to zero when D approaches $\sqrt{2}/3$. In order to obtain accurate eigenvalues for decreasing D it was sufficient to increase the size of matrix (number N in (3.27)). For example, for $D = 5$, $D = 2$, and $D = 1$ we used matrices of the size $N = 17$; for $D = 0.8$, $N = 33$; and for $D = 0.5$, $N = 33$ and $N = 129$. For each case we tried to find the minimum size of the matrix that allowed us to obtain the correct result for a given value of D. This was done by comparing the solutions obtained for different N and checking the convergence.

We have found that the right boundary of the stability interval always coincides with the global maximum of the function $A(K)$, which corresponds to (3.23) with $\langle f_1 \rangle > 0$. Therefore, on the right boundary of the stability interval a monotonic

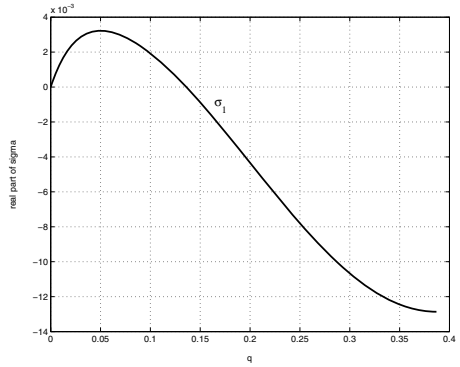


FIG. 10. *Dependence of σ_1, the eigenvalue with the maximal real part, on q; $D = 5$, $K = 0.775$. The case of long-wave monotonic instability.*

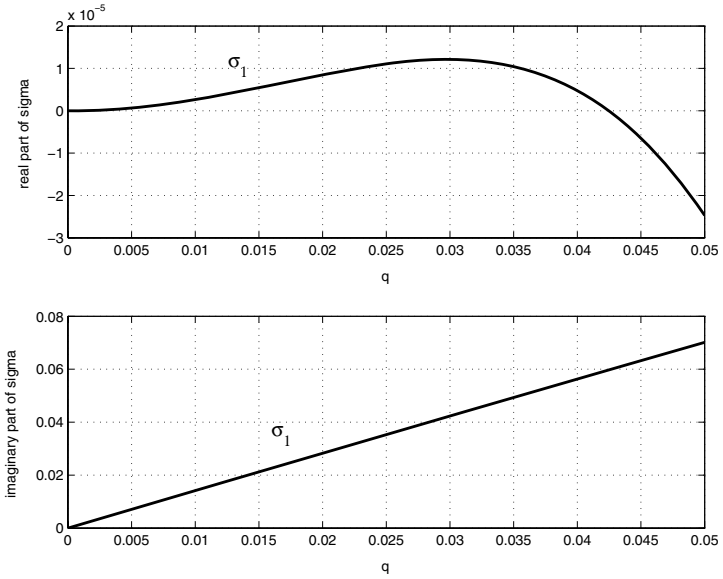


FIG. 11. *The real and imaginary parts of σ_1, for small q; $D = 5$, $K = 0.677$. The oscillatory long-wave instability is found.*

long-wave instability always occurs (see Figure 10).

The type of the instability on the left boundary of the stability interval depends on D. For sufficiently large values of D ($D = 5$), as in the case of the KS equation, the destabilization of the stationary solutions for $K < 0.677$ is due to the oscillatory instability with $q \to 0$ (Figure 11). One can see from Figure 11 that for $q \ll 1$ $\mathrm{Im}[\sigma_1(q)] = O(q)$, $\mathrm{Re}[\sigma_1(q)] = O(q^2)$ for $q \ll 1$. This is in accordance with the results described in section 3.2.

For smaller values of D ($D = 2$ and $D = 1$), the dependence of $\mathrm{Re}[\sigma]$ on the quasi wavenumber q has two maxima near the left stability boundary. We have found that the destabilization scenario in this case is different: now the most "dangerous" oscillatory perturbation (the one with the largest real part of σ_1) at the left end of the stability interval corresponds to $q \to K/2$ (see Figure 12).

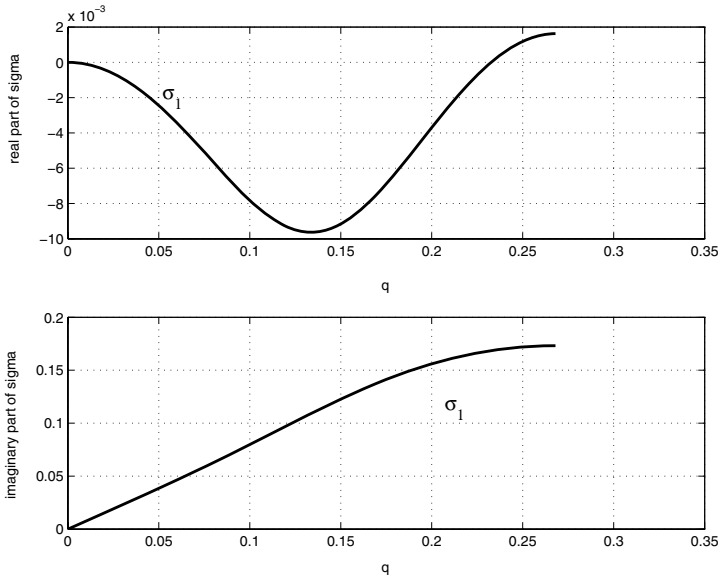


FIG. 12. *Dependence of the real and imaginary parts of σ_1 on q; $D = 2$, $K = 0.537$. The short-wave oscillatory instability is found.*

For $D = 0.8$, we have obtained a monotonic instability with $q \to 0$ at *both* ends of the interval (see Figure 13). Note that the left boundary of the monotonic instability is not connected with the extrema of $A(K)$. As a matter of fact, in the point $K = 0.314059$ the eigenvalue problem (3.17) has one more eigenfunction with $\sigma^{(0)} = 0$, in addition to the solution $v_0 = U'$. This eigenfunction has a nonzero mean value and is related to the bifurcation *sb* of another family of asymmetric stationary solutions with a nonzero mean value (see Figure 8).

For smallest values of D ($D = 0.5$), at the left end of the interval we have a monotonic instability for $q \to K/2$.

In [30], equation (2.1) was solved numerically in a long domain with periodic boundary conditions. It was found that stationary solutions with a definite wavelength are selected in the interval $1 < D < 5$ (see Figure 14). It is clear that the black points which correspond to the stationary solutions obtained in simulations [30] are inside the stability intervals obtained in our computations.

4. Conclusions. We have studied bifurcations of stationary solutions of a convective Cahn–Hilliard equation (2.1) that is generic for a wide class of systems exhibiting phase separation in the presence of external driving (e.g., growth or external field) and have analyzed the stability of the main family of the periodic stationary solutions.

The stability analysis was done for odd stationary periodic solutions of (2.1) (the *main family* of solutions) that bifurcate from the trivial state $u \equiv 0$ as a result of a spinodal decomposition instability. The effect of the driving force, characterized by the parameter D in (2.1), has been investigated. We have found that in the region $0 < D < D_0 = \sqrt{2}/3$ these solutions are unstable. One can expect that the coarsening (Ostwald ripening) of domains with two different phases occurs in this case. For $D > D_0$ the periodic stationary solutions are stable if their wavelength belongs to

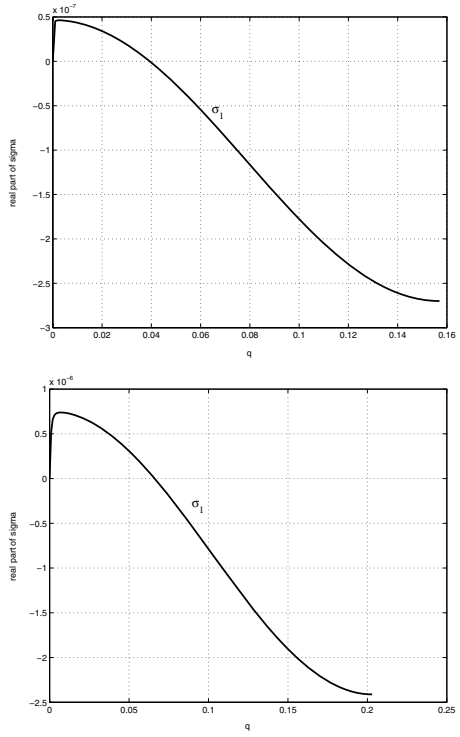


FIG. 13. *Dependence of σ_1 on q, for $D = 0.8$, at the two ends of the stability interval: $K = 0.314059$ (upper figure) and $K = 0.4065$ (lower figure). The case of long-wave monotonic instability.*

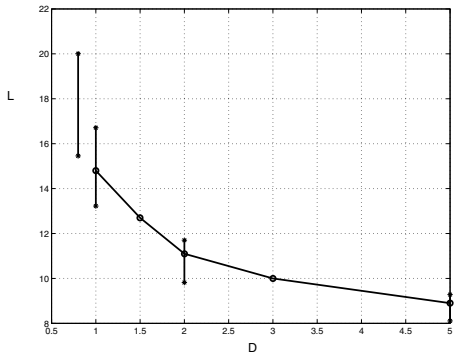


FIG. 14. *Dependence of the solution wavelength on D; • corresponds to the stationary regime, obtained for selected values of D in* [30]. *Vertical lines correspond to the stability intervals for certain D, and $*$ corresponds to the boundaries of the stability intervals obtained in our investigation.*

a certain stability interval. The instabilities of the periodic stationary solutions out of the stability interval can give rise to other classes of solutions that may be stable.

Thus, we have shown that in a driven phase-separating system that undergoes spinodal decomposition, the coarsening can be stopped by the driving force, and the formation of stable periodic structures is possible. This shows an opportunity to control the formation of spatially periodic structures in various phase-separating systems (e.g., phase-separating polymer films) by adjusting the driving force (e.g.,

external field). Note that some two-dimensional generalizations of convective CH models that exhibit the formation of stationary two-dimensional periodic patterns are investigated numerically in [47].

Acknowledgments. The third author acknowledges the hospitality of the ESAM Department at Northwestern University and the support of the Eshbach Society.

REFERENCES

[1] I. S. Aranson and L. Kramer, *The world of the complex Ginzburg–Landau equation*, Rev. Modern Phys., 74 (2002), pp. 99–143.
[2] M. C. Cross and P. C. Hohenberg, *Pattern-formation outside of equilibrium*, Rev. Modern Phys., 65 (1993), pp. 851–1112.
[3] A. A. Nepomnyashchy, *Order parameter equations for long-wavelength instabilities*, Phys. D, 86 (1995), pp. 90–95.
[4] Y. Kuramoto and T. Tsuzuki, *Persistent propagation of concentration waves in dissipative media far from thermal equilibrium*, Progr. Theoret. Phys., 55 (1976), pp. 356–369.
[5] G. I. Sivashinsky, *Nonlinear analysis of hydrodynamic instability in laminar flames.* I. *Derivation of basic equations*, Acta Astronaut., 4 (1977), pp. 1177–1206.
[6] G. M. Homsy, *Model equations for wavy viscous film flow*, in Nonlinear Wave Motion, Lect. Appl. Math. 15, AMS, Providence, RI, 1974, pp. 191–194.
[7] A. A. Nepomnyashchy, *Stability of the wavy regimes in the film flowing down an inclined plane*, Fluid. Dynam., 9 (1974), pp. 354–359.
[8] A. Novick-Cohen, *Interfacial instabilities in directional solidification of dilute binary alloys: The Kuramoto–Sivashinsky equation*, Phys. D, 26 (1987), pp. 403–410.
[9] M. L. Frankel, *On the weakly nonlinear evolution of a perturbed planar solid-liquid interface*, Phys. D, 27 (1987), pp. 260–266.
[10] A. Umantsev and S. H. Davis, *Growth from a hypercooled melt near absolute stability*, Phys. Rev. A, 45 (1992), pp. 7195–7201.
[11] B. I. Cohen, J. A. Krommes, W. M. Tang, and M. N. Rosenbluth, *Nonlinear saturation of dissipative trapped ion mode by mode coupling*, Nuclear Fusion, 16 (1976), pp. 971–992.
[12] U. Frisch, Z. S. She, and O. Thual, *Viscoelastic behavior of cellular solutions to the Kuramoto–Sivashinsky model*, J. Fluid Mech., 168 (1986), pp. 221–240.
[13] T. Yamada and Y. Kuramoto, *Reduced model showing chemical turbulence*, Progr. Theoret. Phys., 56 (1976), pp. 681–683.
[14] T. Bohr, M. H. Jensen, G. Paladin, and A. Vulpiani, *Dynamical System Approach to Turbulence*, Cambridge University Press, Cambridge, UK, 1998.
[15] J. W. Cahn and J. E. Hilliard, *Free energy of nonuniform systems.* I. *Interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–267.
[16] J. W. Cahn and D. W. Hoffman, *Vector thermodynamics for anisotropic surfaces.* II. *Curved and faceted surfaces*, Acta Metall., 22 (1974), pp. 1205–1214.
[17] J. Stewart and N. Goldenfeld, *Spinodal decomposition of a crystal surface*, Phys. Rev. A, 46 (1992), pp. 6505–6512.
[18] A. A. Nepomnyashchy, *Stability of secondary flows of a viscous fluid in unbounded space*, J. Appl. Math. Mech., 40 (1976), pp. 836–841.
[19] B. A. Malomed, A. A. Nepomnyashchy, and M. I. Tribelsky, *Domain boundaries in convective patterns*, Phys. Rev. A, 42 (1990), pp. 7244–7263.
[20] J. S. Langer, *Theory of spinodal decomposition in alloys*, Ann. Phys. (New York), 65 (1971), pp. 53–86.
[21] K. Kawasaki and T. Ohta, *Kink dynamics in one-dimensional non-linear systems*, Phys. A, 116 (1982), pp. 573–593.
[22] L. Bronsard and D. Hilhorst, *On the slow dynamics for the Cahn–Hilliard equation in one space dimension*, Proc. Roy. Soc. London A, 439 (1992), pp. 669–682.
[23] K. Leung, *Theory on morphological instability in driven systems*, J. Statist. Phys., 61 (1990), pp. 345–364.
[24] C. Yeung, T. Rogers, A. Hernandes-Machado, and D. Jasnow, *Phase separation dynamics in driven diffusive systems*, J. Statist. Phys., 66 (1992), pp. 1071–1088.
[25] C. L. Emmott and A. J. Bray, *Coarsening dynamics of a one-dimensional driven Cahn–Hilliard system*, Phys. Rev. E, 54 (1996), pp. 4568–4575.
[26] Y. Saito and M. Uwaha, *Anisotropy effect on step morphology described by Kuramoto–Sivashinsky equation*, J. Phys. Soc. Japan, 65 (1996), pp. 3576–3581.

[27] F. Liu and H. Metiu, *Dynamics of phase separation of crystal surfaces*, Phys. Rev. B, 48 (1993), pp. 5808–5817.
[28] A. A. Golovin, S. H. Davis, and A. A. Nepomnyashchy, *A convective Cahn–Hilliard model for the formation of facets and corners in crystal growth*, Phys. D, 122 (1998), pp. 202–230.
[29] A. A. Golovin, S. H. Davis, and A. A. Nepomnyashchy, *Model for faceting in a kinetically controlled crystal growth*, Phys. Rev. E, 59 (1999), pp. 803–825.
[30] A. A. Golovin, A. A. Nepomnyashchy, S. H. Davis, and M. A. Zaks, *Convective Cahn–Hilliard models: From coarsening to roughening*, Phys. Rev. Lett., 86 (2001), pp. 1550–1553.
[31] S. J. Watson, F. Otto, B. Y. Rubinstein, and S. H. Davis, *Coarsening dynamics of the convective Cahn–Hilliard equation*, Phys. D, 178 (2003), pp. 127–148.
[32] U. Thiele, M. G. Velarde, K. Neuffer, M. Bestehorn, and Y. Pomeau, *Sliding drops in the diffuse interface model coupled to hydrodynamics*, Phys. Rev. E, 64 (2001), paper 061601.
[33] A. DiCarlo, M. E. Gurtin, and P. Podio-Guidugli, *A regularized equation for anisotropic motion-by-curvature*, SIAM J. Appl. Math., 52 (1992), pp. 1111–1119.
[34] M. E. Gurtin, *Thermomechanics of Evolving Phase Boundaries in the Plane*, Clarendon Press, Oxford, UK, 1993.
[35] A. Novick-Cohen and L. A. Segel, *Nonlinear aspects of the Cahn–Hilliard equation*, Phys. D, 10 (1984), pp. 277–298.
[36] Ye. A. Demekhin, G. Yu. Tokarev, and V. Ya. Shkadov, *Hierarchy of bifurcations of space-periodic structures in a nonlinear model of active dissipative media*, Phys. D, 52 (1991), pp. 338–361.
[37] I. G. Kevrekidis, B. Nicolaenko, and J. C. Scovel, *Back in the saddle again: A computer assisted study of the Kuramoto–Sivashinsky equation*, SIAM J. Appl. Math., 50 (1990), pp. 760–790.
[38] Y.-T. Lau, *The "cocoon" bifurcations in three-dimensional systems with two fixed points*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 2 (1992), pp. 543–558.
[39] K. R. Meyer, *Generic bifurcation of periodic points*, Trans. Amer. Math. Soc., 149 (1970), pp. 95–107.
[40] J. J. Gervais, *Bifurcations of subharmonic solutions in reversible systems*, J. Differential Equations, 75 (1988), pp. 28–42.
[41] P. Kent and J. Elgin, *Noose bifurcation of periodic orbits*, Nonlinearity, 4 (1991), pp. 1045–1061.
[42] P. Kent and J. Elgin, *Traveling-waves of the Kuramoto–Sivashinsky equation—Period-multiplying bifurcations*, Nonlinearity, 5 (1992), pp. 899–919.
[43] P. Kuchment, *Floquet Theory for Partial Differential Equations*, Oper. Theory Adv. Appl. 60, Birkhäuser-Verlag, Basel, Switzerland, 1993.
[44] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, Springer, New York, 1999.
[45] C. Kittel, *Quantum Theory of Solids*, John Wiley & Sons, New York, 1987.
[46] P. Politi and C. Misbah, *When does coarsening occur in the dynamics of one-dimensional fronts?*, Phys. Rev. Lett., 92 (2004), paper 090601.
[47] A. A. Golovin and L. M. Pismen, *Dynamic phase separation: From coarsening to turbulence via structure formation*, Chaos, 14 (2004), pp. 845–854.

SIAM J. APPL. MATH.
Vol. 66, No. 2, pp. 721–735

CALIBRATION OF THE EXTENDED CIR MODEL*

HONGTAO YANG†

Abstract. In this paper we shall prove that the calibration problem for the extended CIR model in [J. Hull and A. White, *Rev. Financial Studies*, 3 (1990), pp. 573–592] has a unique solution. The constructive proof leads to a numerical algorithm for computing the approximations of the time-dependent parameters and the zero-coupon bond prices. The results are also extended to multifactor CIR (Cox–Ingersoll–Ross) models. Numerical results are presented to examine the accuracy of our algorithm and to compare the extended CIR model with the Vasicek models.

Key words. extended CIR model, calibration, inverse problem, solution uniqueness and existence, numerical solutions

AMS subject classifications. 15A15, 15A09, 15A23

DOI. 10.1137/S0036139903437357

1. Introduction.

The CIR model of the term structure of interest rates was proposed by Cox, Ingersoll, and Ross in [4]. Unlike the Vasicek model [18], the CIR model does not permit negative interest rates. The interest rates under the CIR model can always be positive under certain conditions, and thus American calls and European calls on zero-coupon bonds have the same value [4], [15]. However, critics of this model (and other models with constant parameters) note that it does not provide a perfect fit to the initial term structure of interest rates. In [7] and [8], Hull and White extended the Vasicek and CIR models to allow time-dependent parameters. The extended models are consistent with both the current term structure of interest rates and either the current volatilities of all spot interest rates or the current volatilities of all forward interest rates. Recent theoretical study on the CIR model and the extended CIR model can be found in [5], [14], and references cited therein.

The important problem both in practice and in theory is to determine the time-dependent parameters of the extended models. Since both the extended Vasicek model and the extended CIR model are affine term structure models, the bond prices can be determined by a system of ordinary differential equations with final conditions (see section 2 for the Extended CIR model). By using the current market date, we can add several initial conditions to the system and then formulate an inverse problem to determine its solution and time-dependent coefficients [7]. The inverse problem can be solved analytically in the case of the extended Vasicek model. However, for the extended CIR model, it seems that numerical methods have to be used. Additionally, to the best of our knowledge, the following is still an open problem: Does the inverse problem proposed in [7] have a solution? The main objective of this paper is to prove that the inverse problem for the extended CIR model has a unique solution under appropriate assumptions. Since our proof is a constructive one, its numerical implementation provides a numerical algorithm for computing the approximations of the time-dependent parameters and the discount bond prices. This makes it possible

*Received by the editors November 4, 2003; accepted for publication (in revised form) August 23, 2005; published electronically January 26, 2006. This research was supported by the Louisiana Board of Regents through the Board of Regents Support Fund under grant LEQSF(2003-06)-RD-A-38.
http://www.siam.org/journals/siap/66-2/43735.html
†Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70504-1010 (hyang@louisiana.edu).

for practitioners to evaluate interest rate derivatives under the extended CIR model when it is fitted to the current market data.

The outline of the paper is as follows. In section 2, we introduce the inverse problem for determining the time-dependent parameters of the extended CIR model. The solution existence and uniqueness of the inverse problem are established by a constructive proof in section 3. We extend the results to the multifactor CIR models in section 4. In section 5, a fourth order numerical algorithm is proposed to compute the approximations of the time-dependent parameters and the bond prices. Numerical results are presented in section 6 to examine the accuracy of our algorithm and to compare the extended CIR model with the Vasicek models. Conclusions are in section 7.

2. An inverse problem to calibrate the extended CIR model. In this section, we introduce the inverse problem for the CIR model proposed in [7]. There are two different economic approaches to the bond market: the equilibrium modeling and the martingale modeling, which lead to the same partial differential equations. Here we shall adopt the latter approach and assume the existence of a risk-neutral measure [1] since we consider only the related mathematical problems.

For the extended CIR model, the process of interest rates, $r(t)$, is assumed to follow the stochastic differential equation

$$dr(t) = (\phi(t) - \psi(t)r(t))dt + \sigma(t)\sqrt{r(t)}dW(t), \tag{2.1}$$

under the risk-neutral measure, where $\psi(t)$ is the speed of adjustment at time t, $\phi(t)/\psi(t)$ is the interest rate of mean reversion at time t, $\sigma(t)$ is a function of t related to the volatility of short rates, and $W(t)$ is a Wiener process. Consider a discount bond with face value \$1 and expiration date T not greater than some positive number T^*. The bond price $P(r,t,T)$ is the solution of the following partial differential equation of the parabolic type (see Proposition 3.4 of [1]):

$$P_t + (\phi(t) - \psi(t)r)P_r + d(t)rP_{rr} - rP = 0, \qquad r > 0,\ 0 < t \le T,$$

with the final condition

$$P(r,T,T) = 1, \quad r \ge 0,$$

where $d(t) = \frac{1}{2}\sigma(t)^2$. It has been proved that $P(r,t,T)$ takes the following form (see [1] and [16]):

$$P(r,t,T) = A(t,T)e^{-B(t,T)r}. \tag{2.2}$$

Then $A(t,T)$ and $B(t,T)$ are the solution of the final value problem

$$A_t(t,T) - \phi(t)A(t,T)B(t,T) = 0, \quad 0 \le t \le T, \tag{2.3}$$

$$B_t(t,T) - \psi(t)B(t,T) - d(t)B^2(t,T) + 1 = 0, \quad 0 \le t \le T, \tag{2.4}$$

$$A(T,T) = 1, \tag{2.5}$$

$$B(T,T) = 0. \tag{2.6}$$

As shown in [7], $A(0,T)$ and $B(0,T)$ can be determined by the current term structure of interest rates and the current term structure of spot rate or forward rate volatilities. The function $\sigma(t)$ can be chosen to reflect the current and future

volatilities of the short-term interest rate. Therefore, we have the following inverse problem:

$$\text{(IP)} \quad \begin{cases} \textit{For given functions } \sigma(t),\ a(t),\ \textit{and } b(t) \textit{ defined on } [0,T^*],\ \textit{determine functions } A(t,T),\ B(t,T),\ \phi(t),\ \textit{and } \psi(t) \textit{ such that they satisfy } (2.3)\text{–}(2.6) \textit{ and } A(0,T)=a(T) \textit{ and } B(0,T)=b(T) \textit{ for } T\in[0,T^*]. \end{cases}$$

In the next section, we shall show that the inverse problem has a unique solution. The solution existence means that the model can be fitted to the current term structure of interest rates. In particular, our constructive proof leads to a natural numerical algorithm for computing the time-dependent parameters $\phi(t)$ and $\psi(t)$ and the bond prices.

3. Solution existence and uniqueness of the inverse problem. First, we study how to determine $B(t,T)$ and $\psi(t)$. Since (2.4) is a Riccati equation, by using the standard transform

$$B(t,T) = -\frac{u_t(t,T)}{d(t)u(t,T)},$$

we have the following second order linear differential equation for the new unknown function $u(t,T)$:

$$-u_{tt}(t,T) + \nu(t)u_t(t,T) + d(t)u(t,T) = 0, \quad 0 < t < T, \tag{3.1}$$

where $\nu(t) = \psi(t) + d'(t)/d(t)$. Let $u_1(t)$ and $u_2(t)$, respectively, be solutions to the following initial value problems:

$$-u_1''(t) + \nu(t)u_1'(t) + d(t)u_1(t) = 0, \qquad 0 < t \le T^*,\ u_1(0) = 1,\ u_1'(0) = 0, \tag{3.2}$$

$$-u_2''(t) + \nu(t)u_2'(t) + d(t)u_2(t) = 0, \qquad 0 < t \le T^*,\ u_2(0) = 0,\ u_2'(0) = 1. \tag{3.3}$$

It is apparent that $u_1(t)$ and $u_2(t)$ are two linearly independent solutions. The general solution of (3.1) is thus a linear combination of $u_1(t)$ and $u_2(t)$. Hence we have

$$B(t,T) = -\frac{c(T)u_1'(t) + u_2'(t)}{d(t)(c(T)u_1(t) + u_2(t))},$$

where $c(T)$ is a function of T. It follows from (2.6) that $c(T)u_1'(T) + u_2'(T) = 0$; i.e., $c(T) = -u_2'(T)/u_1'(T)$. Then we have

$$B(t,T) = -\frac{u_1'(T)u_2'(t) - u_2'(T)u_1'(t)}{d(t)(u_1'(T)u_2(t) - u_2'(T)u_1(t))}, \qquad 0 \le t \le T,\ 0 \le T \le T^*. \tag{3.4}$$

Letting $t = 0$, we get

$$u_1'(T) = g(T)u_2'(T), \quad 0 \le T \le T^*, \tag{3.5}$$

where

$$g(T) = d(0)B(0,T) = d(0)b(T), \quad 0 \le T \le T^*.$$

Substituting (3.5) into (3.4), we have

$$B(t,T) = -\frac{g(T)u_2'(t) - u_1'(t)}{d(t)(g(T)u_2(t) - u_1(t))}, \qquad 0 \le t \le T,\ 0 \le T \le T^*. \tag{3.6}$$

Integrating (3.5) from 0 to t plus integration by parts gives

$$u_1(t) = 1 + g(t)u_2(t) - \int_0^t g'(s)u_2(s)ds, \quad 0 \le t \le T^*. \tag{3.7}$$

Eliminating $\nu(t)$ from (3.2) and (3.3), we obtain

$$\frac{u_1'(t)}{u_2'(t)} = \frac{u_1''(t) - d(t)u_1(t)}{u_2''(t) - d(t)u_2(t)}, \quad 0 \le t \le T^*.$$

Substituting (3.5) and (3.7) into the above equation, we have after simplification

$$u_2'(t) = -\frac{d(t)}{g'(t)} \int_0^t g'(s)u_2(s)ds + \frac{d(t)}{g'(t)}, \quad 0 \le t \le T^*. \tag{3.8}$$

Integration leads to

$$u_2(t) + \int_0^t K(t,s)u_2(s)ds = f(t), \quad 0 \le t \le T^*, \tag{3.9}$$

where

$$f(t) = \int_0^t \frac{d(s)}{g'(s)}ds, \quad K(t,s) = (f(t) - f(s))g'(s).$$

Equation (3.9) is a Volterra integral equation of the second kind for $u_2(t)$. By Theorem 3.1 of [11], it has a unique solution $C[0,T^*]$ (the set of all continuous functions on $[0,T^*]$) if $f(t)$ and $g'(t)$ are continuous in $0 \le t \le T^*$.

Once $u_2(t)$ is determined by (3.9), we can compute $u_1(t)$ by (3.7) and then $B(t,T)$ by (3.6). It follows from (3.3) that

$$\nu(t) = \frac{u_2''(t) - d(t)u_2(t)}{u_2'(t)}.$$

Thus parametric function $\psi(t)$ is determined by

$$\psi(t) = \nu(t) - \frac{d'(t)}{d(t)} = \frac{u_2''(t) - d(t)u_2(t)}{u_2'(t)} - \frac{d'(t)}{d(t)}.$$

By using (3.8) twice, we get the following simple formula for computing $\psi(t)$ in terms of $u_2(t)$ and $u_2'(t)$:

$$\psi(t) = -\frac{g''(t)}{g'(t)} - 2d(t)\frac{u_2(t)}{u_2'(t)}, \quad 0 \le t \le T^*. \tag{3.10}$$

Next we discuss how to determine $A(t,T)$ and $\phi(t)$. Solving (2.3) with the final condition (2.5), we get

$$A(t,T) = \exp\left(-\int_t^T \phi(s)B(s,T)ds\right), \qquad 0 \le t \le T,\ 0 \le T \le T^*. \tag{3.11}$$

Letting $t = 0$ and taking the natural logarithm, we have

$$\int_0^T \phi(s)B(s,T)ds = -\ln(a(T)), \quad T \in [0,T^*], \tag{3.12}$$

where $A(0,T)$ was replaced by $a(T)$. Recall that $B(T,T)=0$ on $[0,T^*]$. The above equation is a Volterra integral equation of the first kind with a degenerate kernel. To remove this difficulty, we differentiate both sides of (3.12) to get

$$\int_0^T \phi(s)B_T(s,T)ds = -\frac{a'(T)}{a(T)}, \quad 0 \le T \le T^*. \tag{3.13}$$

This is still a Volterra integral equation of the first kind. It follows from $B(T,T)=0$ that $B_T(T,T) = -B_t(T,T)$. Letting $t=T$ in (2.4), we have $B_t(T,T)=-1$ and thus $B_T(T,T)=1$ for $0 \le T \le T^*$. Therefore, (3.13) has a nice kernel. Differentiating (3.13), we get the following Volterra integral equation of the second kind for $\phi(t)$:

$$\phi(T) + \int_0^T \phi(s)B_{TT}(s,T)ds = \frac{a'(T)^2 - a(T)a''(T)}{a(T)^2}, \quad T \in [0,T^*]. \tag{3.14}$$

Again, by Theorem 3.1 of [11], equation (3.14) has a unique solution in $C[0,T^*]$ if $B_{TT}(s,T)$ is continuous in $0 \le s \le T \le T^*$ and $a''(T)$ is continuous in $0 \le T \le T^*$.

In order to validate the above calculations, we make the following assumptions:

(H1) $a(T),\ b(T) \in C^2[0,T^*]$, $d(T) \in C^1[0,T^*]$.

(H2) *$a(T)$, $d(T)$, and $b'(T)$ have positive lower bounds on $[0,T^*]$.*

(H3) $a(0)=1$, $b(0)=0$, $b'(0)=1$

(H4) *The derivative of the solution $u_2(t)$ of (3.9) is positive on $[0,T^*]$.*

Here we use $C^k[0,T^*]$ to denote the space of all kth continuously differentiable functions on $[0,T^*]$ for a given positive integer k.

Assumption (H1) is for the regularity requirements of the calculations. Assumption (H3) follows from (2.5), (2.6), and $B_T(T,T)=1$. It is necessary to assume that $a(T)$ and $d(T)$ have positive lower bounds on $[0,T^*]$. Recall that the current bond price $P(r(0),0,T) = A(0,T)\exp(-B(0,T)r(0))$ is a decreasing function of the expiration date T. We may assume that $B(0,T)=b(T)$ is an increasing function of T, which means that $b'(T)$ has a positive lower bound on $[0,T^*]$. This assumption is also implied by $b'(0)=B_T(0,0)=1$. If we evaluate options under the extended CIR (ECIR) model, we need only to know the bond price $P(r,t,T)$ for $t \in [0,S]$, where S is the option expiration date. In this case, it follows from (3.6) that $b'(t)>0$ on $t \in [0,S]$ is sufficient. Assumption (H4) is needed to define $\psi(t)$ by (3.10).

Now we are in a position to show the solution existence and uniqueness of the inverse problem (IP).

THEOREM 1. *Under assumptions* (H1)–(H4)*, the inverse problem* (IP) *has a unique classic solution: $B(t,T),\ A(t,T) \in C^1(\Omega)$, $\phi(t),\ \psi(t) \in C[0,T^*]$, where $\Omega = \{(t,T):\ 0 \le t \le T \le T^*\}$.*

Proof. It is not difficult to check that (3.9) satisfies the conditions of Theorem 3.6 of [11] when assumptions (H1)–(H2) hold. So it has a unique solution $u_2(t)$ in $C^2[0,T^*]$ which also satisfies (3.8). Letting $t=0$ in (3.8) and (3.9), we have $u_2'(0)=1$ and $u_2(0)=0$, where the assumption (H3) was used. Then by assumption (H4), $\psi(t)$ defined by (3.10) is in $C[0,T^*]$, and thus $u_2(t)$ is the unique solution of the initial value problem (3.3). It is easy to check that $u_1(t)$ defined by (3.7) is the unique solution of the initial value problem (3.2) in $C^2[0,T^*]$. Therefore, function $B(t,T)$ defined by (3.6) is the unique solution of (2.4) with the final condition (2.6) and the initial condition $B(0,T)=b(T)$. Notice from (3.6) that $B_{TT}(t,T)$ is a continuous function on Ω. By Theorem 3.1 of [11], the integral equation (3.14) has a unique solution $\phi(T)$ in $C[0,T^*]$, and thus $A(t,T) \in C^1(\Omega)$ determined by (3.11) is the unique solution of (2.3) with the final condition (2.5) and the initial condition $A(0,T)=a(T)$. ∎

We may relax the conditions about the smoothness of functions $a(t)$, $b(t)$, and $d(t)$. Since such conditions are quite complicated, we refer the interested reader to Chapter 3 of [11] for a general consideration. Define

$$K\phi(t) = \int_0^t K(t,s)\phi(s)ds, \qquad L\phi(t) = \int_0^t g'(s)\phi(s)ds\,.$$

Then we have from (3.9)

$$u_2(t) = (I+K)^{-1}f(t) = \sum_{n=0}^{\infty}(-K)^n f(t).$$

Since

$$(K\phi)'(t) = f'(t)L\phi(t),$$

we have after some calculation

$$u_2'(t) = f'(t)(1-Lf(t)) + f'(t)L\sum_{m=0}^{\infty} K^{2m+1}(I-K)f(t).$$

Notice that integral operators K and L have positive kernels and $f(t)$ and $f'(t)$ are positive. We propose the following sufficient conditions for $u_2'(t) > 0$ on $t \in [0, T^*]$:

$$1 - Lf(T^*) > 0, \qquad (I-K)f(T^*) > 0.$$

The first inequality implies the second one and can be rewritten as

$$\int_0^{T^*} \frac{d(t)(g(T^*)-g(t))}{g'(t)}dt < 1. \tag{3.15}$$

Hence assumption (H4) can be replaced by this stronger condition.

The extended CIR model can also be calibrated by assuming that $\psi(t)$ and $\sigma(t)$ are constant or that $\phi(t)$ and $\psi(t)/\sigma^2(t)$ are constant. In the first case, we need to determine only $\psi(t)$ by (3.13) or (3.14) (see [4]). For the latter case, $\sigma(t)$ and $\psi(t)$ can be determined analytically [17].

4. The multifactor CIR models. In this section, we extend the results in the previous section to the multifactor CIR models. Empirical studies have shown that multifactor models have more flexibility to capture more variability of interest rates (see [3], [6], [13], and references therein). Without loss of generality, we consider only the two-factor CIR model, which has the models in [3] and [13] as two specific cases mathematically. The model specifies that the short rate process $r(t)$ is the sum of two square root processes $x_1(t)$ and $x_2(t)$; i.e.,

$$\begin{aligned} r(t) &= x_1(t) + x_2(t), \\ x_i(t) &= (\phi_i(t) - \psi_i(t)x_i(t))dt + \sigma_i(t)\sqrt{x_i(t)}dW_i(t), \quad i = 1, 2, \end{aligned}$$

where $(W_1(t), W_2(t))$ is a two-dimensional standard Brownian motion under the risk-neural measure. Let $P(x_1, x_2, t, T)$ be the price of the zero-coupon bond which pays \$1 at time T. Then it is the solution of the following fundamental partial differential equation (see [1], [13], and [16]):

$$P_t + \sum_{i=1}^{2}\left(\frac{1}{2}\sigma_i(t)^2 x_i P_{ii} + (\phi_i(t) - \psi_i(t)x_i)P_i - x_i P\right) = 0,$$

where P_i and P_{ii} denote the first and second order derivatives with respect to x_i. Write

$$P(x_1, x_2, t, T) = A(t, T) \exp(-B(t, T)x_1 - C(t, T)x_2). \tag{4.1}$$

Then we have the following ordinary differential equations for $A(t, T)$, $B(t, T)$, and $C(t, T)$:

$$A_t(t, T) - (\phi_1(t)B(t, T) + \phi_2(t)C(t, T))A(t, T) = 0, \quad 0 \le t \le T, \tag{4.2}$$

$$B_t(t, T) - \psi_1(t)B(t, T) - \frac{1}{2}\sigma_1(t)^2 B^2(t, T) + 1 = 0, \quad 0 \le t \le T, \tag{4.3}$$

$$C_t(t, T) - \psi_2(t)C(t, T) - \frac{1}{2}\sigma_2(t)^2 C^2(t, T) + 1 = 0, \quad 0 \le t \le T. \tag{4.4}$$

Similarly to the one-factor model, we have the following final conditions:

$$A(T, T) = 1, \tag{4.5}$$

$$B(T, T) = C(T, T) = 0. \tag{4.6}$$

When all parametric functions are constant, the above final value problem can be solved analytically (see Example 6.2).

Assume that $\sigma_1(t)$, $\sigma_2(t)$, $A(0, T)$, $B(0, T)$, and $C(0, T)$ are determined by using the current market data. Then, as in section 3, we can determine $\psi_1(t)$, $B(t, T)$, $\psi_2(t)$, and $C(t, T)$ from (4.3), (4.4), and (4.6). Solving (4.2) for $A(t, T)$, we get

$$A(t, T) = \exp\left(-\int_t^T (\phi_1(s)B(s, T) + \phi_2(s)C(s, T))ds\right).$$

Letting $t = 0$ and taking the natural logarithm, we get

$$\int_0^T (\phi_1(s)B(s, T) + \phi_2(s)C(s, T))\, ds = -\ln(A(0, T)).$$

Since only one function can be determined by this integral equation, we need one of the following assumptions:

(H5) $\phi_1(t)$ or $\phi_2(t)$ is constant.

(H6) $\phi_1(t)$ and $\phi_2(t)$ are linear dependent; i.e., for some constant λ, $\phi_1(t) = \lambda\phi_2(t)$.

(H7) The mean reversion value, $\phi_i(t)/\psi_i(t)$, is constant for $i = 1$ or 2.

Then we have the integral equations similar to (3.13) and (3.14) for $\phi_1(t)$ or $\phi_2(t)$.

Based on the above discussion, we can prove the following result similar to Theorem 1.

THEOREM 2. *The parametric functions $\phi_i(t)$ and $\psi_i(t)$ $(i = 1, 2)$ of the two-factor CIR model can be uniquely determined in $C([0, T^*])$ from* (4.2)–(4.6) *together with known $\sigma_1(t)$, $\sigma_2(t)$, $A(0, T)$, $B(0, T)$, and $C(0, T)$ under the assumptions similar to* (H1)–(H4) *for $A(0, T)$, $B(0, T)$, and $C(0, T)$ and assumption* (H5), (H6), *or* (H7).

Here we have assumed that $dW_1(t)$ and $dW_2(t)$ are uncorrelated. If their correlation is ρdt for some nonzero constant ρ, then the zero-coupon bond price $P(x_1, x_2, t, T)$ can not be expressed in the form of (4.1). Indeed, it satisfies the following partial differential equation:

$$P_t + \sum_{i=1}^{2}\left(\frac{1}{2}\sigma_i(t)^2 x_i P_{ii} + (\phi_i(t) - \psi_i(t)x_i)P_i - x_i P\right) + \rho\sqrt{x_1 x_2}P_{12} = 0,$$

where P_{12} are the mixed second derivatives. In this case, the inverse problem is to determine $P(x_1, x_2, t, T)$ and the parametric functions from the above parabolic partial differential equation together with the current term structure and the other market information. Usually, it is difficult to show the solution existence of such inverse problems and to solve them numerically. Interested readers are referred to [9] for the theory of inverse problems for partial differential equations, and [2] and [10] for the inverse problems for stock option problems.

5. Numerical solutions of the inverse problem (IP). In this section we shall consider approximate solutions to the inverse problem (IP). To this end, we summarize the proof of Theorem 1 as the following algorithm:

Step 1. Solve (3.9) for $u_2(t)$.
Step 2. Compute $u_2'(t)$ by (3.8).
Step 3. Compute $u_1(t)$ and $u_1'(t)$ by (3.7) and (3.5), respectively.
Step 4. Compute $B(t,T)$, $B_T(t,T)$, and $B_{TT}(t,T)$ by using (3.6).
Step 5. Compute $\psi(t)$ by (3.10).
Step 6. Solve (3.13) or (3.14) for $\phi(t)$.
Step 7. Compute $A(t,T)$ by (3.11).
Step 8. Compute the bond price $P(r,t,T)$ by (2.2).

The crucial steps of the above algorithm are Steps 1 and 6, in which we need to solve Volterra equations. There is a large literature in numerical methods and their error analysis for Volterra equations. Interested readers are referred to [11] and references cited therein. In the following, we outline the block-by-block method in section 7.6 of [11] for (3.7) and (3.14) and propose a numerical implementation of our algorithm with the accuracy of order 4.

For an even positive integer M, let the step size $h = T^*/M$, $t_m = mh$ for $m = 0, 1, \ldots, M$, and $t_{m+1/2} = t_m + h/2$ for $m = 0, 1, \ldots, M-1$. Denote by $u_{2,m}$ the approximation of $u_2(t_m)$. Then the block-by-block method for (3.9) reads as follows: for $m = 0, 1, \ldots, M/2 - 1$, compute $u_{2,2m+1}$ and $u_{2,2m+2}$ by

$$\begin{cases} a_m u_{2,2m+1} + b_m u_{2,2m+2} = p_m, \\ c_m u_{2,2m+1} + d_m u_{2,2m+2} = q_m, \end{cases} \tag{5.1}$$

where

$$a_m = 1 + \frac{h}{2}K(t_{2m+1}, t_{2m+1/2}) + \frac{h}{6}K(t_{2m+1}, t_{2m+1}), \qquad b_m = -\frac{h}{12}K(t_{2m+1}, t_{2m+1/2}),$$

$$c_m = \frac{4h}{3}K(t_{2m+2}, t_{2m+1}), \qquad d_m = 1 + \frac{h}{2}K(t_{2m+2}, t_{2m+2}),$$

$$p_m = f(t_{2m+1}) + \frac{h}{6}K(t_{2m+1}, t_{2m})u_{2,2m} - \frac{h}{4}K(t_{2m+1}, t_0)u_{2,0} - \frac{h}{3}\sum_{i=0}^{2m} w_{m,i}K(T_{2m+1}, t_i),$$

$$q_m = f(t_{2m+2}) - \frac{h}{3}\sum_{i=0}^{2m} w_{m,i}K(T_{2m+2}, t_i),$$

where $\{w_{m,0}, w_{m,1}, \ldots, w_{m,m-1}, w_{m,m}\} = \{1, 4, 2, \ldots, 2, 4, 1\}$. The above block-by-block method is derived by using Simpson's rule as the numerical integration formula. System (5.1) has a unique solution when h is small enough. Furthermore, for sufficiently smooth $K(t,s)$ and $f(t)$, we have the following error estimate:

$$\max_{1 \le m \le M} |u_2(t_m) - u_{2,m}| \le Ch^4$$

for some constant C independent of h (see Chapter 7 of [11]).

We can also apply the above block-by-block method to (3.14) for $\phi(t)$. Since $b''(T)$ ($g''(T)$) has to be computed in order to compute $B_{TT}(t,T)$ by (3.6), the accuracy of computation would be reduced when $b(t)$ is obtained by using the current market data through curve fitting. Instead, we may solve (3.13) by block-by-block methods, for example, the method in section 9.4 of [11].

Now consider the approximations of $u_1(t_m)$, $u_1'(t_m)$, and $u_2'(t_m)$. When m is even, the integral in (3.8) can be approximated by Simpson's rule at the mesh points $t_0, t_1, \ldots, t_m$, and thus $u_2'(t_m)$ can be computed with the error of order 4. When m is odd, $u_2'(t_m)$ will be computed by a fourth order interpolation formula on even mesh points. The approximation of $u_1'(t_m)$ is then obtained by (3.5) with the same accuracy. By (3.7) and (3.8), we have the following formula to compute $u_1(t)$ in terms of $u_2'(t)$:

$$u_1(t) = g(t)u_2(t) - \frac{g'(t)u_2'(t)}{d(t)}, \qquad 0 \le t \le T^*.$$

Then we can compute the approximations of $\psi(t_m)$ and $B(t_i,t_m)$ by (3.10) and by (3.6), respectively. Differentiating (3.6) with respect to T, we get the following formulas for computing $B_T(t_i,t_m)$ and $B_{TT}(t_i,t_m)$:

$$B_T(t,T) = -\frac{g'(T)(u_1'(t)u_2(t) - u_1(t)u_2'(t))}{d(t)(g(T)u_2(t) - u_1(t))^2},$$

$$B_{TT}(t,T) = -\frac{(g''(T)(g(T)u_2(t) - u_1(t)) - 2g'(T)^2u_2(t))(u_1'(t)u_2(t) - u_1(t)u_2'(t))}{d(t)(g(T)u_2(t) - u_1(t))^3}.$$

Finally, we discuss how to compute $A(t_i,t_m)$. When $m-i$ is even, the integral in (3.11) can be approximated by Simpson's rule on mesh points. When i is even and m is odd, we rewrite (3.11) as

$$A(t,T) = A(0,T)\exp\left(\int_0^t \phi(s)B(s,T)ds\right), \qquad 0 \le t \le T,\ 0 \le T \le T^*.$$

The integral in the above formula can also be approximated by Simpson's rule on mesh points. Therefore, $A(t_i,T_m)$ can be computed with an error of order 4 in these two cases. For the other cases, $A(t_i,T_m)$ will be computed by a fourth order interpolation formula.

To sum up, our algorithm in the beginning of this section can be implemented numerically with fourth order accuracy.

6. Numerical examples. In this section, we will present numerical results to examine the accuracy of the numerical implementation of our algorithm and to compare the extended CIR model (ECIR) with the Vasicek and the extended Vasicek models. Our C++ codes were run on a Dell 2.4 GHz PC with Red Hat 7.3. Although the numerical experiments are also done for computation of $\phi(t)$ using (3.13), we present the numerical results when the approximations of $\phi(t)$ are computed using (3.14). We observed that both approaches give very accurate approximations for the model problems considered below.

Example 6.1. Consider the CIR model; i.e., $\sigma(t)$, $\psi(t)$, and $\phi(t)$ are constant functions in (2.1). The constants are denoted by σ_0, ψ_0, and ϕ_0, respectively. Then we have (see [4])

$$A(t,T) = \left(\frac{c_1e^{c_2s}}{c_2(e^{c_1s}-1)+c_1}\right)^{c_3}, \qquad B(t,T) = \frac{e^{c_1s}-1}{c_2(e^{c_1s}-1)+c_1}, \qquad s = T^* - t,$$

TABLE 6.1
Maximum errors.

M	$\psi(t)$	$\phi(t)$	$A(t,T)$	$B(t,T)$	CPU
20	$1.39E-4$	$1.13E-3$	$1.90E-4$	$1.04E-3$	0.00
40	$1.03E-5$	$8.01E-5$	$1.33E-5$	$7.52E-5$	0.01
80	$7.08E-7$	$5.33E-6$	$8.39E-7$	$5.11E-6$	0.02
160	$4.64E-8$	$3.44E-7$	$5.19E-8$	$3.34E-7$	0.10
320	$2.97E-9$	$2.19E-8$	$3.22E-9$	$2.13E-8$	0.60
640	$1.88E-10$	$1.38E-9$	$2.00E-10$	$1.35E-9$	4.15

$$c_1 = (\psi_0^2 + 2\sigma_0^2)^{\frac{1}{2}}, \quad c_2 = \frac{(\psi_0 + c_1)}{2}, \quad c_3 = \frac{2\phi_0}{\sigma_0^2}.$$

Let $a(T) = A(0,T)$, $b(T) = B(0,T)$, and $\sigma(t) = \sigma_0$ in the inverse problem (IP). It is apparent that $\psi(t) \equiv \psi_0$, $\phi(t) \equiv \phi_0$, $A(t,T)$, and $B(t,T)$ are the exact solution of the inverse problem (IP).

Take $\sigma = 0.2$, the long-term interest rate $\theta = 10\%$, the adjustment speed $\kappa = 0.4$, and $T^* = 10$. Then $\psi = \kappa\theta = 0.04$ and $\phi = \kappa = 0.4$. In Table 6.1, we display the relative root mean squared errors and the CPU times in seconds. The CPU time includes computing the approximations of $\psi(t_i)$, $\phi(t_i)$, $A(t_j,t_i)$, $B(t_j,t_i)$ for all $i = 0,1,\ldots,M$, $j = 0,1,\ldots,i$, and errors. The results show that the proposed numerical algorithm is very accurate and fast. One can check that the error ratios in each column of Table 6.1 are approximately 16, which agrees with the theoretical prediction of the convergence order 4.

Example 6.2. In this example, we shall calibrate the ECIR model by the Vasicek model. The interest rate process $r(t)$ of the Vasicek model follows

$$dr(t) = \kappa(\theta - r(t))dt + \sigma dW(t).$$

The bond price is given by (2.2) with

$$B(t,T) = \frac{1 - e^{-\kappa(T-t)}}{\kappa},$$
$$A(t,T) = \exp\left(\frac{(2\kappa^2\theta - \sigma^2)(B(t,T) - T + t)}{2\kappa^2} - \frac{\sigma^2}{4\kappa}B^2(t,T)\right).$$

Take $\sigma = 0.06$, the adjustment speed $\kappa = 0.4$, and the long-term interest rate $\theta = 10\%$. The face value is \$100. Assume that the initial short rate $r(0)$ is 9%. Let $a(t) = A(0,T)$ and $b(t) = B(0,T)$ in the inverse problem (IP). For the ECIR model, we set $\sigma(t) = 0.06/\sqrt{0.09} = 0.2$. As indicated in [7], this ensures that the initial short rate volatility equals that in the Vasicek model.

Numerical results show that $u_2'(t)$ is a strictly increasing function, and its graph looks like an exponential function, while the stronger condition (3.15) holds only for $T^* < 22.5$. We display the graphs of $\psi(t)$ and $\phi(t)/\psi(t)$ in Figure 6.1. The figure shows that $\psi(t)$ is a decreasing function having a horizontal asymptote $\psi = 0.2828$ and that $\phi(t)/\psi(t)$ is an increasing function having a horizontal asymptote $\phi/\psi = 0.1071$. Thus the long-term adjustment speed for the ECIR model is approximately 0.2828, which is smaller than that (0.4) for the Vasicek model, and the long-term interest rate for the ECIR model is approximately 10.71%, which is greater than that (10%) for the Vasicek model. In Figure 6.2, we display the contour maps of the differences of the

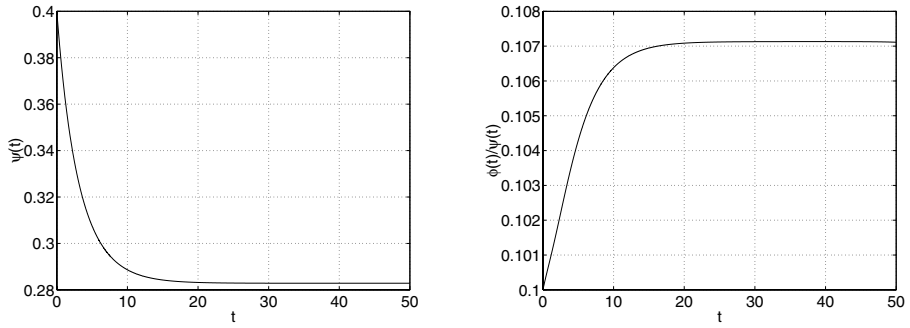


FIG. 6.1. *The graphs of $\psi(t)$ and $\phi(t)/\psi(t)$.*

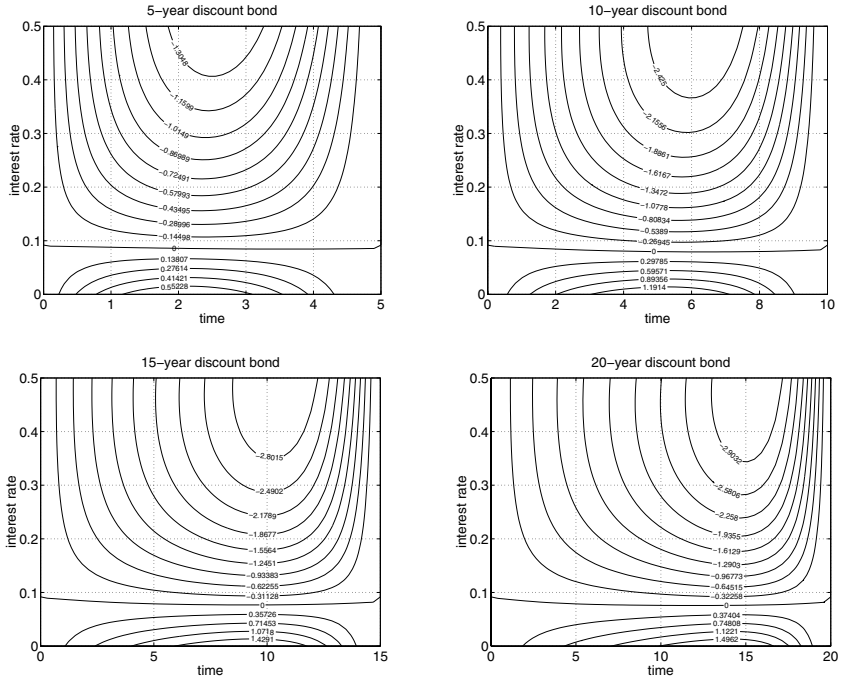


FIG. 6.2. *Contour plots of price differences.*

bond prices under the extended Vasicek model and the ECIR model. We observed that the whole computing domain is divided into two parts. In the part corresponding to small interest rates, the bond price under the extended Vasicek model is greater than that under the ECIR model, and the reverse is true in the other part corresponding to large interest rates. This is because high interest rates have a greater chance of occurring under the ECIR model. The graphs of bond prices are displayed in Figure 6.3. We observe that the bond price is an increasing function of time and a decreasing function of interest rates, which is as expected both in theory and in practice.

Example 6.3. As in [7], in this example we assume that the initial term structure is determined by the two-factor CIR model with constant parameters. Then the discount bond price is given by

$$P(x_1, x_2, t, T) = A(t, T) \exp(-x_1 B_1(t, T) - x_2 B_2(t, T)),$$

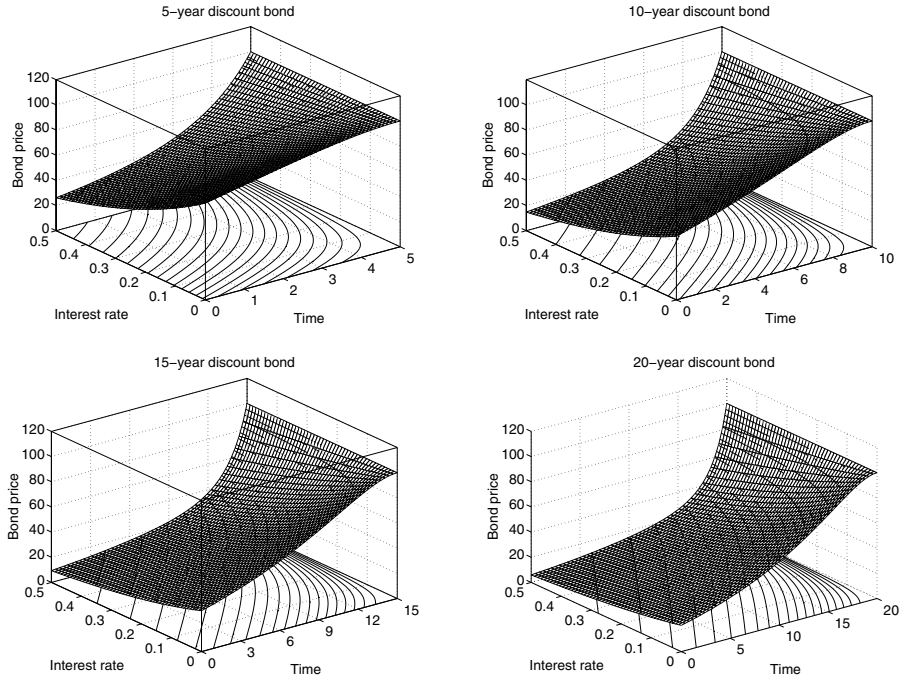


FIG. 6.3. *Bond prices.*

where $A(t,T) = A_1(t,T)A_2(t,T)$, and $A_i(t,T)$ and $B_i(t,T)$ are given by the corresponding formulas, as in Example 6.1. The extended Vasicek model is fitted to the two-factor CIR model as follows (see [7]):

$$\sigma = \sqrt{\sigma_1^2 x_1(0) + \sigma_2^2 x_2(0)},$$
$$\sigma B(0,T) = \sqrt{\sigma_1^2 x_1(0) B_1(0,T)^2 + \sigma_2^2 x_2(0) B_2(0,T)^2},$$
$$P(r(0),0,T) = A(0,T)e^{-B(0,T)r(0)}.$$

Similarly, we fit the ECIR model as follows:

$$\sigma\sqrt{r(0)} = \sqrt{\sigma_1^2 x_1(0) + \sigma_2^2 x_2(0)},$$
$$\sigma\sqrt{r(0)}B(0,T) = \sqrt{\sigma_1^2 x_1(0) B_1(0,T)^2 + \sigma_2^2 x_2(0) B_2(0,T)^2},$$
$$P(r(0),0,T) = A(0,T)e^{-B(0,T)r(0)},$$

where $r(0) = x_1(0) + x_2(0)$. The parameter values are as follows:

$$\sigma_1 = 0.1, \quad \psi_1 = 0.2, \quad \phi_1 = 0.01, \quad x_1(0) = 0.06,$$
$$\sigma_2 = 0.1, \quad \psi_2 = 0.3, \quad \phi_2 = 0.009, \quad x_2(0) = 0.02.$$

The bond face value is \$100.

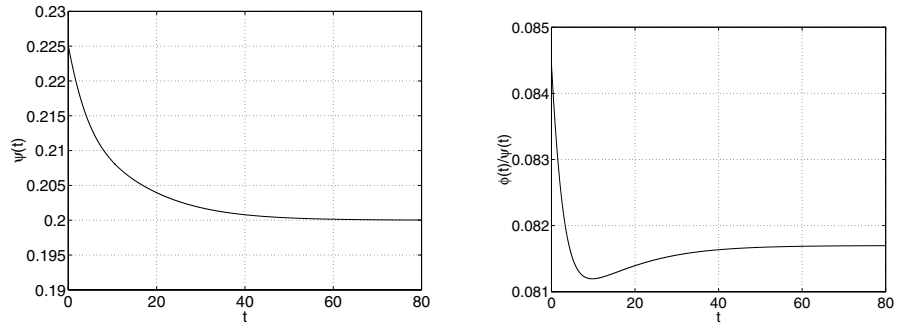


FIG. 6.4. *The graphs of $\psi(t)$ and $\phi(t)/\psi(t)$.*

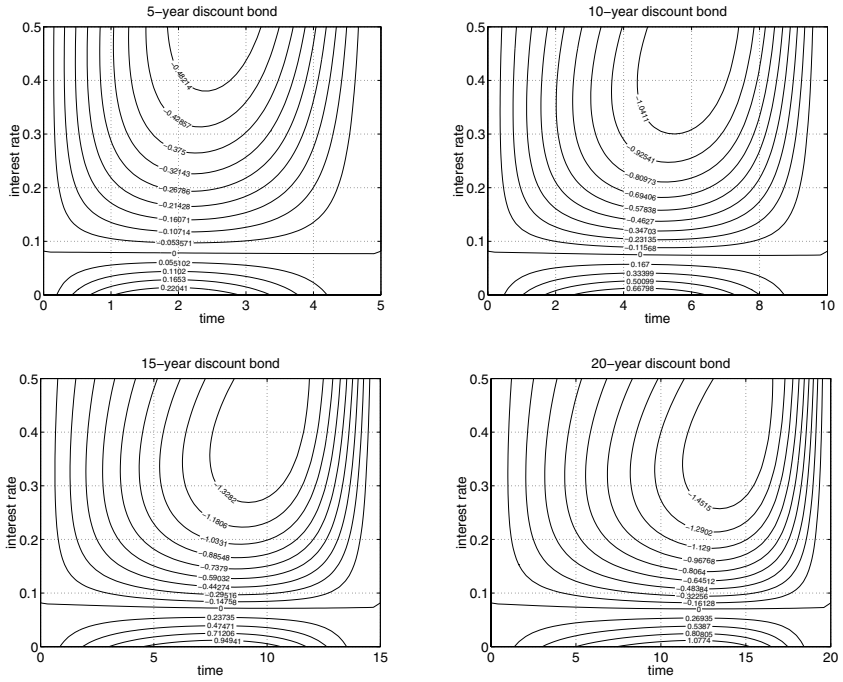


FIG. 6.5. *Contour plots of price differences.*

Numerical results show that $u_2'(t)$ is a strictly increasing function, and its graph looks like an exponential function, while the stronger condition (3.15) holds only for $T^* < 53.29$. The graphs of $\psi(t)$ and $\phi(t)/\psi(t)$ are displayed in Figure 6.4, which shows that the long-term adjustment speed of ECIR model is approximately 0.2, which is the same as the smaller one of the TCIR model. The figure also shows that $\psi(t)$ is a decreasing function while $\phi(t)/\psi(t)$ is not monotone. We observed that the long-term interest rate of the ECIR model is approximately 8.17%, which is smaller than that (9%) of the TCIR model. We display the contour maps of the differences of the bond prices under the extended Vasicek model and the ECIR model and graphs of bond prices in Figure 6.5 and 6.6, respectively. We have the same observations from these pictures as in Example 1.

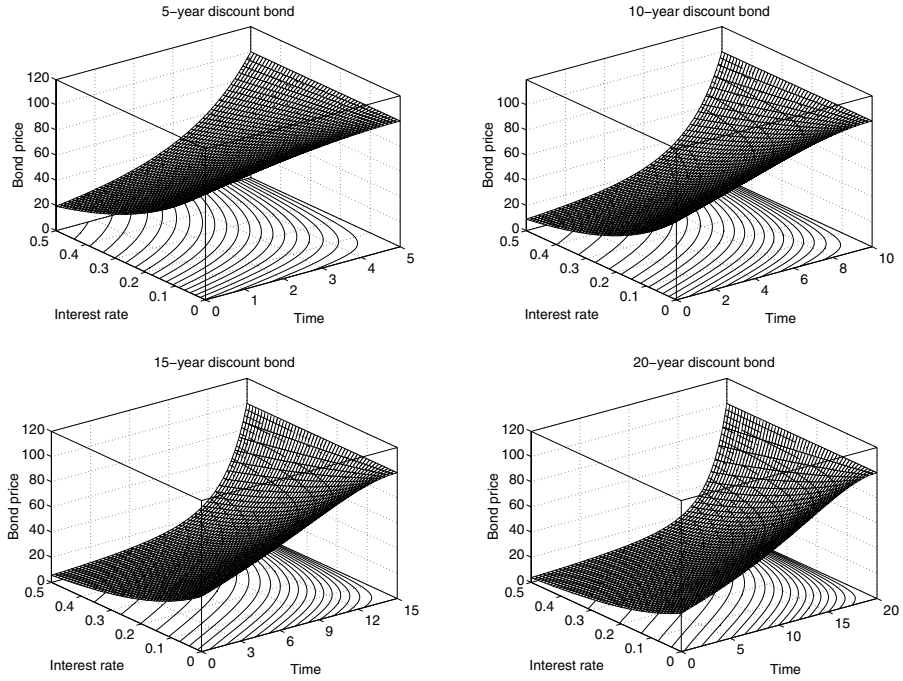


FIG. 6.6. *Bond prices.*

7. Conclusions. In this paper we studied the inverse problem for calibrating the extended CIR model. The solution existence and uniqueness of the problem are established under appropriate assumptions. In other words, the time-dependent parameters of the extended CIR model can be uniquely determined by the current term structure of interest rates and either the current volatilities of all spot interest rates or the current volatilities of all forward interest rates. A fourth order algorithm is proposed to compute the approximations of the time-dependent parameters and the discount bond prices. Numerical results have shown that our algorithm is very accurate and rapid. Comparisons with the Vasicek model and the extended Vasicek model are also presented, and we find that the extended CIR model gives higher/lower discount bond prices for high/lower interest rates. As expected, the bond price is an increasing function of time and a decreasing function of interest rate. Work is ongoing to study numerical methods for pricing other interest rate derivatives, for example, bond options and caps, under the extended CIR model.

Acknowledgment. The author is grateful to the editor William Morokoff and the referees for their insightful comments and suggestions, which have significantly improved the paper.

REFERENCES

[1] T. BJÖRK, *Interest Rate Theory*, Lecture Notes in Math. 1656, Springer, New York, 1996, pp. 53–122.

[2] I. BOUCHOUEV AND V. ISAKOV, *The inverse problem of option pricing*, Inverse Problems, 13 (1997), pp. L11–L17.

[3] R. R. Chen and L. Scott, *Pricing interest rate options in a two-factor Cox–Ingersoll–Ross model of the term structure*, Rev. Financial Studies, 5 (1992), pp. 613–636.
[4] J. C. Cox, J. E. Ingersoll, and S. A. Ross, *A theory of the term structure of interest rates*, Econometrica, 53 (1985), pp. 385–407.
[5] M. Chesney, R. Elliott, and R. Gibson, *Analytical solutions for the pricing of American bond and yield options*, Math. Finance, 3 (1993), pp. 277–294.
[6] Q. Dai and K. J. Singleton, *Specification analysis of affine term structure models*, J. Finance, 55 (2000), pp. 1943–1978.
[7] J. Hull and A. White, *Pricing interest-rate-derivative securities*, Rev. Financial Studies, 3 (1990), pp. 573–592.
[8] J. Hull and A. White, *One-factor interest-rate models and the valuation of interest-rate derivative securities*, J. Fin. Quan. Anal., 28 (1993), pp. 235–254.
[9] V. Isakov, *Inverse Problems for Partial Differential Equations*, Springer-Verlag, New York, 1998.
[10] L. Jiang and Y. Tao, *Identifying the volatility of underlying assets from option prices*, Inverse Problems, 17 (2001), pp. 135–155.
[11] P. Linz, *Analytical and Numerical Methods for Volterra Equations*, SIAM Stud. Appl. Math. 7, SIAM, Philadelphia, 1985.
[12] F. A. Longstaff and E. S. Schwartz, *Interest rate volatility and the term structure: A two-factor general equilibrium model*, J. Finance, 47 (1992), pp. 1259–1282.
[13] F. A. Longstaff and E. S. Schwartz, *Implementation of the Longstaff–Schwartz interest rate model*, J. Fixed Income, 3 (1993), pp. 7–14.
[14] Y. Maghsoodi, *Solution of the extended CIR term structure and bond option valuation*, Math. Finance, 6 (1996), pp. 89–109.
[15] R. C. Merton, *Theory of rational option pricing*, Bell J. Econ. Manage. Sci., 4 (1973), pp. 141–183.
[16] R. Rebonato, *Interest-Rate Option Models: Understanding, Analyzing and Using Models for Exotic Interest-Rate Options*, John Wiley & Sons, Chichester, UK, and Toronto, 1998.
[17] L. G. G. Rogers, *Which model for term-structure of interest rates should one use?*, in Mathematical Finance, IMA Vol. Math. Appl. 65, M. H. A. Davis, D. Duffie, W. Fleming, and S. E. Shreve, eds., Springer, New York, 1995, pp. 93–116.
[18] O. A. Vasicek, *An equilibrium characterization of the term structure*, J. Financial Economics, 5 (1977), pp. 177–188.

SIAM J. APPL. MATH.
Vol. 66, No. 3, pp. 736–757

THE CONVERGENCE OF REGULARIZED MINIMIZERS FOR CAVITATION PROBLEMS IN NONLINEAR ELASTICITY*

JEYABAL SIVALOGANATHAN†, SCOTT J. SPECTOR‡, AND VIVEKA TILAKRAJ§

Abstract. Consider a nonlinearly elastic body which occupies the region $\Omega \subset \mathbb{R}^m$ $(m = 2, 3)$ in its reference state and which is held in tension under prescribed boundary displacements on $\partial\Omega$. Let $\mathbf{x}_0 \in \Omega$ be any fixed point in the body. It is known from variational arguments that, for sufficiently large boundary displacements, there may exist discontinuous weak solutions of the equilibrium equations corresponding to a hole forming at $\mathbf{x}_0$ in the deformed body (this is the phenomenon of cavitation). For each $\epsilon > 0$, define the regularized domains $\Omega_\epsilon = \Omega \backslash \overline{B_\epsilon(\mathbf{x}_0)}$ which contain a preexisting hole of radius $\epsilon > 0$ centered on $\mathbf{x}_0$. Now consider the corresponding mixed displacement/traction problem on Ω_ϵ in which the boundary $\partial\Omega$ is subject to the same boundary displacements and the deformed cavity surface (i.e., the image of ∂B_ϵ) is required to be stress-free. It follows from variational arguments that there exists a weak solution $\mathbf{u}_\epsilon$ of this problem for each $\epsilon > 0$. In this paper we prove convergence of these regularized minimizers $\mathbf{u}_\epsilon$ in the limit as $\epsilon \to 0$. In particular, we show that if $\epsilon_n \to 0$, then, passing to a subsequence, $\mathbf{u}_{\epsilon_n} \to \mathbf{u}$, where $\mathbf{u}$ is a minimizer for the original pure displacement problem on Ω.

Finally, we study the effect on cavitation of regularizing the variational problem by introducing a surface energy term which penalizes the formation and growth of cavities.

Key words. cavitation, convergence, elastic, equilibrium, regular minimizers, singular minimizers, surface energy, weak limit

AMS subject classifications. Primary, 74B20; Secondary, 49K20, 74G65

DOI. 10.1137/040618965

1. Introduction.

Let $\Omega \subset \mathbb{R}^m$ $(m = 2, 3)$ denote the region occupied by a nonlinearly elastic body in its reference configuration. A deformation of the body corresponds to a map $\mathbf{u} : \Omega \to \mathbb{R}^m$ that lies in the Sobolev space $W^{1,1}(\Omega; \mathbb{R}^m)$, is one-to-one a.e., and satisfies

$$\det \nabla \mathbf{u}(\mathbf{x}) > 0 \quad \text{for a.e. } \mathbf{x} \in \Omega. \tag{1.1}$$

In hyperelasticity the total energy stored under such a deformation is given by

$$E(\mathbf{u}) = \int_\Omega W(\mathbf{x}, \nabla \mathbf{u}(\mathbf{x}))\, d\mathbf{x}, \tag{1.2}$$

where $W : \overline{\Omega} \times \mathrm{M}^{m\times m}_+ \to [0, \infty)$ is the stored energy function of the material and $\mathrm{M}^{m\times m}_+$ denotes the set of real $m \times m$ matrices with positive determinant. We consider

*Received by the editors November 15, 2004; accepted for publication (in revised form) September 8, 2005; published electronically February 3, 2006. This work was supported in part by the National Science Foundation under grants 0072414 and 0405646. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

http://www.siam.org/journals/siap/66-3/61896.html

†Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK (js@maths.bath.ac.uk). This author's work was partially supported by a Leverhulme Fellowship.

‡Department of Mathematics, Southern Illinois University, Carbondale, IL 62901 (sspector@math.siu.edu).

§Abingdon Technology Centre, Schlumberger Geoquest, Abingdon OX14 1DZ, UK (vtilakraj@abingdon.oilfield.slb.com). This author's work was supported by a studentship from the University of Bath.

the displacement problem in which we require

$$\mathbf{u}(\mathbf{x}) = \mathbf{A}\mathbf{x} \quad \text{for } \mathbf{x} \in \partial\Omega, \tag{1.3}$$

where $\mathbf{A} \in \mathrm{M}^{m\times m}_{+}$ is fixed.

The radial problem. Following Ball's seminal paper [1], much work has been carried out in the variational setting on the existence of discontinuous radial minimizers for (1.2) corresponding to cavitation (see, e.g., the review article [12] and the references therein). In this approach $\mathbf{A} = \lambda\mathbf{I}$, $\lambda > 0$, $\Omega = B$ is the unit ball in $\mathbb{R}^m$, and deformations are of the form

$$\mathbf{u}(\mathbf{x}) = r(R)\frac{\mathbf{x}}{|\mathbf{x}|},$$

where $r : [0,1] \to [0,\infty)$ and $R = |\mathbf{x}|$. It is known that there are large classes of physically reasonable W for which there exists a critical value of the boundary displacement λ_{crit} such that

(i) for $\lambda \le \lambda_{\text{crit}}$, the unique radial energy minimizer is the homogeneous deformation corresponding to $r(R) \equiv \lambda R$;

(ii) for $\lambda > \lambda_{\text{crit}}$, the unique radial energy minimizer satisfies $r(0) > 0$, corresponding to a deformation that produces a hole at the center of the initially perfect ball (this is the phenomenon of cavitation).

Example 1.1. A typical class of stored energy functions for which the above results hold is given by

$$W(\mathbf{x},\mathbf{F}) = c|\mathbf{F}|^p + \Gamma(\det\mathbf{F}) \quad \forall\,\mathbf{F} \in \mathrm{M}^{m\times m}_{+} \text{ and } \mathbf{x} \in \overline{\Omega}, \tag{1.4}$$

where $c > 0$, $1 \le p < m$, and Γ is a convex function that grows superlinearly and satisfies $\Gamma(d) \to +\infty$ as $d \to 0^+$.

The discontinuous minimizers with r as in (ii) are weak solutions of the corresponding equilibrium equations (see, e.g., [1]). Earlier approaches to cavitation had not been variational and tended to model cavitation as the growth of small preexisting voids in the material (see, e.g., [9] in the context of nonlinear elasticity and [10] in the context of elastoplasticity).

There are various ways of reconciling the two approaches. In particular, results of [18] in the variational setting (see also the example in [11]) show that radial equilibrium solutions

$$\mathbf{u}_\epsilon(\mathbf{x}) = r_\epsilon(R)\frac{\mathbf{x}}{|\mathbf{x}|}$$

for the mixed displacement/zero-traction problem in which B is replaced by $B_\epsilon = \{\mathbf{x} : \epsilon < |\mathbf{x}| < 1\}$ (i.e., a ball with a preexisting hole of radius ϵ in the reference configuration) converge to the radial minimizer for B studied in [1] as $\epsilon \to 0$. In particular, $\sup_{R\in[\epsilon,1]} |r_\epsilon(R) - r(R)| \to 0$ as $\epsilon \to 0$, where $r(R)$ is the minimizer given in (i) and (ii) above (see Figure 1 and also the discussion in [1] for the case of incompressible elasticity).

The nonsymmetric case. The first results extending Ball's original variational approach to nonsymmetric situations while allowing cavitation (i.e., discontinuities) to occur at any point in the body are contained in Müller and Spector [16]. A key element in their approach is an analytical restriction on the class of admissible

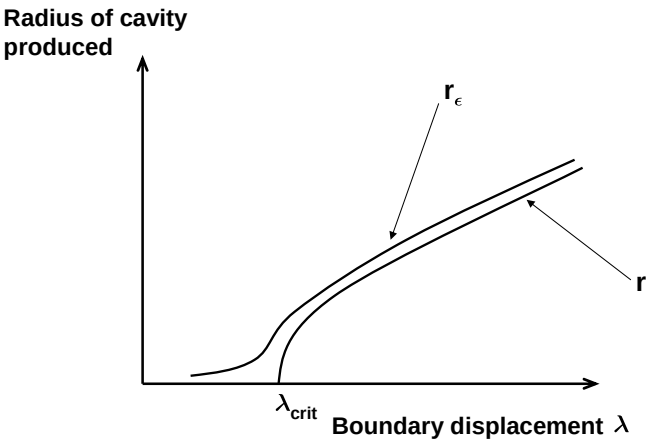


FIG. 1. *Bifurcation diagram for solid and punctured balls.*

deformations $\mathbf{u}$ called condition (INV). Condition (INV) is the requirement that $\mathbf{u}$ be monotone[1] in the sense of Lebesgue and that, roughly speaking, holes created in one part of the body cannot be filled by material from elsewhere (see section 2.3 for the precise definition). Subsequent work in [19], in the variational setting, proposed an alternative model in which cavitation could occur only at a, possibly large, number of infinitesimal flaws in the material. This was modelled mathematically by using admissible deformations whose possible point discontinuities are constrained to be at the specified flaw points: let $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n \in \Omega$ be the flaw points and minimize the total energy (1.2) in the class of deformations that satisfy (1.1), (1.3), condition (INV), and

$$\mathrm{Det}\nabla\mathbf{u} = (\det \nabla\mathbf{u})\mathcal{L}^m + \sum_{i=0}^{n} \alpha_i \delta_{\mathbf{x}_i},$$

where $\mathrm{Det}\nabla\mathbf{u}$ denotes the distributional Jacobian of $\mathbf{u}$, $\mathcal{L}^m$ denotes m-dimensional Lebesgue measure, $\delta_{\mathbf{x}_i}$ is the Dirac measure supported at $\mathbf{x}_i$, and $\alpha_i \geq 0$ is the volume of the hole formed at $\mathbf{x}_i$ by the deformation $\mathbf{u}$ (see section 2.4 for further details). The existence of a minimizer in this class follows from [19], and it is a consequence of a result in [21] that if the matrix $\mathbf{A}$ in the boundary condition (1.3) is "sufficiently large,"[2] then any minimizer $\mathbf{u}$ must satisfy $\alpha_i > 0$ for at least one i.

In analogy with the radial problems outlined earlier, an alternative approach is to regularize by replacing the flaw points with preexisting voids in the reference configuration of maximum radius $\epsilon > 0$. The purpose of this paper is to examine the behavior of energy minimizers for these more regular problems in the limit as $\epsilon \to 0$. In particular we consider the case of one flaw point at $\mathbf{x}_0 \in \Omega$ and study the convergence of minimizers for these regularized problems to the minimizers obtained in [19]. For the convenience of the reader, we next outline the main convergence result in the case of the class of stored energy functions (1.4), deferring precise technical details to later in the paper. However, it should be noted that the results of this paper apply to much more general polyconvex energy functions (see hypotheses (H1)–(H4) in section 3).

[1] Condition (INV) requires, in particular, that $\mathbf{u} \in W^{1,p}(\Omega, \mathbb{R}^m)$, $p > m - 1$.

[2] This is the case, in particular, if $\mathbf{A} = t\mathbf{B}$, where $\mathbf{B} \in \mathrm{M}^{m\times m}_+$ is fixed and $t > 0$ is sufficiently large.

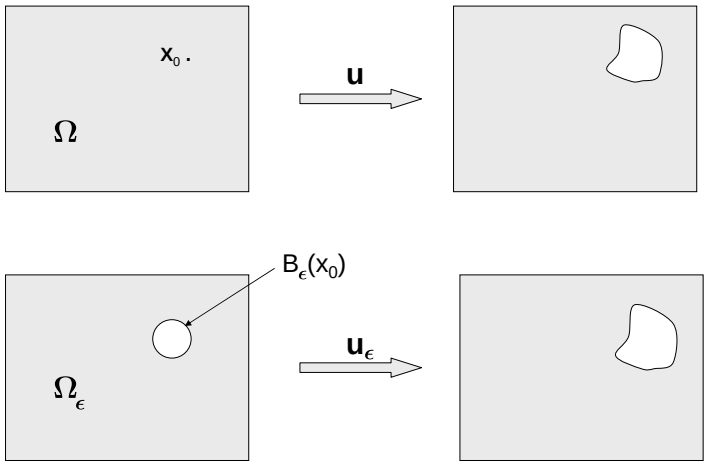


FIG. 2. *Deformations of the original and regularized domains.*

The underlying pure displacement problem. In essence, the underlying problem is to minimize the integral functional E given by (1.2) on a class of deformations[3,4] $\mathbf{u} \in W^{1,p}(\Omega, \mathbb{R}^m)$, $p \in (m-1, m)$, that satisfy (1.1), (1.3), condition (INV) on Ω, and

$$\mathrm{Det}\nabla\mathbf{u} = (\det \nabla\mathbf{u})\mathcal{L}^m + \alpha_0\delta_{\mathbf{x}_0},\ \alpha_0 \geq 0.$$

The existence of a minimizer for this problem follows from [19].

The regularized mixed displacement/traction problem. For each $\epsilon > 0$, define the domains $\Omega_\epsilon = \Omega\backslash\overline{B_\epsilon(\mathbf{x}_0)}$ which contain a preexisting hole of radius $\epsilon > 0$ centered on $\mathbf{x}_0$ (see Figure 2). Now consider the mixed displacement/traction problem on Ω_ϵ in which the outer boundary $\partial\Omega$ is subject to the same boundary displacements (1.3) and the deformed cavity surface (i.e., the image of ∂B_ϵ) is required to be stress-free. The corresponding variational problem is to minimize

$$E_\epsilon(\mathbf{u}_\epsilon) = \int_{\Omega_\epsilon} W(\mathbf{x}, \nabla\mathbf{u}_\epsilon(\mathbf{x}))\, d\mathbf{x}$$

on a class of deformations[5] $\mathbf{u}_\epsilon \in W^{1,p}(\Omega_\epsilon, \mathbb{R}^m)$, $p \in (m-1, m)$, that satisfy (1.3), condition (INV) on Ω_ϵ,

$$\det \nabla\mathbf{u}_\epsilon > 0 \text{ a.e. in } \Omega_\epsilon,$$

and

$$\mathrm{Det}\nabla\mathbf{u}_\epsilon = (\det \nabla\mathbf{u}_\epsilon)\mathcal{L}^m.$$

[3] In fact, for technical reasons, we work with the homogeneous extension $\mathbf{u}^e$ of $\mathbf{u}$ which is defined on a slightly larger domain $\Omega^e \supset \Omega$. It is obtained by extending $\mathbf{u}$ by the homogeneous deformation $\mathbf{Ax}$ (see section 2.3 and Remark 3.2).

[4] For interesting results in the borderline case, $p = m - 1$, see Conti and De Lellis [4].

[5] As in the pure displacement problem, we work with homogeneous extensions of these maps.

In Theorem 4.2 we prove convergence of minimizers for these regularized problems in the limit as $\epsilon \to 0$. In particular, we show that if $\epsilon_n \to 0$ and $(\mathbf{u}_{\epsilon_n})$ is a corresponding sequence of minimizers, then, passing to a subsequence if necessary, $\mathbf{u}_{\epsilon_n} \to \mathbf{u}$, where $\mathbf{u}$ is a minimizer for the pure displacement problem on the original domain Ω.

The above results are quite general and do not depend on the shape of the excluded regions used to obtain Ω_ϵ. In particular, the balls $B_\epsilon(\mathbf{x}_0)$ which are removed to produce Ω_ϵ could be replaced by a nested sequence of (nonspherical) regions around $\mathbf{x}_0$ whose diameters converge to zero as $\epsilon \to 0$.

In section 5, we consider the effect of regularizing the pure displacement variational problem by adding a surface energy term to the total energy which penalizes the formation and growth of cavities. In the standard model, in which surface energy is proportional to the new surface area created, cavitation is still energetically favorable for sufficiently severe boundary displacements. However, in this case we show that there is no longer bifurcation from a homogeneous deformation in the sense that there do not exist discontinuous energy minimizers producing cavities of arbitrarily small volume. Our results thus generalize the recent results of Dollhofer et al. [5], who show that the addition of surface energy in radial cavitation of an incompressible neo-Hookean material will induce the sudden formation of a single cavity of finite radius, rather than the gradual opening of a hole from zero volume.

2. Background.

2.1. Notation. Let Ω denote a nonempty, bounded, connected, open subset of $\mathbb{R}^m$ with Lipschitz boundary $\partial\Omega$ (see [6] or [13]). We denote by $L^p(\Omega)$ and $W^{1,p}(\Omega)$ the usual spaces of p-summable and Sobolev functions, respectively. We use the notation $L^p(\Omega;\mathbb{R}^m)$, etc., for vector-valued maps. A function ϕ is in $L^p_{\text{loc}}(\Omega)$ if $\phi \in L^p(U)$ for all open sets $U \subset\subset \Omega$; i.e., $U \subset K_U \subset \Omega$ for some compact set K_U. Weak convergence in these spaces will be indicated by the half arrow $\rightharpoonup$.

We denote m-dimensional Lebesgue measure by $\mathcal{L}^m$ and k-dimensional Hausdorff measure by $\mathcal{H}^k$. We write

$$B(\mathbf{z},\epsilon) := \{\mathbf{x} \in \mathbb{R}^m : |\mathbf{x}-\mathbf{z}| < \epsilon\}$$

for the open ball of radius $\epsilon > 0$ centered at $\mathbf{z} \in \mathbb{R}^m$ (we also use the notation $B_\epsilon(\mathbf{z})$ for $B(\mathbf{z},\epsilon)$).

Let $\mathbf{u} \in W^{1,p}(\Omega;\mathbb{R}^m)$ with $1 \le p < m$. We will be interested in pointwise properties of $\mathbf{u}$ as well as restrictions of $\mathbf{u}$ to lower-dimensional sets. We will not identify maps that are equal a.e. and choose to work with the precise representative $\mathbf{u}^* : \Omega \to \mathbb{R}^m$ defined by

$$\mathbf{u}^*(\mathbf{x}) := \begin{cases} \lim\limits_{\rho\to 0^+} \dfrac{1}{\mathcal{L}^m(B(\mathbf{x},\rho))} \displaystyle\int_{B(\mathbf{x},\rho)} \mathbf{u}(\mathbf{z})\, d\mathbf{z} & \text{if the limit exists,} \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

We shall make use of the fact that if $\mathbf{u} \in W^{1,p}(\Omega;\mathbb{R}^m)$ with $1 \le p < m$, then the above limit exists for every $\mathbf{x} \in \Omega\backslash P$, where $\mathcal{H}^{m-1}(P) = 0$. Thus, in particular, one can use the precise representative as a representative of the trace on $(m-1)$-dimensional surfaces. Moreover, if $p > m-1$, then $\mathcal{H}^1(P) = 0$, and consequently for each $\mathbf{z} \in \Omega$ the above limit is defined at *every* point on $\partial B(\mathbf{z},r)$ for almost every $r \in (0, r_{\mathbf{z}})$, where $r_{\mathbf{z}} = \text{dist}(\mathbf{z},\partial\Omega)$. For a thorough discussion of precise representative we refer the reader to [6].

2.2. The topological image. In this section we briefly recall some facts about the Brouwer degree (see, e.g., [7] or [22] for more details). Let $\mathbf{u} : \overline{\Omega} \to \mathbb{R}^m$ be a C^1 map. If $\mathbf{y}_0 \in \mathbb{R}^m \backslash \mathbf{u}(\partial\Omega)$ is such that $\det \nabla \mathbf{u}(\mathbf{x}) \neq 0$ for all $\mathbf{x} \in \mathbf{u}^{-1}(\mathbf{y}_0)$, then the degree is defined by

$$\deg(\mathbf{u}, \Omega, \mathbf{y}_0) := \sum_{\mathbf{x} \in \mathbf{u}^{-1}(\mathbf{y}_0)} \operatorname{sgn}[\det \nabla \mathbf{u}(\mathbf{x})], \tag{2.1}$$

where $\operatorname{sgn}(t) = 1$ for $t > 0$ and $\operatorname{sgn}(t) = -1$ for $t < 0$. In particular, if $\mathbf{g} : \overline{\Omega} \to \mathbb{R}^m$ is a diffeomorphism with $\det \nabla \mathbf{g} > 0$ on Ω, then from (2.1) we conclude that

$$\deg(\mathbf{g}, \Omega, \mathbf{y}_0) = \begin{cases} 1 & \text{if } \mathbf{y}_0 \in \mathbf{g}(\Omega), \\ 0 & \text{if } \mathbf{y}_0 \in \mathbb{R}^m \backslash \mathbf{g}(\overline{\Omega}). \end{cases}$$

If ϕ is a C^∞ function supported in the connected component of $\mathbb{R}^m \backslash \mathbf{u}(\partial\Omega)$ that contains $\mathbf{y}_0$, then one can show that

$$\int_\Omega \phi(\mathbf{u}(\mathbf{x})) \det \nabla \mathbf{u}(\mathbf{x})\, d\mathbf{x} = \deg(\mathbf{u}, \Omega, \mathbf{y}_0) \int_{\mathbb{R}^m} \phi(\mathbf{y})\, d\mathbf{y}.$$

One can now define $\deg(\mathbf{u}, \Omega, \mathbf{y})$ for any continuous function $\mathbf{u} : \Omega \to \mathbb{R}^m$ and any $\mathbf{y} \in \mathbb{R}^m \backslash \mathbf{u}(\partial\Omega)$ by using this formula and approximating by C^∞ functions. Moreover, the degree depends only on $\mathbf{u}|_{\partial\Omega}$. Accordingly we can write $\deg(\mathbf{u}, \partial\Omega, \mathbf{y})$ instead of $\deg(\mathbf{u}, \Omega, \mathbf{y})$.

DEFINITION 2.1. *Let $B(\mathbf{z}, r) \subset \Omega$ and suppose that $\overline{\mathbf{u}} : \partial B(\mathbf{z}, r) \to \mathbb{R}^m$ is continuous. We define the topological image of $B(\mathbf{z}, r)$ under $\overline{\mathbf{u}}$ by*

$$\operatorname{im}_{\mathrm{T}}(\overline{\mathbf{u}}, B(\mathbf{z}, r)) := \{\mathbf{y} \in \mathbb{R}^m \backslash \overline{\mathbf{u}}(\partial B(\mathbf{z}, r)) : \ \deg(\overline{\mathbf{u}}, \partial B(\mathbf{z}, r), \mathbf{y}) \neq 0\}. \tag{2.2}$$

Remark 2.2. Let $\mathbf{g} : \overline{B(\mathbf{z}, r)} \to \mathbb{R}^m$ be a homeomorphism. If $\overline{\mathbf{u}} : \partial B(\mathbf{z}, r) \to \mathbb{R}^m$ is such that $\overline{\mathbf{u}}(\partial B(\mathbf{z}, r)) = \mathbf{g}(\partial B(\mathbf{z}, r))$, then

$$\operatorname{im}_{\mathrm{T}}(\overline{\mathbf{u}}, B(\mathbf{z}, r)) = \mathbf{g}(B(\mathbf{z}, r)).$$

2.3. Invertibility condition (INV). In nonlinear elasticity one is interested in globally invertible maps since, in general, matter cannot interpenetrate itself. We say that $\mathbf{u} \in W^{1,p}(\Omega, \mathbb{R}^m)$, $p \geq 1$, is *one-to-one* a.e. if there is a Lebesgue null set $N \subset \Omega$ such that $\mathbf{u}|_{\Omega \backslash N}$ is injective. Unfortunately, if $p < m$, the weak limit of a sequence of maps which are one-to-one a.e. need not be one-to-one a.e. (see, e.g., [16, section 11]). A property that is slightly stronger than one-to-one a.e. is therefore needed.

DEFINITION 2.3. *Let $r_{\mathbf{z}} = \operatorname{dist}(\mathbf{z}, \partial\Omega)$. We say that $\mathbf{u} : \Omega \to \mathbb{R}^m$ satisfies invertibility condition* (INV) *on Ω, provided that for every $\mathbf{z} \in \Omega$ there exists an $\mathcal{L}^1$ null set $N_{\mathbf{z}}$ such that, for all $r \in (0, r_{\mathbf{z}}) \backslash N_{\mathbf{z}}$,*

(o) *$\mathbf{u}|_{\partial B(\mathbf{z}, r)}$ is continuous;*

(i) *$\mathbf{u}(\mathbf{x}) \in \operatorname{im}_{\mathrm{T}}(\mathbf{u}, B(\mathbf{z}, r)) \cup \mathbf{u}(\partial B(\mathbf{z}, r))$ for $\mathcal{L}^m$ a.e. $\mathbf{x} \in \overline{B(\mathbf{z}, r)}$;*

(ii) *$\mathbf{u}(\mathbf{x}) \in \mathbb{R}^m \backslash \operatorname{im}_{\mathrm{T}}(\mathbf{u}, B(\mathbf{z}, r))$ for $\mathcal{L}^m$ a.e. $\mathbf{x} \in \Omega \backslash \overline{B(\mathbf{z}, r)}$.*

The next results show that condition (INV) is preserved under weak convergence and that mappings that satisfy condition (INV) and have nonzero Jacobian a.e. are one-to-one a.e.

PROPOSITION 2.4 (see [16, Lemma 3.3]). *Let $p > m-1$ and suppose that $(\mathbf{u}_n^*)$ is a sequence in $W^{1,p}(\Omega;\mathbb{R}^m)$ that satisfies condition* (INV). *Suppose also that*

$$\mathbf{u}_n \rightharpoonup \mathbf{u} \quad \text{in} \quad W^{1,p}(\Omega;\mathbb{R}^m).$$

Then $\mathbf{u}^*$ satisfies condition (INV).

PROPOSITION 2.5 (see [16, Lemma 3.4]). *Let $\mathbf{u} \in W^{1,p}(\Omega;\mathbb{R}^m)$ with $p > m-1$. Suppose that $\det \nabla \mathbf{u} \neq 0$ a.e. and that $\mathbf{u}^*$ satisfies condition* (INV). *Then $\mathbf{u}$ is one-to-one a.e.*

DEFINITION 2.6. *We say that the function $\mathbf{u}$ satisfies condition* (INV) *on $\Omega\backslash\{\mathbf{x}_0\}$ if $\mathbf{u}$ satisfies condition* (INV) *on $\Omega\backslash\overline{B(\mathbf{x}_0,\delta)}$ for every sufficiently small $\delta > 0$.*

Homogeneous extensions of maps. Let $\Omega \subset\subset \Omega^{\mathrm{e}}$, where Ω^{e} is a bounded, open, connected set with smooth boundary and suppose that $\mathbf{u}^{\mathrm{h}} : \Omega^{\mathrm{e}} \to \mathbb{R}^m$ is the orientation preserving homogeneous map

$$\mathbf{u}^{\mathrm{h}}(\mathbf{x}) \equiv \mathbf{A}\mathbf{x},$$

where $\mathbf{A} \in \mathrm{M}^{m\times m}_+$. If $\mathbf{u} \in W^{1,p}(\Omega;\mathbb{R}^m)$ satisfies $\mathbf{u} = \mathbf{u}^{\mathrm{h}}$ on $\partial\Omega$, then we define its homogeneous extension $\mathbf{u}^e : \Omega^{\mathrm{e}} \to \mathbb{R}^m$ by

$$\mathbf{u}^e(\mathbf{x}) := \begin{cases} \mathbf{u}(\mathbf{x}) & \text{if } \mathbf{x} \in \Omega, \\ \mathbf{A}\mathbf{x} & \text{if } \mathbf{x} \in \Omega^{\mathrm{e}}\backslash\Omega \end{cases} \tag{2.3}$$

and note that $\mathbf{u}^e \in W^{1,p}(\Omega^{\mathrm{e}};\mathbb{R}^m)$. More generally, let $\mathbf{x}_0 \in \Omega$ and $\epsilon \geq 0$ satisfy $B(\mathbf{x}_0,\epsilon) \subset\subset \Omega \subset\subset \Omega^{\mathrm{e}}$ and define

$$\Omega_\epsilon := \Omega\backslash\overline{B(\mathbf{x}_0,\epsilon)}, \qquad \Omega^{\mathrm{e}}_\epsilon := \Omega^{\mathrm{e}}\backslash\overline{B(\mathbf{x}_0,\epsilon)} \tag{2.4}$$

(this corresponds to a preexisting void of radius ϵ in the reference configuration). If $\mathbf{u}_\epsilon \in W^{1,p}(\Omega_\epsilon;\mathbb{R}^m)$ satisfies $\mathbf{u}_\epsilon = \mathbf{u}^{\mathrm{h}}$ on $\partial\Omega$, then we define its homogeneous extension $\mathbf{u}^e_\epsilon : \Omega^{\mathrm{e}}_\epsilon \to \mathbb{R}^m$ by

$$\mathbf{u}^e_\epsilon(\mathbf{x}) := \begin{cases} \mathbf{u}_\epsilon(\mathbf{x}) & \text{if } \mathbf{x} \in \Omega_\epsilon, \\ \mathbf{A}\mathbf{x} & \text{if } \mathbf{x} \in \Omega^{\mathrm{e}}\backslash\Omega \end{cases} \tag{2.5}$$

and note that[6] $\mathbf{u}^e_\epsilon \in W^{1,p}(\Omega^{\mathrm{e}}_\epsilon;\mathbb{R}^m)$. The use of such an extension allows us to obtain the following result, whose proof we omit since it is similar to that of Theorem TL in [17].

LEMMA 2.7. *Let $\mathbf{u}_\epsilon \in W^{1,p}(\Omega_\epsilon;\mathbb{R}^m)$, $p > m-1$, $\epsilon \geq 0$, satisfy $\mathbf{u}_\epsilon = \mathbf{u}^{\mathrm{h}}$ on $\partial\Omega$. Suppose that its homogeneous extension $\mathbf{u}^e_\epsilon$, given by* (2.5), *is one-to-one a.e. on $\Omega^{\mathrm{e}}_\epsilon$. Then*

$$\mathbf{u}_\epsilon(\mathbf{x}) \in \mathbf{u}^{\mathrm{h}}(\overline{\Omega}) \quad \textit{for a.e. } \mathbf{x} \in \Omega_\epsilon.$$

[6] If $\epsilon = 0$, then there is no preexisting hole. In this case, $\Omega_0 = \Omega$, $\Omega^{\mathrm{e}}_0 = \Omega^{\mathrm{e}}$, and thus $\mathbf{u}_0 = \mathbf{u} \in W^{1,p}(\Omega;\mathbb{R}^m)$ and $\mathbf{u}^e_0 = \mathbf{u}^e \in W^{1,p}(\Omega^{\mathrm{e}};\mathbb{R}^m)$ in accordance with the earlier definition (2.3).

2.4. The distributional Jacobian. Given a mapping $\mathbf{u} \in W^{1,p}(\Omega, \mathbb{R}^m)$, with $p > \frac{m^2}{m+1}$, it follows (using the Sobolev embedding theorem) that the distributional Jacobian defined by

$$(\mathrm{Det}\nabla\mathbf{u})(\phi) := -\int_\Omega \frac{1}{m}\left([\mathrm{adj}\,\nabla\mathbf{u}]\mathbf{u}\right)\cdot\nabla\phi\,d\mathbf{x} \quad \forall\phi \in C_0^\infty(\Omega) \tag{2.6}$$

is a well-defined distribution (where $\mathrm{adj}\,\nabla\mathbf{u}$ denotes the adjugate matrix of $\nabla\mathbf{u}$, that is, the transposed matrix of cofactors of $\nabla\mathbf{u}$). The definition follows from the well-known formula for expressing $\det\nabla\mathbf{u}$ as a divergence (see, e.g., [14] for further details and references).

Next, suppose that $\mathbf{u} \in W^{1,p}(\Omega;\mathbb{R}^m)$, $p > m-1$, satisfies (INV) on Ω. Then $\mathbf{u} \in L^\infty_{\mathrm{loc}}(\Omega)$, and hence $\mathrm{Det}\,\nabla\mathbf{u}$ is again a well-defined distribution. Moreover, it follows from [16, Lemma 8.1] that if $\mathbf{u}$ further satisfies $\det\nabla\mathbf{u} > 0$ a.e., then $\mathrm{Det}\,\nabla\mathbf{u}$ is a Radon measure and

$$\mathrm{Det}\,\nabla\mathbf{u} = (\det\nabla\mathbf{u})\mathcal{L}^m + \mu^s,$$

where μ^s is singular with respect to $\mathcal{L}^m$. In this paper we will be interested in the case when μ^s is a Dirac measure[7] of the form $\alpha\delta_{\mathbf{x}_0}$ (where $\alpha > 0$ and $\mathbf{x}_0 \in \Omega$) which corresponds to $\mathbf{u}$ creating a cavity of volume α at the point $\mathbf{x}_0$.[8]

Example 2.8. Let $\Omega = B$ be the unit ball in $\mathbb{R}^3$ (i.e., $m = 3$) and let

$$\mathbf{u}(\mathbf{x}) = (|\mathbf{x}| + c)\frac{\mathbf{x}}{|\mathbf{x}|}, \quad c > 0. \tag{2.7}$$

Then $\mathbf{u}$ produces a hole of radius c at the center of the deformed ball. In this case it can be shown that

$$\mathrm{Det}\,\nabla\mathbf{u} = (\det\nabla\mathbf{u})\mathcal{L}^3 + \frac{4}{3}\pi c^3\delta_{\mathbf{0}}. \tag{2.8}$$

The last expression (2.8) is to be interpreted in the sense of distributions so that

$$(\mathrm{Det}\,\nabla\mathbf{u})(\phi) = \int_\Omega (\det\nabla\mathbf{u}(\mathbf{x}))\phi(\mathbf{x})\,d\mathbf{x} + \frac{4}{3}\pi c^3\phi(0) \quad \forall\phi \in C_0^\infty(\Omega)$$

(notice that the coefficient of $\delta_{\mathbf{0}}$ in (2.8) is the volume of the hole that is formed at the origin under the deformation (2.7)).

3. The energy: Existence of minimizers. We consider an m-dimensional elastic body which, in its reference state, occupies the region $\Omega \subset \mathbb{R}^m$. We let $W \in C(\overline{\Omega} \times \mathrm{M}^{m\times m}_+; [0,\infty))$ denote the stored energy function for the body. For ease of exposition we state the following conditions on W in the case of three dimensions (i.e., $m = 3$). Let $p > 2 = m-1$, $D = \mathrm{M}^{3\times3} \times \mathrm{M}^{3\times3} \times (0,\infty)$; then we will refer to the following hypotheses on W:

(H1) (polyconvexity) there exists $\Phi : \Omega \times D \to \overline{\mathbb{R}}$ such that for a.e. $\mathbf{x} \in \Omega$

$$W(\mathbf{x},\mathbf{F}) = \Phi(\mathbf{x}, (\mathbf{F}, \mathrm{adj}\,\mathbf{F}, \det\mathbf{F})) \quad \forall\,\mathbf{F} \text{ such that } \det\mathbf{F} > 0,$$

where $\Phi(\mathbf{x},\cdot) : D \to \overline{\mathbb{R}}$ is convex for a.e. $\mathbf{x} \in \Omega$;

[7] Other assumptions on the support of the singular measure μ^s may be relevant for modelling different forms of fracture. See also [15] for further results on the singular support of the distributional Jacobian.

[8] Note that such a cavity need not be spherical.

(H2) (continuity) $\Phi(\mathbf{x},\cdot) : D \to \overline{\mathbb{R}}$ is continuous for a.e. $\mathbf{x} \in \Omega$ and $\Phi(\cdot, N) : \Omega \to \overline{\mathbb{R}}$ is measurable for every $N \in D$;

(H3) (coercivity) $W(\mathbf{x},\mathbf{F}) \geq C|\mathbf{F}|^p + \Gamma(\det \mathbf{F}) + K$ for a.e. $\mathbf{x} \in \Omega$, where $C > 0$, K are constants, and $\Gamma : (0,\infty) \to \mathbb{R}$ is a convex function satisfying $\Gamma(t)/t \to +\infty$ as $t \to +\infty$;

(H4) $\Gamma(t) \to +\infty$ as $t \to 0^+$.

Now fix $\mathbf{x}_0 \in \Omega$ and for each $0 \leq \epsilon < \operatorname{dist}(\mathbf{x}_0, \partial\Omega)$ recall the definition of the homogeneous extension of a map given in section 2.3 (see (2.3)–(2.5) in particular). For each such ϵ we seek a minimizer for the total elastic energy

$$E_\epsilon(\mathbf{u}_\epsilon) = \int_{\Omega_\epsilon} W(\mathbf{x}, \nabla \mathbf{u}_\epsilon(\mathbf{x}))\, d\mathbf{x}$$

in the class of admissible functions

$$\mathcal{A}_\epsilon(\mathbf{x}_0) = \{\mathbf{u}_\epsilon \in W^{1,p}(\Omega_\epsilon;\mathbb{R}^m) : \mathbf{u}_\epsilon|_{\partial\Omega} = \mathbf{u}^{\mathrm{h}},\ (\mathbf{u}_\epsilon^e)^* \text{ satisfies (INV) on } \Omega_\epsilon^{\mathrm{e}},\ \det \nabla \mathbf{u}_\epsilon > 0 \text{ a.e.},\ \operatorname{Det} \nabla \mathbf{u}_\epsilon^e = (\det \nabla \mathbf{u}_\epsilon^e)\mathcal{L}^m\}$$

if $\epsilon > 0$, and in the class

$$\mathcal{A}_0(\mathbf{x}_0) = \{\mathbf{u} \in W^{1,p}(\Omega;\mathbb{R}^m) : \mathbf{u}|_{\partial\Omega} = \mathbf{u}^{\mathrm{h}},\ (\mathbf{u}^e)^* \text{ satisfies (INV) on } \Omega^{\mathrm{e}},\ \det \nabla \mathbf{u} > 0 \text{ a.e.},\ \operatorname{Det} \nabla \mathbf{u}^e = (\det \nabla \mathbf{u}^e)\mathcal{L}^m + \alpha_{\mathbf{u}} \delta_{\mathbf{x}_0}\} \tag{3.1}$$

if $\epsilon = 0$, where $\alpha_{\mathbf{u}} \geq 0$ is a scalar depending on the map $\mathbf{u}$ and $\delta_{\mathbf{x}_0}$ denotes the Dirac measure with support at $\mathbf{x}_0$. Thus, $\mathcal{A}_0(\mathbf{x}_0)$ contains maps $\mathbf{u}$ that produce a cavity of volume $\alpha_{\mathbf{u}}$ located at $\mathbf{x}_0 \in \Omega$.

PROPOSITION 3.1 (see [19, Theorem 4.1]). *Let $p > m-1$ and $\epsilon \geq 0$. Suppose that $W \in C(\overline{\Omega} \times \mathrm{M}^{m\times m}_+; [0,\infty))$ satisfies hypotheses* (H1)–(H4). *Then E_ϵ attains its infimum on $\mathcal{A}_\epsilon(\mathbf{x}_0)$.*[9]

Thus both the *mixed displacement/traction problem* ($\epsilon > 0$) and the *pure displacement problem* ($\epsilon = 0$) have energy minimizers. In the next section we will show that a subsequence of the minimizers of the mixed problems converges to a minimizer of the displacement problem as the preexisting hole size shrinks to zero.

Remark 3.2. The reasons for requiring that the homogeneous extensions $\mathbf{u}^e, \mathbf{u}_\epsilon^e$, rather than the original maps $\mathbf{u}, \mathbf{u}_\epsilon$, satisfy condition (INV) are to, first, prevent the phenomenon of cavitation at the boundary of Ω (see [16, p. 55]) and, second, to prevent leakage at the boundary (see [16, p. 56] and [17, p. 975]). For interesting related results see Swanson and Ziemer [25, 26].

Remark 3.3. Suppose that in the above theorem $\mathbf{A} = \lambda \mathbf{I}$ and $\lambda > \lambda_{\mathrm{crit}}$ (where λ_{crit} is the critical boundary displacement after which radial cavitation occurs). Suppose further that the energy grows sufficiently slowly with respect to $|\mathbf{F}|$; e.g., W is given by (1.4) for some $c > 0$ and $p \in (m-1, m)$ with Γ as in (H3) and (H4). It then follows from results on radial cavitation (see [20]) that any minimizer $\mathbf{u}$ given by the above result with $\epsilon = 0$ must satisfy $\alpha_{\mathbf{u}} > 0$; i.e., it must form a new cavity.

4. Convergence of minimizers. In this section we show that, as $\epsilon \to 0^+$, minimizers $\mathbf{u}_\epsilon$ of the mixed displacement/traction problem given by Proposition 3.1 converge to a minimizer of the pure displacement problem $\mathbf{u} \in \mathcal{A}_0(\mathbf{x}_0)$. More precisely, we prove the following main theorem of the paper.

[9] We refer the reader to [4] for interesting analytical difficulties that arise in the borderline case $p = m - 1$.

THEOREM 4.1. *Let $\mathbf{x}_0 \in \Omega$ be fixed and let $W \in C(\overline{\Omega} \times \mathrm{M}^{m\times m}_+; [0,\infty))$ satisfy* (H1)–(H4). *Suppose that (ϵ_n) is a monotone decreasing sequence converging to zero and let $(\mathbf{u}_{\epsilon_n})$ be a corresponding sequence of minimizers whose existence is given by Proposition* 3.1. *Then there exists $\mathbf{u} \in \mathcal{A}_0(\mathbf{x}_0)$ and a subsequence $(\mathbf{u}_{\epsilon_{n_j}})$ such that*

$$\mathbf{u}_{\epsilon_{n_j}} \rightharpoonup \mathbf{u} \quad \text{as} \quad j \to \infty \quad \text{in} \quad W^{1,p}(\Omega^{\mathrm{e}}_\delta; \mathbb{R}^m)$$

for any $\delta > 0$. Moreover, $\mathbf{u}$ is a minimizer of E_0 on $\mathcal{A}_0(\mathbf{x}_0)$.

The proof of this theorem is contained in the remainder of this section. Throughout this section (ϵ_n) will denote a fixed monotone decreasing sequence converging to zero, while $(\mathbf{u}_{\epsilon_n})$ will denote a corresponding sequence of minimizers whose existence is given by Proposition 3.1.

The convergence proof is split into three parts: We first identify a map $\mathbf{u} \in W^{1,p}(\Omega; \mathbb{R}^m)$ such that for a subsequence $(\mathbf{u}_{\epsilon_{n_j}})$ of $(\mathbf{u}_{\epsilon_n})$ we have $\mathbf{u}_{\epsilon_{n_j}} \rightharpoonup \mathbf{u}$ in $W^{1,p}(\Omega_\delta; \mathbb{R}^m)$ as $j \to \infty$ for any $\delta > 0$ sufficiently small. It then follows from Proposition 2.4 and Definition 2.6 that $\mathbf{u}$ satisfies condition (INV) on $\Omega\backslash\{\mathbf{x}_0\}$. In section 4.2 we show in Theorem 4.3 that it then follows that $\mathbf{u}$ satisfies (INV) on Ω and in Lemma 4.5 that $\operatorname{Det}\nabla\mathbf{u} = (\det\nabla\mathbf{u})\mathcal{L}^m + \alpha_{\mathbf{u}}\delta_{\mathbf{x}_0}$. Thus we can conclude that $\mathbf{u} \in \mathcal{A}_0(\mathbf{x}_0)$. Finally, in section 4.3 in Theorem 4.6 we prove that $\mathbf{u}$ is a minimizer of the energy E_0 on $\mathcal{A}_0(\mathbf{x}_0)$ using lower semicontinuity arguments.

4.1. Identifying a weak limit for the sequence of minimizers.

THEOREM 4.2. *Let $(\mathbf{u}_{\epsilon_n})$ be a sequence of minimizers given by Proposition* 3.1. *Then there is a subsequence $(\mathbf{u}_{\epsilon_{n_j}})$ and a mapping $\mathbf{u} : \Omega \to \mathbb{R}^m$ such that, for any $\delta > 0$, $\mathbf{u} \in W^{1,p}(\Omega\backslash\overline{B_\delta}; \mathbb{R}^m)$ and*

$$\mathbf{u}_{\epsilon_{n_j}} \rightharpoonup \mathbf{u} \quad \text{as} \quad j \to \infty \quad \text{in} \quad W^{1,p}(\Omega\backslash\overline{B(\mathbf{x}_0,\delta)}; \mathbb{R}^m).$$

Proof. Let $(\mathbf{u}_{\epsilon_n})$ be a sequence of minimizers given by Proposition 3.1. We first note that, since W is continuous and nonnegative on the compact set $\overline{\Omega} \times \{\mathbf{A}\}$, the homogeneous deformation $\mathbf{u}^{\mathrm{h}}$ has finite energy. Hence for any $n \in \mathbb{N}$,

$$E_{\epsilon_n}(\mathbf{u}^{\mathrm{h}}) = \int_{\Omega_{\epsilon_n}} W(\mathbf{x}, \mathbf{A})\, d\mathbf{x} \le \int_\Omega W(\mathbf{x}, \mathbf{A})\, d\mathbf{x} < \infty. \tag{4.1}$$

Next, by the convexity of Γ and its growth at zero and infinity ((H3) and (H4)) it follows that Γ is bounded below. Thus, by the coercivity condition (H3) and the Poincaré inequality, we find that for any $N \in \mathbb{N}$ and $n > N$

$$E_{\epsilon_N}(\mathbf{u}_{\epsilon_n}) \ge C_1 \|\mathbf{u}_{\epsilon_n}\|^p_{W^{1,p}(\Omega_{\epsilon_N})} - C_2\mathcal{L}^m(\Omega_{\epsilon_N}), \tag{4.2}$$

where C_1 and C_2 are positive constants. In addition, W is nonnegative and $\Omega_{\epsilon_N} \subset \Omega_{\epsilon_n}$ so that

$$\begin{aligned} E_{\epsilon_N}(\mathbf{u}_{\epsilon_n}) &= \int_{\Omega_{\epsilon_N}} W(\mathbf{x}, \nabla\mathbf{u}_{\epsilon_n}(\mathbf{x}))\, d\mathbf{x} \\ &\le \int_{\Omega_{\epsilon_n}} W(\mathbf{x}, \nabla\mathbf{u}_{\epsilon_n}(\mathbf{x}))\, d\mathbf{x} \quad = \quad E_{\epsilon_n}(\mathbf{u}_{\epsilon_n}). \end{aligned} \tag{4.3}$$

Also, $\mathbf{u}^{\mathrm{h}} \in \mathcal{A}_{\epsilon_n}(\mathbf{x}_0)$ and $\mathbf{u}_{\epsilon_n}$ is a minimizer of E_{ϵ_n} on $\mathcal{A}_{\epsilon_n}(\mathbf{x}_0)$, and so

$$E_{\epsilon_n}(\mathbf{u}_{\epsilon_n}) \le E_{\epsilon_n}(\mathbf{u}^{\mathrm{h}}). \tag{4.4}$$

Therefore by (4.1)–(4.4) the sequence $(\mathbf{u}_{\epsilon_n})$ is bounded in the reflexive Banach space $W^{1,p}(\Omega_{\epsilon_N};\mathbb{R}^m)$, and consequently there exists a subsequence, still labeled $(\mathbf{u}_{\epsilon_n})$, converging weakly to some function $\mathbf{u}^N$ in $W^{1,p}(\Omega_{\epsilon_N};\mathbb{R}^m)$; that is,

$$\mathbf{u}_{\epsilon_n} \rightharpoonup \mathbf{u}^N \quad \text{in} \quad W^{1,p}(\Omega_{\epsilon_N};\mathbb{R}^m) \quad \text{as} \quad n \to \infty.$$

Now inductively take successive subsequences with $N = 1, 2, 3, \dots$ and then choose a diagonal sequence to obtain a subsequence, labeled $(\mathbf{u}_{\epsilon_{n_j}})$, of $(\mathbf{u}_{\epsilon_n})$, that satisfies

$$\mathbf{u}_{\epsilon_{n_j}} \rightharpoonup \mathbf{u} \quad \text{in} \quad W^{1,p}(\Omega_{\epsilon_N};\mathbb{R}^m) \quad \text{as} \quad j \to \infty,$$

where $\mathbf{u} : \Omega\backslash\{\mathbf{x}_0\} \to \mathbb{R}^m$ is defined by

$$\mathbf{u}(\mathbf{x}) := \begin{cases} \mathbf{u}^1(\mathbf{x}) & \text{if } \mathbf{x} \in \Omega_{\epsilon_1}, \\ \mathbf{u}^N(\mathbf{x}) & \text{if } \mathbf{x} \in \Omega_{\epsilon_N}\backslash\Omega_{\epsilon_{N-1}}, \end{cases}$$

for $N = 2, 3, 4, \dots$. By construction $\mathbf{u} \in W^{1,p}(\Omega_{\epsilon_N};\mathbb{R}^m)$ for any N. ∎

Note that $\mathbf{u}$ is well defined by the uniqueness of weak limits. We will henceforth use $(\mathbf{u}_{\epsilon_n})$ to denote the subsequence of minimizers $(\mathbf{u}_{\epsilon_{n_j}})$ obtained in Theorem 4.2. It now follows from standard arguments that the limit function $\mathbf{u}$ identified in the above theorem lies in $W^{1,p}(\Omega;\mathbb{R}^m)$ for $p > m - 1$. For example, let $\psi \in C^\infty(\mathbb{R})$ be a fixed monotone increasing function that satisfies

$$\psi(t) = \begin{cases} 0 & \text{if } t \le 1, \\ 1 & \text{if } t \ge 4/3. \end{cases}$$

For each $n \in \mathbb{N}$, extend $\mathbf{u}_{\epsilon_n} : \Omega_{\epsilon_n} \to \mathbb{R}^m$ to a mapping $\tilde{\mathbf{u}}_{\epsilon_n} : \Omega \to \mathbb{R}^m$ by defining

$$\tilde{\mathbf{u}}_{\epsilon_n}(\mathbf{x}) := \begin{cases} \psi\left(\frac{2|\mathbf{x}-\mathbf{x}_0|}{|\mathbf{x}-\mathbf{x}_0|+\epsilon_n}\right)\mathbf{u}_{\epsilon_n}(\mathbf{x}) & \text{for } \mathbf{x} \in \Omega_{\epsilon_n}, \\ 0 & \text{otherwise.} \end{cases}$$

Then verify that $\tilde{\mathbf{u}}_{\epsilon_n} \rightharpoonup \mathbf{u}$ in $W^{1,p}(\Omega;\mathbb{R}^m)$ as $n \to \infty$.

4.2. The weak limit u lies in $\mathcal{A}_0(\mathbf{x}_0)$. In this section we prove that the weak limit $\mathbf{u}$ of the sequence $(\mathbf{u}_{\epsilon_n})$ is in the class of admissible functions $\mathcal{A}_0(\mathbf{x}_0)$. Since $\mathbf{u}_{\epsilon_n} \rightharpoonup \mathbf{u}$ in $W^{1,p}(\Omega_{\epsilon_N};\mathbb{R}^m)$ for any N, standard results imply that $\mathbf{u} = \mathbf{u}^{\mathrm{h}}$ on $\partial\Omega$, and the arguments in the proof of Theorem 4.1 in [19] show that $\det\nabla\mathbf{u}^e > 0$ a.e. Since $\mathbf{u}^{\mathrm{h}}$ is a diffeomorphism, it follows that $\mathbf{u}^{\mathrm{h}}(\overline{\Omega})$ is closed, and, since we may assume $\mathbf{u}_{\epsilon_n} \to \mathbf{u}$ a.e. in Ω, we conclude that $\mathbf{u}(\mathbf{x}) \in \mathbf{u}^{\mathrm{h}}(\overline{\Omega})$ for a.e. $\mathbf{x} \in \Omega$ and consequently

$$\mathbf{u}^e(\mathbf{x}) \in \mathbf{u}^{\mathrm{h}}(\Omega^{\mathrm{e}}) \quad \text{for a.e. } \mathbf{x} \in \Omega^{\mathrm{e}}. \tag{4.5}$$

We next prove that if $\mathbf{u}^e$ satisfies condition (INV) on $\Omega^{\mathrm{e}}\backslash\{\mathbf{x}_0\}$, then $\mathbf{u}^e$ satisfies condition (INV) on Ω^{e}. Following this, we then prove in Lemma 4.5 that the distributional Jacobian of $\mathbf{u}^e$ has the appropriate form, which will complete the proof that $\mathbf{u}^e$ lies in $\mathcal{A}_0(\mathbf{x}_0)$.

THEOREM 4.3. *Let $p > m - 1$ and suppose that $\mathbf{u}_{\epsilon_n} \in \mathcal{A}_{\epsilon_n}(\mathbf{x}_0)$ is the sequence of minimizers given in Proposition* 3.1. *Suppose further that there exists $\mathbf{u}^e \in W^{1,p}(\Omega^{\mathrm{e}};\mathbb{R}^m)$ such that $\det\nabla\mathbf{u}^e > 0$ a.e. and that for any fixed $N \in \mathbb{N}$*

$$\mathbf{u}^e_{\epsilon_n} \rightharpoonup \mathbf{u}^e \quad \text{in} \quad W^{1,p}(\Omega^{\mathrm{e}}_{\epsilon_N};\mathbb{R}^m). \tag{4.6}$$

Then $(\mathbf{u}^e)^*$ satisfies (INV) *on Ω^{e}.*

Proof. Without loss of generality we take $\mathbf{u}^e_{\epsilon_n} = (\mathbf{u}^e_{\epsilon_n})^*$, $\mathbf{u}^e = (\mathbf{u}^e)^*$, and fix $\mathbf{x}_1 \in \Omega^{\mathrm{e}}$. Then we must show that for $\mathcal{L}^1$ a.e. $r \in (0, \operatorname{dist}(\mathbf{x}_1, \partial\Omega^{\mathrm{e}}))$

$$\begin{aligned} &\text{(i)} \quad \mathbf{u}^e(\mathbf{x}) \in \operatorname{im}_{\mathrm{T}}(\mathbf{u}^e, B_r(\mathbf{x}_1)) \cup \mathbf{u}^e(\partial B_r(\mathbf{x}_1)) \text{ for a.e. } \mathbf{x} \in \overline{B_r(\mathbf{x}_1)};\\ &\text{(ii)} \quad \mathbf{u}^e(\mathbf{x}) \in \mathbb{R}^m \backslash \operatorname{im}_{\mathrm{T}}(\mathbf{u}^e, B_r(\mathbf{x}_1)) \text{ for a.e. } \mathbf{x} \in \Omega^{\mathrm{e}} \backslash \overline{B_r(\mathbf{x}_1)}, \end{aligned} \tag{4.7}$$

where $B_r(\mathbf{x}_1) := B(\mathbf{x}_1, r)$.

Let N be fixed and suppose that $n > N$. By definition of $\mathcal{A}_{\epsilon_n}(\mathbf{x}_0)$ the minimizers $\mathbf{u}^e_{\epsilon_n}$ satisfy condition (INV) on $\Omega^{\mathrm{e}}_{\epsilon_N}$ and $\det \nabla \mathbf{u}^e_{\epsilon_n} > 0$ a.e. Consequently, by (4.6) and Proposition 2.4, $\mathbf{u}^e$ satisfies condition (INV) on $\Omega^{\mathrm{e}}_{\epsilon_N}$. Since N is arbitrary, it follows that $\mathbf{u}^e$ satisfies condition (INV) on $\Omega^{\mathrm{e}} \backslash \{\mathbf{x}_0\}$. Next, fix $r_1 > 0$ such that $B(\mathbf{x}_1, r_1) \subset \Omega^{\mathrm{e}}$ and $\mathbf{x}_0 \notin \partial B(\mathbf{x}_1, r_1)$. Then for all r sufficiently close to r_1 either $\mathbf{x}_0 \in B(\mathbf{x}_1, r)$ or $\mathbf{x}_0 \in \Omega^{\mathrm{e}} \backslash \overline{B(\mathbf{x}_1, r)}$. In the former case define

$$U_r := \Omega^{\mathrm{e}} \backslash \overline{B(\mathbf{x}_1, r)} \quad \text{and} \quad V_r := B(\mathbf{x}_1, r) \tag{4.8}$$

and in the latter case define

$$U_r := B(\mathbf{x}_1, r) \quad \text{and} \quad V_r := \Omega^{\mathrm{e}} \backslash \overline{B(\mathbf{x}_1, r)}.$$

Since $\partial\Omega^{\mathrm{e}}$ is smooth, U_r and V_r are each open sets with C^1 boundary.

We prove the result in the first case, $\mathbf{x}_0 \in B(\mathbf{x}_1, r)$ and (4.8), and note that the proof in the second case is similar. Let $\delta > 0$ be sufficiently small. Then by [16, Theorem 9.1] (see the appendix) there exists an $\epsilon_0 > 0$ such that for a.e. $r \in (r_1 - \epsilon_0, r_1 + \epsilon_0)$

$$\begin{aligned} &\text{(o)} \quad \mathbf{u}^e|_{\partial U_r} \in W^{1,p}(\partial U_r, \mathbb{R}^m) \cap C(\partial U_r, \mathbb{R}^m);\\ &\text{(i)} \quad \mathbf{u}^e(\mathbf{x}) \in \operatorname{im}_{\mathrm{T}}(\mathbf{u}^e, U_r) \cup \mathbf{u}^e(\partial U_r) \ \text{ for a.e. } \mathbf{x} \in \overline{U_r};\\ &\text{(ii)} \quad \mathbf{u}^e(\mathbf{x}) \in \mathbb{R}^m \backslash \operatorname{im}_{\mathrm{T}}(\mathbf{u}^e, U_r) \ \text{ for a.e. } \mathbf{x} \in \Omega^{\mathrm{e}} \backslash (\overline{U_r} \cup \overline{B(\mathbf{x}_0, \delta)}\,). \end{aligned} \tag{4.9}$$

Fix one such r. For each $n \in \mathbb{N}$ we successively take $\delta = \epsilon_n$ in $(4.9)_3$. Then since the countable union of null sets is a null set, as is the set $\{\mathbf{x}_0\}$, we conclude that

$$\begin{aligned} &\text{(i)}' \quad \mathbf{u}^e(\mathbf{x}) \in \operatorname{im}_{\mathrm{T}}(\mathbf{u}^e, U_r) \cup \mathbf{u}^e(\partial U_r) \ \text{ for a.e. } \mathbf{x} \in \overline{U_r} = \overline{\Omega^{\mathrm{e}}} \backslash B_r(\mathbf{x}_1);\\ &\text{(ii)}' \quad \mathbf{u}^e(\mathbf{x}) \in \mathbb{R}^m \backslash \operatorname{im}_{\mathrm{T}}(\mathbf{u}^e, U_r) \ \text{ for a.e. } \mathbf{x} \in \Omega^{\mathrm{e}} \backslash \overline{U_r} = B_r(\mathbf{x}_1). \end{aligned} \tag{4.10}$$

In order to obtain (4.7) we first note that V_r, U_r, and ∂V_r are pairwise disjoint sets with V_r and U_r open, ∂V_r compact, and $\Omega^e = V_r \cup U_r \cup \partial V_r$. Thus standard properties of degree[10] imply

$$\deg(\mathbf{u}^e, \Omega^{\mathrm{e}}, \mathbf{y}) = \deg(\mathbf{u}^e, U_r, \mathbf{y}) + \deg(\mathbf{u}^e, V_r, \mathbf{y}) \tag{4.11}$$

for all $\mathbf{y} \notin \mathbf{u}^e(\partial\Omega^{\mathrm{e}} \cup \partial V_r)$ (note that $\partial\Omega^{\mathrm{e}} \cup \partial V_r = \partial U_r$).

Next, $\mathbf{u}^{\mathrm{h}}$ is an orientation-preserving diffeomorphism whose degree satisfies $\deg(\mathbf{u}^{\mathrm{h}}, \Omega^{\mathrm{e}}, \mathbf{y}) = 1$ if $\mathbf{y} \in \operatorname{im}_{\mathrm{T}}(\mathbf{u}^{\mathrm{h}}, \Omega^{\mathrm{e}}) = \mathbf{u}^{\mathrm{h}}(\Omega^e)$ and $\deg(\mathbf{u}^{\mathrm{h}}, \Omega^{\mathrm{e}}, \mathbf{y}) = 0$ if $\mathbf{y} \in \mathbb{R}^m \backslash \mathbf{u}^{\mathrm{h}}(\overline{\Omega}^{\mathrm{e}})$. Since $\mathbf{u}^{\mathrm{h}}$ and $\mathbf{u}^e$ assume the same boundary values on $\partial\Omega^{\mathrm{e}}$ and since the degree depends only on the boundary values, it follows that their degrees are equal:

$$\deg(\mathbf{u}^e, \Omega^{\mathrm{e}}, \mathbf{y}) = \begin{cases} 1 & \text{if } \mathbf{y} \in \mathbf{u}^{\mathrm{h}}(\Omega^{\mathrm{e}}), \\ 0 & \text{if } \mathbf{y} \in \mathbb{R}^m \backslash \mathbf{u}^{\mathrm{h}}(\overline{\Omega}^{\mathrm{e}}). \end{cases}$$

[10] These are the domain decomposition and excision properties; see, e.g., [7] or [22].

Consequently, in view of (4.11), if $\mathbf{y} \in \mathbf{u}^{\mathrm{h}}(\Omega^{\mathrm{e}})$ and $\mathbf{y} \notin \mathbf{u}^e(\partial U_r)$, then

$$\deg(\mathbf{u}^e, U_r, \mathbf{y}) + \deg(\mathbf{u}^e, V_r, \mathbf{y}) = 1. \tag{4.12}$$

We next recall that $\mathbf{x}_0 \in B_r(\mathbf{x}_1) = V_r$ and that $\mathbf{u}^e$ satisfies (INV) on $U_r = \Omega^{\mathrm{e}} \backslash \overline{V_r} \subset \Omega^{\mathrm{e}} \backslash \{\mathbf{x}_0\}$. Therefore (4.10) and the proof of [16, Theorem 9.1] yield

$$\deg(\mathbf{u}^e, U_r, \mathbf{y}) = \begin{cases} 1 & \text{if } \mathbf{y} \in \mathrm{im}_{\mathrm{T}}(\mathbf{u}^e, U_r), \\ 0 & \text{if } \mathbf{y} \in \mathbb{R}^m \backslash (\mathrm{im}_{\mathrm{T}}(\mathbf{u}^e, U_r) \cup \mathbf{u}^e(\partial U_r)), \end{cases}$$

which, together with (4.12), allows us to conclude that if $\mathbf{y} \in \mathbf{u}^{\mathrm{h}}(\Omega^{\mathrm{e}})$, then

$$\deg(\mathbf{u}^e, V_r, \mathbf{y}) = \begin{cases} 0 & \text{if } \mathbf{y} \in \mathrm{im}_{\mathrm{T}}(\mathbf{u}^e, U_r), \\ 1 & \text{if } \mathbf{y} \in \mathbb{R}^m \backslash (\mathrm{im}_{\mathrm{T}}(\mathbf{u}^e, U_r) \cup \mathbf{u}^e(\partial U_r)). \end{cases} \tag{4.13}$$

We are now ready to prove $(4.7)_1$: by (4.5) we may assume that $\mathbf{u}^e(\mathbf{x}) \in \mathbf{u}^{\mathrm{h}}(\Omega^{\mathrm{e}})$ for a.e. $\mathbf{x} \in \overline{B_r(\mathbf{x}_1)}$. By $(4.10)_2$ and $(4.13)_2$ it follows that for a.e. $\mathbf{x} \in B_r(\mathbf{x}_1)$

$$\mathbf{u}^e(\mathbf{x}) \in \mathbf{u}^e(\partial U_r) \quad \text{or} \quad \deg(\mathbf{u}^e, V_r, \mathbf{u}^e(\mathbf{x})) = 1 \ .$$

However, $\mathbf{u}^e(\partial U_r) = \mathbf{u}^e(\partial V_r) \cup \mathbf{u}^e(\partial \Omega^{\mathrm{e}})$, $\mathbf{u}^e(\partial \Omega^{\mathrm{e}}) \cap \mathbf{u}^{\mathrm{h}}(\Omega^{\mathrm{e}}) = \emptyset$, and by (4.5) we have $\mathbf{u}^e(\mathbf{x}) \in \mathbf{u}^{\mathrm{h}}(\Omega^{\mathrm{e}})$ for a.e. $\mathbf{x} \in B_r(\mathbf{x}_1)$. Therefore $(4.7)_1$ follows from the definition of the topological image (2.2). Similarly, by (4.5), $(4.10)_1$, and $(4.13)_1$ it follows that for a.e. $\mathbf{x} \in U_r = \Omega^e \backslash \overline{B_r(\mathbf{x}_1)}$

$$\mathbf{u}^e(\mathbf{x}) \in \mathbf{u}^e(\partial V_r) \quad \text{or} \quad \deg(\mathbf{u}^e, V_r, \mathbf{u}^e(\mathbf{x})) = 0,$$

and so $(4.7)_2$ follows from (2.2). ∎

Remark 4.4. The containment condition $\mathbf{u}^e(\mathbf{x}) \in \mathbf{u}^{\mathrm{h}}(\Omega^{\mathrm{e}})$ for a.e. $\mathbf{x} \in \Omega^{\mathrm{e}}$, which follows from Lemma 2.7, is crucial to the argument in the last proof.

The following lemma combined with the last two subsections will allow us to conclude that $\mathbf{u}^e$ lies in $\mathcal{A}(\mathbf{x}_0)$.

LEMMA 4.5. *Let $m = 3$, let $p > 2 = m - 1$, and suppose that $\mathbf{u}_{\epsilon_n} \in \mathcal{A}_{\epsilon_n}(\mathbf{x}_0)$ is the sequence of minimizers of the mixed displacement/traction problem. Let $\mathbf{u} \in W^{1,p}(\Omega; \mathbb{R}^m)$ be the weak limit of the sequence as given in Theorem 4.2, so that the precise representative of its homogeneous extension $(\mathbf{u}^e)^* \in W^{1,p}(\Omega^{\mathrm{e}}; \mathbb{R}^m)$ satisfies* (INV) *on Ω^{e} and*

$$\mathbf{u}_{\epsilon_n} \rightharpoonup \mathbf{u}^e \quad \text{in} \quad W^{1,p}(\Omega^{\mathrm{e}}_{\epsilon_N}; \mathbb{R}^m)$$

for any $N \in \mathbb{N}$. Then there exists $\alpha_{\mathbf{u}} \geq 0$ such that

$$\mathrm{Det}\, \nabla \mathbf{u}^e = (\det \nabla \mathbf{u}^e)\mathcal{L}^m + \alpha_{\mathbf{u}} \delta_{\mathbf{x}_0},$$

where $\delta_{\mathbf{x}_0}$ denotes the Dirac measure supported at $\mathbf{x}_0$.

Proof. We first note that $\mathbf{u}^e \in W^{1,p}(\Omega^{\mathrm{e}}; \mathbb{R}^m)$, $\det \nabla \mathbf{u}^e > 0$ a.e., and $(\mathbf{u}^e)^*$ satisfies condition (INV) on Ω^{e}. Therefore, $\mathrm{Det}\, \nabla \mathbf{u}^e$ is a Radon measure and

$$\mathrm{Det}\, \nabla \mathbf{u}^e = (\det \nabla \mathbf{u}^e)\mathcal{L}^m + \mu^s, \tag{4.14}$$

where μ^s is a (nonnegative) Radon measure that is singular with respect to $\mathcal{L}^m$ (see section 2.4).

Next, $\mathbf{u}_{\epsilon_n} \rightharpoonup \mathbf{u}^e$ in $W^{1,p}(\Omega^{\mathrm{e}}_{\epsilon_N};\mathbb{R}^m)$ for any N and $\mathbf{u}_{\epsilon_n}$ satisfies (INV) on $\Omega^{\mathrm{e}}_{\epsilon_N}$. Therefore, by Lemma 3.3 in [19], there exists a subsequence (still labeled $\mathbf{u}_{\epsilon_n}$) that satisfies

$$\mathbf{u}_{\epsilon_n} \to \mathbf{u}^e \ \text{ in } L^q_{\mathrm{loc}}(\Omega^{\mathrm{e}}_{\epsilon_N};\mathbb{R}^m) \tag{4.15}$$

for every $1 < q < \infty$. Moreover, since $p > m-1$,

$$\operatorname{adj}\nabla\mathbf{u}_{\epsilon_n} \rightharpoonup \operatorname{adj}\nabla\mathbf{u}^e \quad \text{in } L^{\frac{p}{m-1}}(\Omega^{\mathrm{e}}_{\epsilon_N};\mathbb{R}^m) \tag{4.16}$$

(see [2, Theorem 3.4]). Consequently, by (4.15) and (4.16) (see [19, Lemma 6.7]),

$$(\operatorname{adj}\nabla\mathbf{u}_{\epsilon_n})\mathbf{u}_{\epsilon_n} \rightharpoonup (\operatorname{adj}\nabla\mathbf{u}^e)\mathbf{u}^e \quad \text{in } L^1_{\mathrm{loc}}(\Omega^{\mathrm{e}}_{\epsilon_N};\mathbb{R}^m),$$

and hence, in view of (2.6),

$$(\operatorname{Det}\nabla\mathbf{u}_{\epsilon_n})(\phi) \to (\operatorname{Det}\nabla\mathbf{u}^e)(\phi) \tag{4.17}$$

for every $\phi \in C_0^\infty(\Omega^{\mathrm{e}}_{\epsilon_N})$. Since $\mathbf{u}_{\epsilon_n} \in \mathcal{A}_{\epsilon_n}(\mathbf{x}_0)$, the sequence $\mathbf{u}_{\epsilon_n}$ satisfies

$$\operatorname{Det}\nabla\mathbf{u}_{\epsilon_n} = (\det\nabla\mathbf{u}_{\epsilon_n})\mathcal{L}^m \ \text{ on } \ \Omega^{\mathrm{e}}_{\epsilon_N}. \tag{4.18}$$

Now, by (4.1), (4.3), (4.4), and hypothesis (H3), for any $N \in \mathbb{N}$ and all $n > N$

$$\int_{\Omega_{\epsilon_N}} \Gamma(\det\nabla\mathbf{u}_{\epsilon_n})\,d\mathbf{x} \le \int_\Omega W(\mathbf{x},\mathbf{A})\,d\mathbf{x} < \infty,$$

where Γ is the convex, superlinear function given in (H3). Thus, by the de la Vallée Poussin and Dunford–Pettis criteria (see, e.g., [19, Theorem 4.1]), there is a $\theta \in L^1(\Omega_{\epsilon_N})$ such that (for a subsequence) $\det\nabla\mathbf{u}_{\epsilon_n} \rightharpoonup \theta$ in $L^1(\Omega^{\mathrm{e}}_{\epsilon_N})$. Moreover, Lemma 3.2 in [19] implies that $\theta = \det\nabla\mathbf{u}^e$ a.e. in $\Omega^{\mathrm{e}}_{\epsilon_N}$. Therefore,

$$\det\nabla\mathbf{u}_{\epsilon_n} \rightharpoonup \det\nabla\mathbf{u}^e \quad \text{in } L^1(\Omega^{\mathrm{e}}_{\epsilon_N};\mathbb{R}^m), \tag{4.19}$$

and in view of (4.17)–(4.19) we find that, for every $\phi \in C_0^\infty(\Omega^{\mathrm{e}}_{\epsilon_N})$,

$$(\operatorname{Det}\nabla\mathbf{u}^e)(\phi) = \int_{\Omega^{\mathrm{e}}_{\epsilon_N}} \phi(\mathbf{x})\det\nabla\mathbf{u}^e(\mathbf{x})\,d\mathbf{x}$$

and consequently that

$$(\operatorname{Det}\nabla\mathbf{u}^e)(\Omega^{\mathrm{e}}_{\epsilon_N}) = \int_{\Omega^{\mathrm{e}}_{\epsilon_N}} \det\nabla\mathbf{u}^e(\mathbf{x})\,d\mathbf{x}. \tag{4.20}$$

Finally, (4.14) and (4.20) imply that for every $N \in \mathbb{N}$

$$\mu^s(\Omega^{\mathrm{e}}_{\epsilon_N}) = 0.$$

Thus, since

$$\Omega^{\mathrm{e}} \subset \{\mathbf{x}_0\} \cup \bigcup_{N=1}^{\infty} \Omega^{\mathrm{e}}_{\epsilon_N},$$

we find that

$$\mu^s(\Omega^{\mathrm{e}}) \le \mu^s(\{\mathbf{x}_0\}) + \sum_{N=1}^{\infty} \mu^s(\Omega^{\mathrm{e}}_{\epsilon_N}) = \mu^s(\{\mathbf{x}_0\}),$$

which yields the desired result (set $\alpha_{\mathbf{u}} = \mu^s(\{\mathbf{x}_0\})$). ∎

4.3. u is a minimizer for the pure displacement problem. Thus far we have shown that the weak limit $\mathbf{u}$ of a subsequence of the minimizers $\mathbf{u}_{\epsilon_n}$ of the mixed displacement/traction boundary value problem lies in $\mathcal{A}(\mathbf{x}_0)$. Next, we will prove that $\mathbf{u}$ is a minimizer of the pure displacement boundary value problem considered in Proposition 3.1. Throughout this section we will use $\mathbf{u}_{\epsilon_n}$ to denote this convergent subsequence identified in sections 4.1 and 4.2.

THEOREM 4.6. *Suppose that E_0 and $\mathcal{A}_0(\mathbf{x}_0)$ are as defined in section 3 and that $\mathbf{u} \in \mathcal{A}(\mathbf{x}_0)$ is the weak limit of the sequence of minimizers $(\mathbf{u}_{\epsilon_n})$ of the mixed displacement/traction problems. Then $E_0(\mathbf{u}) \leq E_0(\tilde{\mathbf{u}})$ for all $\tilde{\mathbf{u}} \in \mathcal{A}_0(\mathbf{x}_0)$.*

Proof. Define

$$\lambda := \liminf_{n\to\infty} \int_{\Omega_{\epsilon_n}} W(\mathbf{x}, \nabla \mathbf{u}_{\epsilon_n}(\mathbf{x}))\, d\mathbf{x}$$

and let N be fixed. Then, since W is nonnegative, for any $n > N$

$$\int_{\Omega_{\epsilon_N}} W(\mathbf{x}, \nabla \mathbf{u}_{\epsilon_n}(\mathbf{x}))\, d\mathbf{x} \leq \int_{\Omega_{\epsilon_n}} W(\mathbf{x}, \nabla \mathbf{u}_{\epsilon_n}(\mathbf{x}))\, d\mathbf{x}. \tag{4.21}$$

Next, by sequential weak lower semicontinuity of E_{ϵ_N} (see [3, Theorem 7.1] and [19, pp. 93–100]) and since $\mathbf{u}_{\epsilon_n} \rightharpoonup \mathbf{u}$ in $W^{1,p}(\Omega_{\epsilon_N}; \mathbb{R}^m)$ as $n \to \infty$, we have

$$\int_{\Omega_{\epsilon_N}} W(\mathbf{x}, \nabla \mathbf{u}(\mathbf{x}))\, d\mathbf{x} \leq \liminf_{n\to\infty} \int_{\Omega_{\epsilon_N}} W(\mathbf{x}, \nabla \mathbf{u}_{\epsilon_n}(\mathbf{x}))\, d\mathbf{x}. \tag{4.22}$$

If we combine (4.21) and (4.22) we find that

$$\int_{\Omega_{\epsilon_N}} W(\mathbf{x}, \nabla \mathbf{u}(\mathbf{x}))\, d\mathbf{x} \leq \lambda$$

and hence by the monotone convergence theorem that

$$\int_{\Omega} W(\mathbf{x}, \nabla \mathbf{u}(\mathbf{x}))\, d\mathbf{x} \leq \lambda. \tag{4.23}$$

Now, suppose that $\tilde{\mathbf{u}} \in \mathcal{A}_0(\mathbf{x}_0)$. Then $\tilde{\mathbf{u}}|_{\Omega_\epsilon} \in \mathcal{A}_\epsilon(\mathbf{x}_0)$ for every sufficiently small ϵ, and thus

$$\int_{\Omega_{\epsilon_n}} W(\mathbf{x}, \nabla \mathbf{u}_{\epsilon_n}(\mathbf{x}))\, d\mathbf{x} \leq \int_{\Omega_{\epsilon_n}} W(\mathbf{x}, \nabla \tilde{\mathbf{u}}(\mathbf{x}))\, d\mathbf{x}.$$

Then since W is nonnegative,

$$\int_{\Omega_{\epsilon_n}} W(\mathbf{x}, \nabla \mathbf{u}_{\epsilon_n}(\mathbf{x}))\, d\mathbf{x} \leq \int_{\Omega} W(\mathbf{x}, \nabla \tilde{\mathbf{u}}(\mathbf{x}))\, d\mathbf{x},$$

and if we take the liminf we conclude that

$$\lambda = \liminf_{n\to\infty} \int_{\Omega_{\epsilon_n}} W(\mathbf{x}, \nabla \mathbf{u}_{\epsilon_n}(\mathbf{x}))\, d\mathbf{x} \leq \int_{\Omega} W(\mathbf{x}, \nabla \tilde{\mathbf{u}}(\mathbf{x}))\, d\mathbf{x},$$

which together with (4.23) yields the desired result. ∎

We conclude from the results in section 4 that minimizers of E_{ϵ_n} on $\mathcal{A}_{\epsilon_n}(\mathbf{x}_0)$ converge to a minimizer of E_0 on $\mathcal{A}_0(\mathbf{x}_0)$ as $\epsilon_n \to 0$, i.e., that (a subsequence of) the minimizers of the mixed displacement/traction boundary value problem converge weakly to a minimizer of the pure displacement boundary value problem. This completes the proof of Theorem 4.1.

Remark 4.7. We cannot conclude in general that any sequence of minimizers $(\mathbf{u}_{\epsilon_n})$ of the mixed problem necessarily converges to a minimizer $\mathbf{u}$ of the pure displacement problem; this follows only after passing to a subsequence, since we do not know whether the minimizers given by Theorem 3.1 are unique.

5. Surface energy. In this section we examine the effect of regularizing our pure displacement variational problem by adding a surface energy term which penalizes the formation and growth of cavities. In particular, we consider the case of a typical surface energy term proportional to the surface area[11] of the holes produced. This modification is partly motivated by experimental observations of Gent [8], which indicate that such surface energy effects are relevant in certain situations involving cavitation. In this case, the original energy functional is replaced by the augmented functional:

$$\tilde{E}(\mathbf{u}) = \int_\Omega W(\nabla \mathbf{u}(\mathbf{x}))\, d\mathbf{x} \;+\; \kappa \operatorname{Per}(\operatorname{im}(\mathbf{u}, \Omega)) = E(\mathbf{u}) + \kappa \operatorname{Per}(\operatorname{im}(\mathbf{u}, \Omega)), \tag{5.1}$$

where $\kappa > 0$ is a constant and $\operatorname{Per}(\operatorname{im}(\mathbf{u}, \Omega))$ corresponds to the perimeter or surface area of the holes produced (see [16] for a precise definition of the perimeter of the measure-theoretic image of Ω under $\mathbf{u}$). Again we consider the displacement boundary value problem in which the admissible deformations are required to satisfy the boundary condition (1.3). If W satisfies the hypotheses of section 3 (with $1 \le p < 3$), then analogous arguments to those given in [21, Theorem 2] show that, for "large $\mathbf{A}$,"[12] the homogeneous deformation $\mathbf{u}^{\mathrm{h}}(\mathbf{x}) \equiv \mathbf{A}\mathbf{x}$ is no longer the global minimizer of the energy (5.1). In particular, the energy can be lowered through the introduction of holes, or cavities, in the material. However, in contrast to the bifurcation diagram (Figure 1) for the radial problem without surface energy, we will show that these discontinuous minimizers do not bifurcate[13] from the trivial homogeneous deformations (i.e., that there do not exist discontinuous minimizers producing holes of arbitrarily small volume).

Our proof will consist of a simple energy estimate; however, for ease of presentation we will restrict our attention to the following class of homogeneous energy functions (but our arguments apply to much more general[14] stored energy functions):

$$W(\mathbf{F}) = c|\mathbf{F}|^p + \Gamma(\det \mathbf{F}),$$

where $2 < p < 3$, $c > 0$, and Γ is convex and differentiable. Let $\mathbf{u} \in \tilde{\mathcal{A}}$,

$$\tilde{\mathcal{A}} = \{\mathbf{u} \in W^{1,1}(\Omega; \mathbb{R}^m) : \mathbf{u}|_{\partial\Omega} = \mathbf{u}^{\mathrm{h}},\ (\mathbf{u}^{\mathrm{e}})^* \text{ satisfies (INV) on } \Omega^{\mathrm{e}},\ \det \nabla \mathbf{u} > 0 \text{ a.e.}\},$$

[11] As argued in [16], this may be criticized on the grounds that it assigns the same energy to creating the surface of a new cavity as it does to stretching the surface of a preexisting cavity.

[12] This is the case, in particular, if $\mathbf{A} = t\mathbf{B}$ for large $t > 0$, where $\mathbf{B} \in \mathrm{M}^{3\times 3}_+$ is any fixed matrix.

[13] See also [1, p. 608], which studies the addition of surface energy for radial deformations of a ball of incompressible material.

[14] In particular, our arguments clearly extend to stored energy functions of the form $W(\mathbf{F}) = W_0(\mathbf{F}) + c|\mathbf{F}|^p + \Gamma(\det \mathbf{F})$, $2 < p < 3$, where W_0 is $W^{1,p}$-quasiconvex.

and suppose that $\tilde{E}(\mathbf{u}) < \tilde{E}(\mathbf{u}^{\mathrm{h}})$, where $\mathbf{u}^{\mathrm{h}}(\mathbf{x}) \equiv \mathbf{A}\mathbf{x}$ and $\tilde{E}$ is given by (5.1). From the definition of $\tilde{E}$ it then follows that $E(\mathbf{u}) < E(\mathbf{u}^{\mathrm{h}})$. Next, by the convexity of the mapping $\mathbf{F} \mapsto |\mathbf{F}|^p$ and the boundary condition $\mathbf{u}|_{\partial\Omega} = \mathbf{u}^{\mathrm{h}}$,

$$c \int_\Omega |\mathbf{A}|^p \, d\mathbf{x} \leq c \int_\Omega |\nabla \mathbf{u}|^p \, d\mathbf{x},$$

and hence

$$\begin{aligned} 0 < E(\mathbf{u}^{\mathrm{h}}) - E(\mathbf{u}) &= \int_\Omega c|\mathbf{A}|^p - c|\nabla \mathbf{u}|^p + \Gamma(\det \mathbf{A}) - \Gamma(\det \nabla \mathbf{u}) \, d\mathbf{x} \\ &\leq \int_\Omega \Gamma(\det \mathbf{A}) - \Gamma(\det \nabla \mathbf{u}) \, d\mathbf{x} \\ &\leq \int_\Omega \Gamma'(\det \mathbf{A})(\det \mathbf{A} - \det \nabla \mathbf{u}) \, d\mathbf{x}, \end{aligned} \tag{5.2}$$

where we have used the convexity of Γ in the last step. The above estimate will allow us to bound the decrease in the bulk energy due to the formation of a cavity by a constant times the volume of the cavity.

Suppose for a contradiction that there exists a sequence of matrices $\mathbf{A}_n \to \mathbf{A}$ and a corresponding sequence of deformations $\mathbf{u}_n \in \tilde{\mathcal{A}}$, with $\mathbf{u}_n(\mathbf{x}) = \mathbf{A}_n\mathbf{x}$ for all $\mathbf{x} \in \partial\Omega$ and

$$\operatorname{Det} \nabla \mathbf{u}_n = (\det \nabla \mathbf{u}_n)\mathcal{L}^3 + \mu_n, \tag{5.3}$$

such that $\tilde{E}(\mathbf{u}_n) < \tilde{E}(\mathbf{u}_n^{\mathrm{h}})$ for all n and $\mu_n(\Omega) \to 0$ as $n \to \infty$, where $\mathbf{u}_n^{\mathrm{h}}(\mathbf{x}) \equiv \mathbf{A}_n\mathbf{x}$. Then by (5.2), (5.1), and (5.3)

$$\begin{aligned} 0 < \tilde{E}(\mathbf{u}_n^{\mathrm{h}}) - \tilde{E}(\mathbf{u}_n) &\leq \Gamma'(\det \mathbf{A}_n) \int_\Omega (\det \mathbf{A} - \det \nabla \mathbf{u}_n) \, d\mathbf{x} - \kappa \operatorname{Per}(\operatorname{im}(\mathbf{u}_n, \Omega)) \\ &= \Gamma'(\det \mathbf{A}_n)\mu_n(\Omega) - \kappa \operatorname{Per}(\operatorname{im}(\mathbf{u}_n, \Omega)). \end{aligned} \tag{5.4}$$

The intuitive idea now is that $\mu_n(\Omega)$ represents the volume of the holes formed and $\operatorname{Per}(\operatorname{im}(\mathbf{u}_n, \Omega))$ represents the surface areas produced. Since by assumption $\mu_n(\Omega) \to 0$, the above inequality yields a contradiction for large n. Mathematically, this argument is made rigorous through the use of an isoperimetric inequality. By the definition of $\tilde{E}$ (see (5.1)) it follows that if $\tilde{E}(\mathbf{u}) < \infty$, then $\operatorname{Per}(\operatorname{im}(\mathbf{u}, \Omega)) < \infty$, and so by [16, Theorem 8.4]

$$\operatorname{Det} \nabla \mathbf{u}_n = (\det \nabla \mathbf{u}_n)\mathcal{L}^3 + \sum_{i=1}^{\infty} \alpha_i^{(n)} \delta_{\mathbf{x}_i}, \tag{5.5}$$

where $\alpha_i^{(n)} \geq 0$ for all i and n. Moreover,

$$\sum_{i=1}^{\infty} \left(\alpha_i^{(n)}\right)^{2/3} \leq \bar{c} \operatorname{Per}(\operatorname{im}(\mathbf{u}_n, \Omega)),$$

where $\bar{c} > 0$ is the isoperimetric constant.

Without loss of generality we may assume that $0 \leq \Gamma'(\det \mathbf{A}_n) \leq K$ for all n (the first inequality follows from the main result in [23] and the second by the assumed convergence of $(\mathbf{A}_n)$ to $\mathbf{A}$). Hence, by (5.4) and the above expression,

$$0 < K \sum_{i=1}^{\infty} \alpha_i^{(n)} - \frac{\kappa}{\bar{c}} \sum_{i=1}^{\infty} \left(\alpha_i^{(n)}\right)^{2/3}. \tag{5.6}$$

Now, in view of (5.5), $\mu_n(\Omega) = \sum_{i=1}^{\infty} \alpha_i^{(n)} \to 0$ as $n \to \infty$. Therefore, given $l \in \mathbb{N}$, let $N_l \in \mathbb{N}$ be such that $\alpha_i^{(n)} < \frac{1}{2^{3l}}$ for all $n > N_l$ and $i \in \mathbb{N}$. Then, by (5.6),

$$0 < \left(\frac{K}{2^l} - \frac{\kappa}{\bar{c}}\right) \sum_{i=1}^{\infty} \left(\alpha_i^{(n)}\right)^{2/3} \quad \forall n > N_l \,,$$

which yields a contradiction for l sufficiently large. □

Remark 5.1. We note that the arguments used in section 4 can be adapted to show that if $\kappa_n \to 0$ and $(\mathbf{u}_n)$ is a corresponding sequence of minimizers of $\tilde{E}_n$ (given by (5.1) with $\kappa = \kappa_n$) on $\mathcal{A}_0(\mathbf{x}_0)$ (given by (3.1)), then, passing to a subsequence if necessary, $(\mathbf{u}_n)$ converges weakly in $W^{1,p}(\Omega)$ to a minimizer of E on $\mathcal{A}_0(\mathbf{x}_0)$.

6. Concluding remarks. It is interesting to note that the convergence result given in Theorem 4.1 could form the rigorous basis for a numerical method to compute approximations to the singular minimizers whose existence is given in Proposition 3.1. In particular, it is sufficient to compute regular minimizers on Ω_ϵ whose existence is given by Proposition 3.1 for small ϵ.

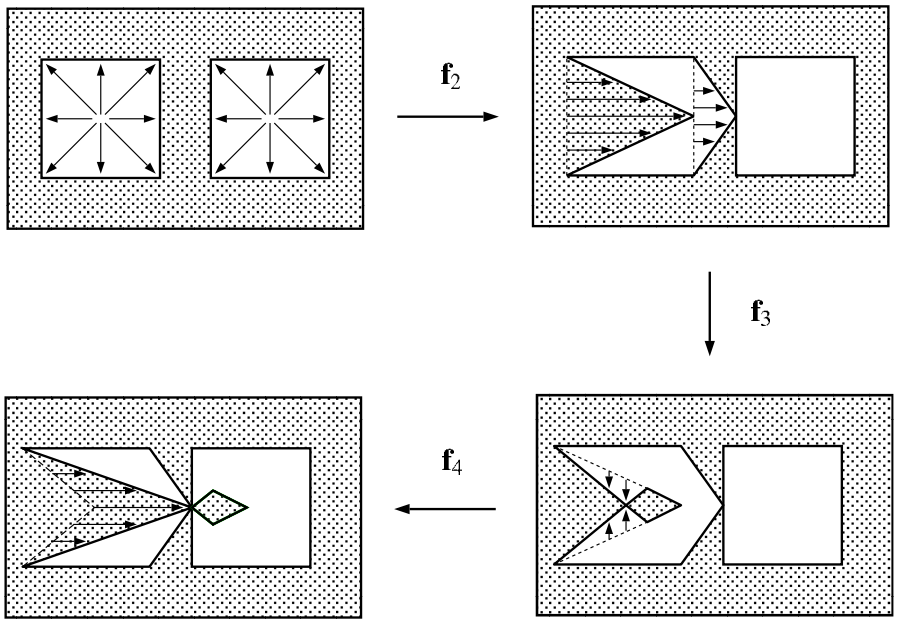


FIG. 3. *Leakage between holes.*

We next mention the main difficulty encountered in trying to extend the convergence result of Theorem 4.1 to the situation where we have, say, two (or more) flaw points located at $\{\mathbf{x}_0, \mathbf{x}_1\}$. Most of the arguments extend to this case; however, a map $\mathbf{u}$ satisfying (INV) on $\Omega \backslash \{\mathbf{x}_0, \mathbf{x}_1\}$ need not satisfy (INV) on Ω. The difficulty is related to the counterexample given in [17]: it is possible to construct a mapping $\mathbf{u} \in W^{1,p}(\Omega, \mathbb{R}^m)$ with $p > m - 1$ which satisfies (INV) on $\Omega^e \backslash \{\mathbf{x}_0, \mathbf{x}_1\}$ but which does not satisfy (INV) on Ω; consider a map which forms two adjacent cavities and which allows leakage from one cavity into the other (see Figure 3). Such a map can be produced, for example, as a composition of a two-hole cavitating map $\mathbf{f}_1$ with three Lipschitz maps $\mathbf{f}_2, \mathbf{f}_3$, and $\mathbf{f}_4$. Explicit formulae for some of the mappings can be found in [16].

Appendix. We note that our results depend crucially on Theorem 9.1 in [16], which extends condition (INV) from balls to other regions. However, the original proof of this result contains a small error. We therefore include here a corrected proof.

THEOREM A.1 (see [16, Theorem 9.1]). *Let* $\mathbf{u} \in W^{1,p}(\Omega;\mathbb{R}^m)$ *with* $p > m-1$. *Suppose that* $\det \nabla \mathbf{u} > 0$ *a.e. and that* $\mathbf{u}^*$ *satisfies condition* (INV). *Assume that* $U \subset\subset \Omega$ *is open with* C^1 *boundary and that there exists an* $\varepsilon_0 > 0$, *an open neighborhood* N *of* ∂U, *and a (surjective) diffeomorphism* $\mathbf{w} : \partial U \times (-\varepsilon_0, \varepsilon_0) \to N$ *that satisfies* $\mathbf{w}(\mathbf{x}, 0) = \mathbf{x}$ *and* $\mathbf{w}_\varepsilon(\mathbf{x}, 0) \cdot \mathbf{n}(\mathbf{x}) > 0$ *for all* $\mathbf{x} \in \partial U$, *where* $\mathbf{n}(\mathbf{x})$ *is the outward unit normal to* U *and* $\mathbf{w}_\varepsilon$ *is the partial derivative of* $\mathbf{w}$ *with respect to its second argument. Define*

$$U_\varepsilon := (U \backslash N) \cup \mathbf{w}\big(\partial U \times (-\varepsilon_0, \varepsilon)\big).$$

Then for a.e. $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$

(o) $\mathbf{u}^*|_{\partial U_\varepsilon} \in W^{1,p}(\partial U_\varepsilon;\mathbb{R}^m) \cap C^0(\partial U_\varepsilon;\mathbb{R}^m)$;

(i) $\mathbf{u}^*(\mathbf{x}) \in (\mathrm{im}_{\mathrm{T}}(\mathbf{u}^*, U_\varepsilon) \cup \mathbf{u}^*(\partial U_\varepsilon))$ *for a.e.* $\mathbf{x} \in \overline{U}_\varepsilon$;

(ii) $\mathbf{u}^*(\mathbf{x}) \in \mathbb{R}^m \backslash \mathrm{im}_{\mathrm{T}}(\mathbf{u}^*, U_\varepsilon)$ *for a.e.* $\mathbf{x} \in \Omega \backslash U_\varepsilon$.

Proof. Let $\theta \in C_0^\infty(\mathbb{R}^m)$ satisfy $\theta \geq 0$ and suppose that $\mathbf{g} \in C^\infty(\mathbb{R}^m;\mathbb{R}^m) \cap L^\infty(\mathbb{R}^m;\mathbb{R}^m)$ satisfies $\operatorname{div} \mathbf{g} = \theta$. Then, as in section 8 in [16], the map $\mu_\theta : C_0^\infty(\Omega) \to \mathbb{R}$ given by

$$\mu_\theta(\varphi) := -\int_\Omega \nabla\varphi \cdot (\operatorname{adj} \nabla \mathbf{u})(\mathbf{g} \circ \mathbf{u})\, d\mathbf{x}$$

is a distribution on Ω. If we let φ_t be a standard sequence of (radial) mollifiers, then we find that the computation that leads to equation (8.4) in [16] will now yield

$$(\varphi_t * \mu_\theta)(\mathbf{x}) = -\int_0^t \psi_t'(r) \int_{\mathbb{R}^m} (\operatorname{div} \mathbf{g}) \deg(\mathbf{u}, S(\mathbf{x}, r), \mathbf{y})\, d\mathbf{y}\, dr,$$

where $\varphi(\mathbf{x}) = \psi(|\mathbf{x}|)$ and we have written $\mathbf{u}$ for $\mathbf{u}^*$. Since $\operatorname{div} \mathbf{g} = \theta \geq 0$, the reasoning used in the proof of [16, Lemma 8.1] therefore implies that μ_θ is a Radon measure, for each θ, and that

$$\begin{aligned} \mu_\theta(\overline{B(\mathbf{b}, r)}) &= \int_{\mathbb{R}^m} \theta(\mathbf{y}) \deg(\mathbf{u}, S(\mathbf{b}, r), \mathbf{y})\, d\mathbf{y} \\ &= \int_{\mathbb{R}^m} \theta \chi_{\mathrm{im}_{\mathrm{T}}(\mathbf{u}, B(\mathbf{b}, r))}\, d\mathbf{y} \end{aligned} \tag{A.1}$$

for $\mathcal{L}^1$ a.e. $r \in (0, r_{\mathbf{b}})$.

We next show that for $\mathcal{L}^1$ a.e. $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$

$$\mu_\theta(U_\varepsilon) = \int_{\mathbb{R}^m} \theta(\mathbf{y}) \deg(\mathbf{u}, \partial U_\varepsilon, \mathbf{y})\, d\mathbf{y}. \tag{A.2}$$

Let $\psi \in C^\infty(\mathbb{R})$ satisfy $\psi(s) = 1$ for $s < -\varepsilon_0/2$ and $\psi(s) = 0$ for $s > \varepsilon_0/2$ and define $\varphi := \psi \circ \omega$, where the function $\omega : N \to (-\varepsilon_0, \varepsilon_0)$ denotes the last component of the diffeomorphism $\mathbf{w}^{-1}$. We note that $\nu := \nabla\omega/|\nabla\omega|$ is the outward unit normal to the surfaces ∂U_ε and apply the coarea formula for the C^1 function ω to get

$$\begin{aligned} \mu_\theta(\varphi) &= -\int_\Omega (\psi' \circ \omega)|\nabla\omega| \nu \cdot (\operatorname{adj} \nabla \mathbf{u})(\mathbf{g} \circ \mathbf{u})\, d\mathbf{x} \\ &= \int_{-\varepsilon_0}^{\varepsilon_0} \psi'(s) \int_{\omega = s} (\operatorname{adj} \nabla \mathbf{u})^T \nu \cdot (\mathbf{g} \circ \mathbf{u})\, d\mathcal{H}^{m-1}\, ds. \end{aligned}$$

An application of the area formula together with the fact that $\mathbf{g}$ is bounded shows that the inner integral is an $\mathcal{L}^1$ function in the variable s. If we choose a suitable increasing sequence $\psi_k \nearrow \chi_{(-\infty,\varepsilon)}$, then we find, with the aid of Proposition 2.1 in [16], that (A.2) is satisfied.

We note that it follows from (A.1) and (8.3) in [16] that

$$\mu_\theta(\overline{B(\mathbf{b},r)}) \le (\sup\theta)(\operatorname{Det}\nabla\mathbf{u})(\overline{B(\mathbf{b},r)})$$

for $\mathcal{L}^1$ a.e. $r \in (0, r_{\mathbf{b}})$. This implies that each of the measures μ_θ is absolutely continuous with respect to $\operatorname{Det}\nabla\mathbf{u}$.

Next, let $\Theta \subset C_0^\infty(\mathbb{R}^m;\mathbb{R}^{\geq})$ be a countable set that is dense in $L^2(\mathbb{R}^m;\mathbb{R}^{\geq})$ and fix $\varepsilon \in (-\varepsilon_0,\varepsilon_0)$ so that (o) is satisfied and (A.2) is satisfied for every $\theta \in \Theta$. Let $N_{\mathbf{b}}$ be the $\mathcal{L}^1$ null set of Lemma 7.3 in [16] and let $\hat{N}_{\mathbf{b}}$ be an $\mathcal{L}^1$ null set such that (A.1) is satisfied for every $\theta \in \Theta$ and every $r \in (0, r_{\mathbf{b}})\backslash\hat{N}_b$. Writing $\tilde{N}_{\mathbf{b}} = N_{\mathbf{b}} \cup \hat{N}_{\mathbf{b}}$ we consider the family of closed balls

$$\mathcal{F} := \{\overline{B(\mathbf{b},r)} : \mathbf{b} \in U_\varepsilon,\ r \in (0,r_{\mathbf{b}})\backslash\tilde{N}_{\mathbf{b}},\ (\operatorname{Det}\nabla\mathbf{u})(\partial B(\mathbf{b},r)) = 0\}.$$

Then, since the set of radii for which $(\operatorname{Det}\nabla\mathbf{u})(\partial B(\mathbf{b},r)) > 0$ is at most countable, for each $\mathbf{b} \in U_\varepsilon$

$$\inf\{r : \overline{B(\mathbf{b},r)} \in \mathcal{F}\} = 0,$$

and hence we can apply the Besicovitch covering theorem to get a sequence of pairwise disjoint closed balls $\overline{B(\mathbf{b}_k,r_k)} \subset U_\varepsilon$ such that

$$(\mathcal{L}^m + \operatorname{Det}\nabla\mathbf{u})\left(U_\varepsilon \setminus \bigcup_{k=1}^{\infty} \overline{B(\mathbf{b}_k,r_k)}\right) = 0. \tag{A.3}$$

Therefore, since the sets $\overline{B(\mathbf{b}_k,r_k)}$ are pairwise disjoint, (A.1)–(A.3) together with the absolute continuity of each measure μ_θ with respect to $\operatorname{Det}\nabla\mathbf{u}$ imply that

$$\int_{\mathbb{R}^m} \theta(\mathbf{y}) \deg(\mathbf{u},\partial U_\varepsilon,\mathbf{y})\, d\mathbf{y} = \sum_{k=1}^{\infty} \int_{\mathbb{R}^m} \theta\chi_{\operatorname{im_T}(\mathbf{u},B(\mathbf{b}_k,r_k))}\, d\mathbf{y}$$

for every $\theta \in \Theta$. Since Θ is dense in $\mathcal{L}^2(\mathbb{R}^m;\mathbb{R}^{\geq})$ and since the sets $\operatorname{im_T}(\mathbf{u},B(\mathbf{b}_k,r_k))$ are pairwise disjoint, the bounded convergence theorem and the last equation yield

$$\deg(\mathbf{u},\partial U_\varepsilon,\cdot\,) = \sum_{k=1}^{\infty} \chi_{\operatorname{im_T}(\mathbf{u},B(\mathbf{b}_k,r_k))} \text{ a.e.}$$

Consequently, since the sets $\operatorname{im_T}(\mathbf{u},B(\mathbf{b}_k,r_k))$ are pairwise disjoint, $\deg(\mathbf{u},\partial U_\epsilon,\mathbf{y})$ assumes only the values 0 and 1, and hence $\deg(\mathbf{u},\partial U_\epsilon,\cdot\,) = \chi_{\operatorname{im_T}(\mathbf{u},U_\epsilon)}$: thus,

$$\chi_{\operatorname{im_T}(\mathbf{u},U_\varepsilon)} = \sum_{k=1}^{\infty} \chi_{\operatorname{im_T}(\mathbf{u},B(\mathbf{b}_k,r_k))} \text{ a.e.} \tag{A.4}$$

We note that, by the area formula, the sets $\mathbf{u}(\partial B(\mathbf{b}_k,r_k))$ and $\mathbf{u}(\partial U_\varepsilon)$ are each Lebesgue null sets. Consequently, since the sets $\operatorname{im_T}(\mathbf{u},B(\mathbf{b}_k,r_k))$ are pairwise disjoint, (A.4) implies that there exist Lebesgue null sets M_1 and M_2 such that

$$\bigcup_{k=1}^{\infty} E(\mathbf{u},B(\mathbf{b}_k,r_k)) \subset M_1 \cup E(\mathbf{u},U_\varepsilon) \tag{A.5}$$

and

$$\text{im}_{\text{T}}(\mathbf{u}, U_\varepsilon) \subset M_2 \cup \bigcup_{k=1}^{\infty} \text{im}_{\text{T}}(\mathbf{u}, B(\mathbf{b}_k, r_k)). \tag{A.6}$$

Finally, since $\det \nabla \mathbf{u} > 0$ a.e., it follows from the area formula (see, e.g., [24, Lemma 2]) that $\mathbf{u}^{-1}(M_1)$ and $\mathbf{u}^{-1}(M_2)$ are Lebesgue null sets. Therefore (i) of this theorem follows from (A.5) and (i) of condition (INV), while (A.6) and (ii) of condition (INV) yield (ii) of this theorem. □

Acknowledgment. J.S. would like to thank Geoffrey Burton for helpful discussions in the course of this work.

REFERENCES

[1] J. M. Ball, *Discontinuous equilibrium solutions and cavitation in nonlinear elasticity*, Philos. Trans. Roy. Soc. London Ser. A, 306 (1982), pp. 557–611.
[2] J. M. Ball, *Constitutive inequalities and existence theorems in nonlinear elastostatics*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, Vol. 1, R. J. Knops, ed., Pitman, London, 1977, pp. 187–241.
[3] J. M. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 63 (1977), pp. 337–403.
[4] S. Conti and C. De Lellis, *Some remarks on the theory of elasticity for compressible Neohookean materials*, Ann. Sc. Norm. Super. Pisa Cl. Sci. (5), 2 (2003), pp. 521–549.
[5] J. Dollhofer, A. Chiche, V. Muralidharan, C. Creton, and C. Y. Hui, *Surface energy effects for cavity growth and nucleation in an incompressible neoHookean material-modeling and experiment*, Internat. J. Solids Structures, 41 (2004), pp. 6111–6127.
[6] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.
[7] I. Fonseca and W. Gangbo, *Degree Theory in Analysis and Applications*, Oxford University Press, New York, 1995.
[8] A. N. Gent, *Cavitation in rubber: A cautionary tale*, Rubber Chem. Tech., 63 (1991), pp. G49–G53.
[9] A. N. Gent and P. B. Lindley, *Internal rupture of bonded rubber cylinders in tension*, Proc. Roy. Soc. London Ser. A, 249 (1958), pp. 195–205.
[10] R. J. Hill, *The Mathematical Theory of Plasticity*, Clarendon Press, Oxford, UK, 1950.
[11] C. O. Horgan and R. Abeyaratne, *A bifurcation problem for a compressible nonlinearly elastic medium: Growth of a microvoid*, J. Elasticity, 16 (1986), pp. 189–200.
[12] C. O. Horgan and D. A. Polignone, *Cavitation in nonlinearly elastic solids: A review*, Appl. Mech. Rev., 48 (1995), pp. 471–485.
[13] C. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, New York, 1966.
[14] S. Müller, *A remark on the distributional determinant*, C. R. Acad. Sci. Paris Sér. I Math., 311 (1990), pp. 13–17.
[15] S. Müller, *On the singular support of the distributional determinant*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 657–696.
[16] S. Müller and S. J. Spector, *An existence theory for nonlinear elasticity that allows for cavitation*, Arch. Rational Mech. Anal., 131 (1995), pp. 1–66.
[17] S. Müller, S. J. Spector, and Q. Tang, *Invertibility and a topological property of Sobolev maps*, SIAM J. Math. Anal., 27 (1996), pp. 959–976.
[18] J. Sivaloganathan, *Uniqueness of regular and singular equilibria for spherically symmetric problems of nonlinear elasticity*, Arch. Rational Mech. Anal., 96 (1986), pp. 97–136.
[19] J. Sivaloganathan and S. J. Spector, *On the existence of minimizers with prescribed singular points in nonlinear elasticity*, J. Elasticity, 59 (2000), pp. 83–113.
[20] J. Sivaloganathan and S. J. Spector, *On the optimal location of singularities arising in variational problems of nonlinear elasticity*, J. Elasticity, 58 (2000), pp. 191–224.
[21] J. Sivaloganathan and S. J. Spector, *A variational approach to modelling initiation of fracture in nonlinear elasticity*, in IUTAM Symposium on Asymptotics, Singularities, and Homogenisation in Problems of Mechanics, A. B. Movchan, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003, pp. 295–306.

[22] J. T. Schwartz, *Nonlinear Functional Analysis*, Gordon and Breach, New York, London, Paris, 1969.
[23] S. J. Spector, *Linear deformations as global minimizers in nonlinear elasticity*, Quart. Appl. Math., 52 (1994), pp. 59–64.
[24] V. Šverák, *Regularity properties of deformations with finite energy*, Arch. Rational Mech. Anal., 100 (1988), pp. 105–127.
[25] D. Swanson and W. P. Ziemer, *The image of a weakly differentiable mapping*, SIAM J. Math. Anal., 35 (2004), pp. 1099–1109.
[26] D. Swanson and W. P. Ziemer, *A topological aspect of Sobolev mappings*, Calc. Var. Partial Differential Equations, 14 (2002), pp. 69–84.

SIAM J. APPL. MATH.
Vol. 66, No. 3, pp. 758–786

DYNAMICS OF ROTATING BOSE–EINSTEIN CONDENSATES AND ITS EFFICIENT AND ACCURATE NUMERICAL COMPUTATION *

WEIZHU BAO†, QIANG DU‡, AND YANZHI ZHANG§

Abstract. In this paper, we study the dynamics of rotating Bose–Einstein condensates (BEC) based on the Gross–Pitaevskii equation (GPE) with an angular momentum rotation term and present an efficient and accurate algorithm for numerical simulations. We examine the conservation of the angular momentum expectation and the condensate width and analyze the dynamics of a stationary state with a shift in its center. By formulating the equation in either the two-dimensional polar coordinate system or the three-dimensional cylindrical coordinate system, the angular momentum rotation term becomes a term with constant coefficients. This allows us to develop an efficient time-splitting method which is time reversible, unconditionally stable, efficient, and accurate for the problem. Moreover, it conserves the position density. We also apply the numerical method to study issues such as the stability of central vortex states and the quantized vortex lattice dynamics in rotating BEC.

Key words. rotating Bose–Einstein condensation, Gross–Pitaevskii equation, angular momentum rotation, time-splitting, ground state, central vortex state, energy, condensate width, angular momentum expectation

AMS subject classifications. 35Q55, 65T99, 65Z05, 65N12, 65N35, 81-08

DOI. 10.1137/050629392

1. Introduction.

Since its realization in dilute bosonic atomic gases [3, 20, 21], Bose–Einstein condensation of alkali atoms and hydrogen has been produced and studied extensively in the laboratory [45] and has permitted an intriguing glimpse into the macroscopic quantum world. In view of potential applications [26, 43, 44], the study of quantized vortices, which are well-known signatures of superfluidity, is one of the key issues. Different research groups have obtained quantized vortices in Bose–Einstein condensates (BEC) experimentally, e.g., the JILA group [39], the ENS group [36, 37], and the MIT group [45]. Currently, there are at least two typical ways to generate quantized vortices from the ground state of BEC: (i) impose a laser beam rotating with an angular velocity on the magnetic trap holding the atoms to create an harmonic anisotropic potential [17, 33, 1, 13]; (ii) add to the stationary magnetic trap a narrow, moving Gaussian potential, representing a far-blue detuned laser [29, 7]. The recent experimental and theoretical advances in the exploration of quantized vortices in BEC have spurred great excitement in the atomic physics community and renewed interest in studying superfluidity.

The properties of BEC in a rotational frame at temperature T much smaller than the critical condensation temperature T_c are well described by the macroscopic

*Received by the editors April 18, 2005; accepted for publication (in revised form) September 12, 2005; published electronically February 3, 2006. This work was partially done while the first two authors were visiting the Institute for Mathematical Sciences, National University of Singapore, in 2005. The visit was supported by the Institute.

http://www.siam.org/journals/siap/66-3/62939.html

†Department of Mathematics, National University of Singapore, 117543 Singapore (bao@cz3.nus.edu.sg, http://www.cz3.nus.edu.sg/∼bao/). The research of this author was supported by the National University of Singapore grant R-151-000-035-112.

‡Department of Mathematics, Penn State University, University Park, PA 16802 (qdu@math.psu.edu, http://www.math.psu.edu/qdu). The research of this author was supported by NSF DMS 0409297 and ITR 0205232.

§Department of Mathematics, National University of Singapore, 117543 Singapore (zhyanzhi@cz3.nus.edu.sg).

wave function $\psi(\mathbf{x},t)$, whose evolution is governed by a self-consistent, mean field nonlinear Schrödinger equation (NLSE) in a rotational frame, also known as the Gross–Pitaevskii equation (GPE) with an angular momentum rotation term [26, 17, 24, 25, 15]:

(1.1)
$$i\hbar\partial_t\psi(\mathbf{x},t) = \left(-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{x}) + NU_0|\psi|^2 - \Omega L_z\right)\psi(\mathbf{x},t), \quad \mathbf{x}\in\mathbb{R}^3, \quad t\geq 0,$$

where $\mathbf{x} = (x,y,z)^T$ is the Cartesian coordinate vector, m is the atomic mass, $\hbar$ is the Planck constant, N is the number of atoms in the condensate, Ω is the angular velocity of the rotating laser beam, and $V(\mathbf{x})$ is an external trapping potential. When an harmonic trap potential is considered, $V(\mathbf{x}) = \frac{m}{2}\left(\omega_x^2x^2+\omega_y^2y^2+\omega_z^2z^2\right)$ with ω_x, ω_y, and ω_z being the trap frequencies in the x-, y-, and z-direction, respectively. $U_0 = \frac{4\pi\hbar^2a_s}{m}$ describes the interaction between atoms in the condensate with a_s (positive for repulsive interaction and negative for attractive interaction) the s-wave scattering length, and $L_z = xp_y - yp_x = -i\hbar\left(x\partial_y - y\partial_x\right)$ is the z-component of the angular momentum $\mathbf{L} = \mathbf{x}\times\mathbf{P}$ with the momentum operator $\mathbf{P} = -i\hbar\nabla = (p_x,p_y,p_z)^T$. It is convenient to normalize the wave function by requiring that

(1.2)
$$\|\psi(\cdot,t)\|^2 := \int_{\mathbb{R}^3}|\psi(\mathbf{x},t)|^2d\mathbf{x} = 1.$$

Under such a normalization, we introduce the dimensionless variables as follows: $t\to t/\omega_m$ with $\omega_m = \min\{\omega_x,\omega_y,\omega_z\}$, $\Omega\to\omega_m\Omega$, $\mathbf{x}\to a_0\mathbf{x}$ with $a_0 = \sqrt{\frac{\hbar}{m\omega_m}}$, and $\psi\to\psi/a_0^{3/2}$. We also let

$$\gamma_x = \frac{\omega_x}{\omega_m}, \quad \gamma_y = \frac{\omega_y}{\omega_m}, \quad \gamma_z = \frac{\omega_z}{\omega_m}, \quad \beta = \frac{U_0N}{a_0^3\hbar\omega_m} = \frac{4\pi a_sN}{a_0}.$$

The dimensionless angular momentum rotational term then becomes

(1.3)
$$L_z = -i(x\partial_y - y\partial_x) = i(y\partial_x - x\partial_y) = -i\partial_\theta$$

with (r,θ) being the polar coordinates in two dimensions (2D) and (r,θ,z) the cylindrical coordinates in three dimensions (3D). In the disk-shaped condensation, i.e., $\omega_y\approx\omega_x$ and $\omega_z\gg\omega_x$ ($\Leftrightarrow\gamma_x = 1$, $\gamma_y\approx 1$, and $\gamma_z\gg 1$ with choosing $\omega_m = \omega_x$), the three-dimensional GPE can be reduced to a two-dimensional GPE [13]. Thus, here we consider the dimensionless GPE with a rotational term in the d-dimensions ($d = 2,3$) [13]:

(1.4)
$$i\partial_t\psi(\mathbf{x},t) = -\frac{1}{2}\nabla^2\psi + V_d(\mathbf{x})\psi + \beta_d|\psi|^2\psi - \Omega L_z\psi, \quad \mathbf{x}\in\mathbb{R}^d, \quad t>0,$$

(1.5)
$$\psi(\mathbf{x},0) = \psi_0(\mathbf{x}), \quad \mathbf{x}\in\mathbb{R}^d, \quad \text{with} \quad \|\psi_0\|^2 := \int_{\mathbb{R}^d}|\psi_0(\mathbf{x})|^2d\mathbf{x} = 1,$$

where

(1.6)
$$\beta_d = \begin{cases}\beta\sqrt{\gamma_z/2\pi}, & \\ \beta, & \end{cases} \qquad V_d(\mathbf{x}) = \begin{cases}(\gamma_x^2x^2+\gamma_y^2y^2)/2, & d=2,\\ (\gamma_x^2x^2+\gamma_y^2y^2+\gamma_z^2z^2)/2, & d=3,\end{cases}$$

with $\gamma_x > 0$, $\gamma_y > 0$, and $\gamma_z > 0$ being constants. Two important invariants of (1.4) are the *normalization of the wave function*

$$N(\psi) = \int_{\mathbb{R}^d} |\psi(\mathbf{x},t)|^2 d\mathbf{x} \equiv \int_{\mathbb{R}^d} |\psi(\mathbf{x},0)|^2 d\mathbf{x} = N(\psi_0) = 1, \qquad t \geq 0, \tag{1.7}$$

and the *energy*

$$\begin{aligned} E_{\beta,\Omega}(\psi) &= \int_{\mathbb{R}^d} \left[\frac{1}{2}|\nabla\psi|^2 + V_d(\mathbf{x})|\psi|^2 + \frac{\beta_d}{2}|\psi|^4 - \Omega \mathrm{Re}(\psi^* L_z \psi) \right] d\mathbf{x} \\ &\equiv E_{\beta,\Omega}(\psi_0), \qquad t \geq 0, \end{aligned} \tag{1.8}$$

where f^* and $\mathrm{Re}f$ denote the conjugate and the real part of the function f, respectively.

In order to study effectively the dynamics of BEC, especially in the strong repulsive interaction regime, i.e., $\beta_d \gg 1$ in (1.4), an efficient and accurate numerical method is one of the key issues. For nonrotating BEC, i.e., $\Omega = 0$ in (1.4), many numerical methods were proposed in the literature. For example, Bao, Jaksch, and Markowich [7], Bao and Jaksch [6], and Bao and Zhang [14] proposed a fourth-order time-splitting sine or Fourier pseudospectral (TSSP) method, and Bao and Shen [11] presented a fourth-order time-splitting Laguerre–Hermite (TSLH) pseudospectral method for the GPE when the external trapping potential is radially or cylindrically symmetric in 2D or 3D. The key ideas for the numerical methods in [7, 4, 6, 14, 11, 9, 10] are based on (i) a time-splitting technique being applied to decouple the nonlinearity in the GPE [7, 6, 9, 10]; (ii) proper spectral basis functions being chosen for a linear Schrödinger equation with a potential such that the ODE system in phase space is diagonalized and thus can be integrated exactly [14, 11]. These methods are explicit, unconditionally stable, and of spectral accuracy in space and fourth-order accuracy in time. Thus they are very efficient and accurate for computing the dynamics of nonrotating BEC in 3D [8] and for multicomponent [4]. Some other numerical methods for nonrotating BEC include the finite difference method [18, 41, 40], the particle-inspired scheme [19, 40], and the Runge–Kutta pseudospectral method [16, 40]. Due to the appearance of the angular momentum rotation term in the GPE (1.4), the TSSP and TSLH methods proposed in [7, 14, 11] can no longer be used for rotating BEC. Currently, the numerical methods proposed in the literature for studying the dynamics of rotating BEC remain limited [1, 22, 33], and they usually are low-order methods. Thus it is of great interest to develop an efficient, accurate, and unconditionally stable numerical method for the GPE (1.4) with an angular momentum rotation term. Such a numerical method is proposed here and is applied to the study of the dynamics of the rotating BEC. The key features of our numerical method are based on (i) the application of a time-splitting technique for decoupling the nonlinearity in the GPE; (ii) the adoption of polar coordinates or cylindrical coordinates so as to make the coefficient of the angular momentum rotation term constant; (iii) the utilization of Fourier pseudospectral discretization in the transverse direction and a second- or fourth-order finite difference or finite element discretization in the radial direction. Our extensive numerical results demonstrate that the method is very efficient and accurate.

The paper is organized as follows. In section 2, the conservation of the angular momentum expectation and the dynamics of condensate widths are first established. We then analyze the stationary state with a shift in its center and provide some study on the decrease of the total density in the presence of dissipation. In section 3, a

numerical method is presented for the efficient and accurate simulation of GPE (1.4) in 2D and 3D. It is then applied to study the vortex state and the dynamics of rotating BEC in section 4. Finally, some conclusions are drawn in section 5.

2. Dynamics of rotating BEC. In this section, we provide some analytical results on the conservation of the angular momentum expectation in a symmetric trap, i.e., $\gamma_x = \gamma_y$ in (1.6), derive a second-order ODE for time evolution of the condensate width, and present some dynamic laws of a stationary state with a shifted center in rotating BEC.

2.1. Conservation of angular momentum expectation. As a measure of the vortex flux, we define the angular momentum expectation:

$$\langle L_z\rangle(t) := \int_{\mathbb{R}^d} \psi^*(\mathbf{x},t) L_z \psi(\mathbf{x},t)\, d\mathbf{x} = i\int_{\mathbb{R}^d} \psi^*(\mathbf{x},t)(y\partial_x - x\partial_y)\psi(\mathbf{x},t) d\mathbf{x} \tag{2.1}$$

for any $t \geq 0$. For the dynamics of angular momentum expectation in rotating BEC, we have the following lemma.

LEMMA 2.1. *Suppose $\psi(\mathbf{x},t)$ is the solution of the problem* (1.4)–(1.5)*; then we have*

$$\frac{d\langle L_z\rangle(t)}{dt} = \left(\gamma_x^2 - \gamma_y^2\right)\delta_{xy}(t), \quad \textit{where } \delta_{xy}(t) = \int_{\mathbb{R}^d} xy|\psi(\mathbf{x},t)|^2 d\mathbf{x}, \quad t \geq 0. \tag{2.2}$$

Consequently, the angular momentum expectation and energy for the nonrotating part are conserved; that is, for any given initial data $\psi_0(\mathbf{x})$ in (1.5)*,*

$$\langle L_z\rangle(t) \equiv \langle L_z\rangle(0), \quad E_{\beta,0}(\psi) \equiv E_{\beta,0}(\psi_0), \quad t \geq 0, \tag{2.3}$$

at least for radially symmetric trap in 2D or cylindrically symmetric trap in 3D, i.e., $\gamma_x = \gamma_y$.

Proof. Differentiating (2.1) with respect to t, noticing (1.4), integrating by parts, and taking into account that ψ decreases to 0 exponentially when $|\mathbf{x}| \to \infty$, we have

$$\begin{aligned}
&\frac{d\langle L_z\rangle(t)}{dt} \\
&= i\int_{\mathbb{R}^d} \left[\psi_t^*(y\partial_x - x\partial_y)\psi + \psi^*(y\partial_x - x\partial_y)\psi_t\right] d\mathbf{x} \\
&= \int_{\mathbb{R}^d} \left[(-i\psi_t^*)\,(x\partial_y - y\partial_x)\psi + (i\psi_t)\,(x\partial_y - y\partial_x)\psi^*\right] d\mathbf{x} \\
&= \int_{\mathbb{R}^d} \Bigg[\left(-\frac{1}{2}\nabla^2\psi^* + V_d(\mathbf{x})\psi^* + \beta_d|\psi|^2\psi^* - i\Omega(x\partial_y - y\partial_x)\psi^*\right)(x\partial_y - y\partial_x)\psi \\
&\qquad + \left(-\frac{1}{2}\nabla^2\psi + V_d(\mathbf{x})\psi + \beta_d|\psi|^2\psi + i\Omega(x\partial_y - y\partial_x)\psi\right)(x\partial_y - y\partial_x)\psi^*\Bigg] d\mathbf{x} \\
&= \int_{\mathbb{R}^d} \Bigg[-\frac{1}{2}\left[\nabla^2\psi^*(x\partial_y - y\partial_x)\psi + \nabla^2\psi(x\partial_y - y\partial_x)\psi^*\right] \\
&\qquad + \left(V_d(\mathbf{x}) + \beta_d|\psi|^2\right)\left[\psi^*(x\partial_y - y\partial_x)\psi + \psi(x\partial_y - y\partial_x)\psi^*\right]\Bigg] d\mathbf{x} \\
&= -\int_{\mathbb{R}^d} |\psi|^2(x\partial_y - y\partial_x)V_d(\mathbf{x}) d\mathbf{x} = \left(\gamma_x^2 - \gamma_y^2\right)\int_{\mathbb{R}^d} xy|\psi|^2 d\mathbf{x}, \quad t \geq 0,
\end{aligned} \tag{2.4}$$

which gives (2.2). As for the conservation properties, since $\gamma_x = \gamma_y$, (2.2) reduces to the first-order ODE:

$$\frac{d\langle L_z\rangle(t)}{dt} = 0, \qquad t \geq 0. \tag{2.5}$$

We thus get the conservation of $\langle L_z\rangle$ immediately.

Noticing $E_{\beta,\Omega}(\psi) = E_{\beta,0}(\psi) - \Omega \mathrm{Re}\langle L_z\rangle$ and $\mathrm{Re}\langle L_z\rangle = \langle L_z\rangle$, we get (2.3) from (2.5) and (1.8). ∎

2.2. Dynamics of condensate widths. Another quantity characterizing the dynamics of rotating BEC is the condensate width defined as

$$\sigma_\alpha(t) = \sqrt{\delta_\alpha(t)}, \quad \text{where } \delta_\alpha(t) = \langle \alpha^2\rangle(t) = \int_{\mathbb{R}^d} \alpha^2 |\psi(\mathbf{x},t)|^2 d\mathbf{x} \tag{2.6}$$

for $t \geq 0$ and α being x, y, or z. For the dynamics of condensate widths, we have the following lemmas.

LEMMA 2.2. *Suppose $\psi(\mathbf{x},t)$ is the solution of problem (1.4)–(1.5); then we have*

$$\begin{aligned}\frac{d^2\delta_\alpha(t)}{dt^2} = \int_{\mathbb{R}^d} \Big[&(\partial_y \alpha - \partial_x \alpha)\left(4i\Omega\psi^*(x\partial_y + y\partial_x)\psi + 2\Omega^2(x^2-y^2)|\psi|^2\right) \\ &+ 2|\partial_\alpha \psi|^2 + \beta_d |\psi|^4 - 2\alpha|\psi|^2 \partial_\alpha(V_d(\mathbf{x}))\Big] d\mathbf{x}, \qquad t \geq 0,\end{aligned} \tag{2.7}$$

$$\delta_\alpha(0) = \delta_\alpha^{(0)} = \int_{\mathbb{R}^d} \alpha^2 |\psi_0(\mathbf{x})|^2 d\mathbf{x}, \qquad \alpha = x, y, z, \tag{2.8}$$

$$\dot{\delta}_\alpha(0) = \delta_\alpha^{(1)} = 2\int_{\mathbb{R}^d} \alpha \left[-\Omega|\psi_0|^2 (x\partial_y - y\partial_x)\alpha + \mathrm{Im}\left(\psi_0^* \partial_\alpha \psi_0\right)\right] d\mathbf{x}, \tag{2.9}$$

where $\mathrm{Im}(f)$ *denotes the imaginary part of* f.

Proof. Differentiating (2.6) with respect to t, applying (1.4), and integrating by parts, we obtain

$$\begin{aligned}\frac{d\delta_\alpha(t)}{dt} &= \frac{d}{dt}\int_{\mathbb{R}^d} \alpha^2 |\psi(\mathbf{x},t)|^2 d\mathbf{x} = \int_{\mathbb{R}^d} \alpha^2 \left(\psi \partial_t \psi^* + \psi^* \partial_t \psi\right) d\mathbf{x} \\ &= \int_{\mathbb{R}^d} \left[\frac{i}{2}\alpha^2 \left(\psi^* \nabla^2 \psi - \psi \nabla^2 \psi^*\right) + \Omega \alpha^2 \left(x\partial_y - y\partial_x\right)|\psi|^2\right] d\mathbf{x} \\ &= \int_{\mathbb{R}^d} \left[i\alpha \left(\psi \partial_\alpha \psi^* - \psi^* \partial_\alpha \psi\right) - 2\Omega\alpha|\psi|^2 \left(x\partial_y - y\partial_x\right)\alpha\right] d\mathbf{x}.\end{aligned} \tag{2.10}$$

Differentiating the above equation again, applying (1.4), and integrating by parts, we get

$$\begin{aligned}&\frac{d^2\delta_\alpha(t)}{dt^2} \\ &= \int_{\mathbb{R}^d} \Big[i\alpha \left(\partial_t \psi \partial_\alpha \psi^* + \psi \partial_{\alpha t} \psi^* - \partial_t \psi^* \partial_\alpha \psi - \psi^* \partial_{\alpha t} \psi\right) \\ &\qquad - 2\Omega\alpha \left(\psi \partial_t \psi^* + \psi^* \partial_t \psi\right)\left(x\partial_y - y\partial_x\right)\alpha\Big] d\mathbf{x} \\ &= \int_{\mathbb{R}^d} \Big[2i\alpha \left(\partial_t \psi \partial_\alpha \psi^* - \partial_t \psi^* \partial_\alpha \psi\right) + i\left(\psi^* \partial_t \psi - \psi \partial_t \psi^*\right) \\ &\qquad - 2\Omega\alpha (x\partial_y - y\partial_x)\alpha \left(\frac{i}{2}\left(\psi^* \nabla^2 \psi - \psi \nabla^2 \psi^*\right) + \Omega (x\partial_y - y\partial_x)|\psi|^2\right)\Big] d\mathbf{x},\end{aligned} \tag{2.11}$$

$$
\begin{aligned}
&\frac{d^2\delta_\alpha(t)}{dt^2}\\
&= \int_{\mathbb{R}^d}\Bigg[-\alpha\left(\partial_\alpha\psi^*\nabla^2\psi + \partial_\alpha\psi\nabla^2\psi^*\right) + 2\alpha\left(V_d(\mathbf{x}) + \beta_d|\psi|^2\right)\left(\psi\partial_\alpha\psi^* + \psi^*\partial_\alpha\psi\right)\\
&\qquad -2i\Omega\alpha\left[\partial_\alpha\psi\left(x\partial_y - y\partial_x\right)\psi^* - \partial_\alpha\psi^*\left(x\partial_y - y\partial_x\right)\psi\right] - \frac{1}{2}\left(\psi^*\nabla^2\psi + \psi\nabla^2\psi^*\right)\\
&\qquad +2\left(V_d(\mathbf{x})|\psi|^2 + \beta_d|\psi|^4\right) - i\Omega\left[\psi\left(x\partial_y - y\partial_x\right)\psi^* - \psi^*\left(x\partial_y - y\partial_x\right)\psi\right]\\
&\qquad -2i\Omega\psi^*\left[\partial_x\alpha\left(\alpha\partial_y + y\partial_\alpha\right)\psi - \partial_y\alpha\left(\alpha\partial_x + x\partial_\alpha\right)\psi\right]\\
&\qquad +2\Omega^2|\psi|^2\left[\left(y^2 - \alpha x\right)\partial_x\alpha + \left(x^2 - \alpha y\right)\partial_y\alpha\right]\Bigg] d\mathbf{x}\\
&= \int_{\mathbb{R}^d}\Big[-4i\Omega\psi^*\left[\partial_x\alpha\left(\alpha\partial_y + y\partial_\alpha\right)\psi - \partial_y\alpha\left(\alpha\partial_x + x\partial_\alpha\right)\psi\right] + 2|\partial_\alpha\psi|^2 + \beta_d|\psi|^4\\
&\qquad +2\Omega^2|\psi|^2\left[\left(y^2 - \alpha x\right)\partial_x\alpha + \left(x^2 - \alpha y\right)\partial_y\alpha\right] - 2\alpha|\psi|^2\partial_\alpha\left(V_d(\mathbf{x})\right)\Big] d\mathbf{x}\\
&= \int_{\mathbb{R}^d}\Big[\left(\partial_y\alpha - \partial_x\alpha\right)\left[4i\Omega\psi^*(x\partial_y + y\partial_x)\psi + 2\Omega^2\left(x^2 - y^2\right)|\psi|^2\right]\\
&\qquad +2|\partial_\alpha\psi|^2 + \beta_d|\psi|^4 - 2\alpha|\psi|^2\partial_\alpha\left(V_d(\mathbf{x})\right)\Big] d\mathbf{x}. \qquad (2.12)
\end{aligned}
$$

Furthermore, noticing (1.5), (2.6), and (2.10) with $t = 0$, we get (2.8) and (2.9) immediately. □

LEMMA 2.3. (i) *In 2D with a radial symmetric trap, i.e., $d = 2$ and $\gamma_x = \gamma_y := \gamma_r$ in* (1.4)*, for any initial data $\psi_0 = \psi_0(x, y)$, we have, for any $t \geq 0$,*

$$
\delta_r(t) = \frac{E_{\beta,\Omega}(\psi_0) + \Omega\langle L_z\rangle(0)}{\gamma_r^2}\left[1 - \cos(2\gamma_r t)\right] + \delta_r^{(0)}\cos(2\gamma_r t) + \frac{\delta_r^{(1)}}{2\gamma_r}\sin(2\gamma_r t), \qquad (2.13)
$$

where $\delta_r(t) = \delta_x(t) + \delta_y(t)$, $\delta_r^{(0)} := \delta_x(0) + \delta_y(0)$, and $\delta_r^{(1)} := \dot\delta_x(0) + \dot\delta_y(0)$. Furthermore, when the initial condition $\psi_0(x, y)$ in (1.5) *satisfies*

$$
\psi_0(x, y) = f(r)e^{im\theta} \quad \text{with} \quad m \in \mathbb{Z} \quad \text{and} \quad f(0) = 0 \quad \text{when} \quad m \neq 0, \qquad (2.14)
$$

we have, for any $t \geq 0$,

$$
\begin{aligned}
\delta_x(t) = \delta_y(t) &= \frac{1}{2}\delta_r(t)\\
&= \frac{E_{\beta,\Omega}(\psi_0) + m\Omega}{2\gamma_x^2}\left[1 - \cos(2\gamma_x t)\right] + \delta_x^{(0)}\cos(2\gamma_x t) + \frac{\delta_x^{(1)}}{2\gamma_x}\sin(2\gamma_x t). \qquad (2.15)
\end{aligned}
$$

This and (2.6) *imply that*

$$
\sigma_x = \sigma_y = \sqrt{\frac{E_{\beta,\Omega}(\psi_0) + m\Omega}{2\gamma_x^2}\left[1 - \cos(2\gamma_x t)\right] + \delta_x^{(0)}\cos(2\gamma_x t) + \frac{\delta_x^{(1)}}{2\gamma_x}\sin(2\gamma_x t)}. \qquad (2.16)
$$

Thus in this case, the condensate widths $\sigma_x(t)$ and $\sigma_y(t)$ are periodic functions with frequency doubling the trapping frequency.

(ii) *For all other cases, we have, for any* $t \geq 0$,

$$\delta_\alpha(t) = \frac{E_{\beta,\Omega}(\psi_0)}{\gamma_\alpha^2} + \left(\delta_\alpha^{(0)} - \frac{E_{\beta,\Omega}(\psi_0)}{\gamma_\alpha^2}\right)\cos(2\gamma_\alpha t) + \frac{\delta_\alpha^{(1)}}{2\gamma_\alpha}\sin(2\gamma_\alpha t) + f_\alpha(t), \tag{2.17}$$

where $f_\alpha(t)$ *is the solution of the following second-order ODE:*

$$\frac{d^2 f_\alpha(t)}{dt^2} + 4\gamma_\alpha^2\, f_\alpha(t) = F_\alpha(t), \qquad f_\alpha(0) = \frac{df_\alpha(0)}{dt} = 0, \tag{2.18}$$

with

$$\begin{aligned} F_\alpha(t) = \int_{\mathbb{R}^d} \Big[& 2|\partial_\alpha\psi|^2 - 2|\nabla\psi|^2 - \beta_d|\psi|^4 + \left(2\gamma_\alpha^2\alpha^2 - 4V_d(\mathbf{x})\right)|\psi|^2 + 4\Omega\psi^* L_z\psi \\ & + (\partial_y\alpha - \partial_x\alpha)\left(4i\Omega\psi^*\left(x\partial_y + y\partial_x\right)\psi + 2\Omega^2(x^2-y^2)|\psi|^2\right)\Big] d\mathbf{x}. \end{aligned}$$

Proof. From (2.7) with $d = 2$, we have

$$\begin{aligned} & \frac{d^2\delta_x(t)}{dt^2} + 2\gamma_x^2\delta_x(t) \\ & = \int_{\mathbb{R}^2}\left[2|\partial_x\psi|^2 + \beta_2|\psi|^4 - 4i\Omega\psi^*\left(x\partial_y + y\partial_x\right)\psi - 2\Omega^2\left(x^2-y^2\right)|\psi|^2\right] d\mathbf{x}, \end{aligned} \tag{2.19}$$

$$\begin{aligned} & \frac{d^2\delta_y(t)}{dt^2} + 2\gamma_y^2\delta_y(t) \\ & = \int_{\mathbb{R}^2}\left[2|\partial_y\psi|^2 + \beta_2|\psi|^4 + 4i\Omega\psi^*\left(x\partial_y + y\partial_x\right)\psi + 2\Omega^2\left(x^2-y^2\right)|\psi|^2\right] d\mathbf{x}. \end{aligned} \tag{2.20}$$

For case (i) with $\gamma_x = \gamma_y := \gamma_r$ in (1.4), summing up (2.19) and (2.20) together and applying (1.8) and (2.3), we have the following ODE for $\delta_r(t)$:

$$\begin{aligned} \frac{d^2\delta_r(t)}{dt^2} &= -2\gamma_r^2\delta_r(t) + \int_{\mathbb{R}^2}\left[2|\nabla\psi|^2 + 2\beta_2|\psi|^4\right] d\mathbf{x} \\ &= -2\gamma_r^2\delta_r(t) - 4\int_{\mathbb{R}^2}\left[V_2(\mathbf{x})|\psi|^2 - \Omega\psi^* L_z\psi\right] d\mathbf{x} \\ &\quad + 4\int_{\mathbb{R}^2}\left[\frac{1}{2}|\nabla\psi|^2 + V_2(\mathbf{x})|\psi|^2 + \frac{\beta_2}{2}|\psi|^4 - \Omega\psi^* L_z\psi\right] d\mathbf{x} \\ &= -2\gamma_r^2\delta_r(t) - 2\gamma_r^2\delta r(t) + 4\Omega\langle L_z\rangle(t) + 4E_{\beta,\Omega}(\psi(\cdot,t)) \\ &= -4\gamma_r^2\delta_r(t) + 4E_{\beta,\Omega}(\psi_0) + 4\Omega\langle L_z\rangle(0), \qquad t \geq 0, \end{aligned} \tag{2.21}$$

$$\delta_r(0) = \delta_r^{(0)}, \qquad \dot{\delta}_r(0) = \delta_r^{(1)}. \tag{2.22}$$

Thus, (2.13) is the unique solution of the second-order ODE (2.21) with the initial data (2.22). Furthermore, when the initial data $\psi_0(\mathbf{x})$ in (1.5) satisfies (2.14), due to symmetry, the solution $\psi(\mathbf{x},t)$ of (1.4)–(1.5) satisfies

$$\psi(x,y,t) = g(r,t)e^{im\theta} \quad \text{with} \quad g(r,0) = f(r). \tag{2.23}$$

This implies

$$\begin{aligned}\delta_x(t) &= \int_{\mathbb{R}^2} x^2|\psi(x,y,t)|^2\,d\mathbf{x} = \int_0^\infty\int_0^{2\pi} r^2\cos^2\theta|g(r,t)|^2 r\,d\theta dr\\ &= \pi\int_0^\infty r^2|g(r,t)|^2 r\,dr = \int_0^\infty\int_0^{2\pi} r^2\sin^2\theta|g(r,t)|^2 r\,d\theta dr\\ &= \int_{\mathbb{R}^2} y^2|\psi(x,y,t)|^2\,d\mathbf{x} = \delta_y(t), \qquad t\ge 0. \end{aligned} \tag{2.24}$$

Since $\gamma_x = \gamma_y$, by Lemma 2.1, we know in this case that

$$\begin{aligned}\langle L_z\rangle(t) &= \langle L_z\rangle(0) = -i\int_{\mathbb{R}^2}\psi_0^*(x,y)\partial_\theta\psi_0(x,y)\,d\mathbf{x}\\ &= 2\pi m\int_0^\infty |f(r)|^2 r\,dr = m\|\psi_0\|^2 = m.\end{aligned} \tag{2.25}$$

Thus, (2.15) is a combination of (2.13), (2.24), and (2.25).

(ii) From (2.7) and noticing the energy conservation (1.8), we have

$$\begin{aligned}\frac{d^2\delta_\alpha(t)}{dt^2} &= \int_{\mathbb{R}^d}\Big[(\partial_y\alpha-\partial_x\alpha)\left[4i\Omega\psi^*(x\partial_y+y\partial_x)\psi+2\Omega^2\left(x^2-y^2\right)|\psi|^2\right]\\ &\qquad +2|\partial_\alpha\psi|^2+\beta_d|\psi|^4-2\gamma_\alpha^2\alpha^2|\psi|^2\Big]\,d\mathbf{x}\\ &= -4\gamma_\alpha^2\delta_\alpha(t)+4\int_{\mathbb{R}^2}\left[\frac{1}{2}|\nabla\psi|^2+V_d(\mathbf{x})|\psi|^2+\frac{\beta_d}{2}|\psi|^4-\Omega\psi^*L_z\psi\right]d\mathbf{x}\\ &\quad+\int_{\mathbb{R}^d}\Big[2|\partial_\alpha\psi|^2-2|\nabla\psi|^2-\beta_d|\psi|^4+\left(2\gamma_\alpha^2\alpha^2-4V_d(\mathbf{x})\right)|\psi|^2+4\Omega\psi^*L_z\psi\\ &\qquad+(\partial_y\alpha-\partial_x\alpha)\left(4i\Omega\psi^*\left(x\partial_y+y\partial_x\right)\psi+2\Omega^2(x^2-y^2)|\psi|^2\right)\Big]d\mathbf{x}\\ &= -4\gamma_\alpha^2\delta_\alpha(t)+4E_{\beta,\Omega}(\psi(\cdot,t))+F_\alpha(t)\\ &= -4\gamma_\alpha^2\delta_\alpha(t)+4E_{\beta,\Omega}(\psi_0)+F_\alpha(t), \qquad t\ge 0.\end{aligned} \tag{2.26}$$

Thus (2.17) is the unique solution of the second-order ODE (2.26) with the initial data (2.8), (2.9). ∎

2.3. Dynamics of a stationary state with its center shifted. Let $\phi_e(\mathbf{x})$ be a stationary state of the GPE (1.4) with a chemical potential μ_e [13, 12], i.e., (μ_e,ϕ_e) satisfying

$$\mu_e\phi_e(\mathbf{x}) = -\frac{1}{2}\nabla^2\phi_e+V_d(\mathbf{x})\phi_e+\beta_d|\phi_e|^2\phi_e-\Omega L_z\phi_e, \qquad \|\phi_e\|^2=1. \tag{2.27}$$

If the initial data $\psi_0(\mathbf{x})$ in (1.5) is chosen as a stationary state with a shift in its center, one can construct an exact solution of the GPE (1.4) with an harmonic oscillator potential (1.6). This kind of analytical construction can be used, in particular, in the benchmark and validation of numerical algorithms for the GPE. In [27], a similar kind of solution was constructed for the GPE and a second order ODE system was derived for the dynamics of the center, but the results there were valid only for nonrotating BEC, i.e., $\Omega = 0$. Modifications must be made for the rotating BEC, i.e., $\Omega \neq 0$. Later, in [15], similar results were extended to the case of a general Hamiltonian but

without specifying the initial data for the ODE system. For the convenience of the reader, here we present a simple derivation of the dynamic laws for rotating BEC.

LEMMA 2.4. *If the initial data* $\psi_0(\mathbf{x})$ *in* (1.5) *is chosen as*

$$\psi_0(\mathbf{x}) = \phi_e(\mathbf{x} - \mathbf{x}_0), \qquad \mathbf{x} \in \mathbb{R}^d, \tag{2.28}$$

where $\mathbf{x}_0$ *is a given point in* $\mathbb{R}^d$*, then the exact solution of* (1.4)–(1.5) *satisfies*

$$\psi(\mathbf{x}, t) = \phi_e(\mathbf{x} - \mathbf{x}(t))\, e^{-i\mu_e t}\, e^{iw(\mathbf{x},t)}, \quad \mathbf{x} \in \mathbb{R}^d, \quad t \geq 0, \tag{2.29}$$

where for any time $t \geq 0$*,* $w(\mathbf{x}, t)$ *is linear for* $\mathbf{x}$*, i.e.,*

$$w(\mathbf{x}, t) = \mathbf{c}(t) \cdot \mathbf{x} + g(t), \quad \mathbf{c}(t) = (c_1(t), \ldots, c_d(t))^T, \quad \mathbf{x} \in \mathbb{R}^d, \quad t \geq 0, \tag{2.30}$$

and $\mathbf{x}(t)$ *satisfies the following second-order ODE system:*

$$\ddot{x}(t) - 2\Omega\dot{y}(t) + \left(\gamma_x^2 - \Omega^2\right) x(t) = 0, \tag{2.31}$$

$$\ddot{y}(t) + 2\Omega\dot{x}(t) + \left(\gamma_y^2 - \Omega^2\right) y(t) = 0, \qquad t \geq 0, \tag{2.32}$$

$$x(0) = x_0, \quad y(0) = y_0, \quad \dot{x}(0) = \Omega y_0, \quad \dot{y}(0) = -\Omega x_0. \tag{2.33}$$

Moreover, if in $3D$*, another ODE needs to be added:*

$$\ddot{z}(t) + \gamma_z^2 z(t) = 0, \quad z(0) = z_0, \quad \dot{z}(0) = 0. \tag{2.34}$$

Proof. For $d = 2$, we introduce

$$J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad A = \begin{pmatrix} \gamma_x^2 & 0 \\ 0 & \gamma_y^2 \end{pmatrix}, \quad \nabla = \begin{pmatrix} \partial_x \\ \partial_y \end{pmatrix}.$$

Differentiating (2.29) with respect to t and $\mathbf{x}$, respectively, plugging into (1.4), changing variable $\mathbf{x} - \mathbf{x}(t) \to \mathbf{x}$, and noticing (2.27), we obtain for $\phi_e = \phi_e(\mathbf{x})$ and $w = w(\mathbf{x} + \mathbf{x}(t), t)$ that

$$\begin{aligned} \phi_e \partial_t w + i\dot{\mathbf{x}}(t) \cdot \nabla\phi_e = \frac{1}{2}\left[i\phi_e \nabla^2 w - \phi_e |\nabla w|^2 - \mathbf{x}(t)^T A(2\mathbf{x} + \mathbf{x}(t))\phi_e\right] \\ + i\nabla\phi_e \cdot \nabla w - \phi_e \Omega(\mathbf{x} + \mathbf{x}(t)) \cdot (J\nabla w) + i\Omega\mathbf{x}(t) \cdot (J\nabla\phi_e). \end{aligned} \tag{2.35}$$

Taking the real and imaginary parts in (2.35) and noticing (2.30), we have

$$[\dot{\mathbf{x}}(t) - \nabla w(\mathbf{x} + \mathbf{x}(t), t) - \Omega J\mathbf{x}(t)] \cdot \nabla\phi_e = 0, \tag{2.36}$$

$$\left[\partial_t w + \frac{1}{2}|\nabla w|^2 + \frac{1}{2}\mathbf{x}(t)^T A(2\mathbf{x} + \mathbf{x}(t)) - \Omega(\mathbf{x} + \mathbf{x}(t)) \cdot (J\nabla w)\right] \phi_e = 0. \tag{2.37}$$

We thus get

$$\dot{\mathbf{x}}(t) = \nabla w(\mathbf{x} + \mathbf{x}(t), t) + \Omega J\mathbf{x}(t), \tag{2.38}$$

$$\begin{aligned} \partial_t w(\mathbf{x} + \mathbf{x}(t), t) = -\frac{1}{2}\left[|\nabla w|^2 + \mathbf{x}(t)^T A(2\mathbf{x} + \mathbf{x}(t))\right] \\ + \Omega(\mathbf{x} + \mathbf{x}(t)) \cdot (J\nabla w). \end{aligned} \tag{2.39}$$

Differentiating (2.38) and (2.39) with respect to t and $\mathbf{x}$, respectively, and noticing (2.30), which implies that $|\nabla w|^2$ is independent of $\mathbf{x}$, we obtain

$$\begin{aligned}
0 &= \ddot{\mathbf{x}}(t) - \partial_t(\nabla w(\mathbf{x}+\mathbf{x}(t),t)) - \Omega J\dot{\mathbf{x}}(t)\\
&= \ddot{\mathbf{x}}(t) - \nabla(\partial_t w(\mathbf{x}+\mathbf{x}(t),t)) - \dot{\mathbf{x}}(t)\,\nabla^2 w(\mathbf{x}+\mathbf{x}(t),t) - \Omega J\dot{\mathbf{x}}(t)\\
&= \ddot{\mathbf{x}}(t) - \nabla(\partial_t w(\mathbf{x}+\mathbf{x}(t),t)) - \Omega J\dot{\mathbf{x}}(t)\\
&= \ddot{\mathbf{x}}(t) + A\mathbf{x}(t) - \Omega J\,[\dot{\mathbf{x}}(t) - \Omega J\mathbf{x}(t)] - \Omega J\dot{\mathbf{x}}(t)\\
&= \ddot{\mathbf{x}}(t) - 2\Omega J\dot{\mathbf{x}}(t) + (A+\Omega^2 J^2)\mathbf{x}(t)\\
&= \ddot{\mathbf{x}}(t) - 2\Omega J\dot{\mathbf{x}}(t) + (A-\Omega^2 I)\mathbf{x}(t), \qquad t\ge 0. \qquad (2.40)
\end{aligned}$$

From (2.29) with $t=0$, we get

$$\mathbf{x}(0)=\mathbf{x}_0, \quad w(\mathbf{x},0)\equiv 0, \quad \mathbf{x}\in\mathbb{R}^d. \qquad (2.41)$$

Thus (2.33) is a combination of (2.41) and (2.38) with $t=0$. For $d=3$, the proof is similar, and the details are omitted here. □

In the literature, constructions similar to the above are often numerically verified by directly simulating the dynamics of the GPE in nonrotating BEC [14, 27]. To our knowledge, the above lemma gives the first rigorous derivation for rotating BEC.

Notice that if $\mathbf{u}=\dot{\mathbf{x}}(t)-\Omega J\mathbf{x}(t)$, then (2.40) gives a coupled first-order system

$$\begin{cases} \dot{\mathbf{x}}(t) = \Omega J\mathbf{x}(t) + \mathbf{u},\\ \dot{\mathbf{u}}(t) = -A\mathbf{x}(t) + \Omega J\mathbf{u},\\ \mathbf{x}(0)=\mathbf{x}_0, \ \ \mathbf{u}(0)=0, \end{cases} \qquad (2.42)$$

which is a Hamiltonian system with the Hamiltonian $H(\mathbf{x},\mathbf{u}) = \Omega\mathbf{u}^T J\mathbf{x} + (\mathbf{u}^T\mathbf{u} + \mathbf{x}^T A\mathbf{x})/2$. The characteristic roots λ of the system are given by the equation

$$\lambda^4 + (\gamma_x^2+\gamma_y^2+2\Omega^2)\lambda^2 + (\gamma_x^2-\Omega^2)(\gamma_y^2-\Omega^2) = 0. \qquad (2.43)$$

The exact solutions of (2.42) may thus be completely determined.

We note not only that results on the dynamics of a stationary state with its center shifted are physically interesting this type of exact solution of the time-dependent GPE can also serve as a good benchmark for numerical algorithms and is useful in the mathematical studies of the dynamic stabilities of the vortex state in BEC. In section 4.2, we will study this kind of dynamics by directly simulating the GPE in a rotational frame and explore different motion patterns of a stationary state center under different rotation speed Ω.

2.4. Dynamics of the total density in the presence of dissipation. Consider a more general GPE of the form

$$(i-\lambda)\partial_t\psi(\mathbf{x},t) = -\frac{1}{2}\nabla^2\psi + V(\mathbf{x},t)\psi + \beta_d|\psi|^2\psi - \Omega L_z\psi, \quad \mathbf{x}\in\mathbb{R}^d, \quad t>0, \qquad (2.44)$$

$$\psi(\mathbf{x},0)=\psi_0(\mathbf{x}), \quad \mathbf{x}\in\mathbb{R}^d, \qquad (2.45)$$

where $\lambda\ge 0$ is a real parameter that models a dissipation mechanism [2, 30, 32] and $V(\mathbf{x},t) = V_d(\mathbf{x}) + W(\mathbf{x},t)$ with $W(\mathbf{x},t)$ an external driven field [16, 31]. Typical external driven fields used in physics literature include a Delta kicked potential [31]

$$W(x,t) = K_s\cos(k_s x)\sum_{n=-\infty}^{\infty}\delta(t-n\tau), \qquad (2.46)$$

with K_s being the kick strength, k_s the wavenumber, τ the time interval between kicks, and $\delta(\tau)$ the Dirac delta function, or a far-blue detuned Gaussian laser beam stirrer [16]

$$W(\mathbf{x},t) = W_s(t)\exp\left[-\left(\frac{|\mathbf{x}-\mathbf{x}_s(t)|^2}{w_s/2}\right)\right] \tag{2.47}$$

with $W_s(t)$ being the height, w_s the width, and $\mathbf{x}_s(t)$ the position of the stirrer. In addition, we note that to study the onset of energy dissipation in BEC stirred by a laser field, another possibility is to view the beam as an translating *obstacle* [2] instead of introducing the Gaussian potential.

While the total density remains constant with $\lambda = 0$, in the more general case, we have the following lemma for the dynamics of the total density.

LEMMA 2.5. *Let $\psi(\mathbf{x},t)$ be the solution of (2.44)–(2.45); then the total density satisfies*

$$\dot{N}(\psi)(t) = \frac{d}{dt}\int_{\mathbb{R}^d}|\psi(\mathbf{x},t)|^2\,d\mathbf{x} = -\frac{2\lambda}{1+\lambda^2}\mu_{\beta,\Omega}(\psi), \qquad t\geq 0, \tag{2.48}$$

where

$$\mu_{\beta,\Omega}(\psi) = \int_{\mathbb{R}^d}\left[\frac{1}{2}|\nabla\psi|^2 + V(\mathbf{x},t)|\psi|^2 + \beta_d|\psi|^4 - \Omega\mathrm{Re}(\psi^* L_z\psi)\right]d\mathbf{x}.$$

Consequently, the total density decreases when $\lambda > 0$ and $|\Omega| \leq \gamma_{xy} := \min\{\gamma_x,\gamma_y\}$.

Proof. Dividing (2.44) by $(i-\lambda)$, multiplying it by ψ^* and summing with its complex conjugate, and integrating by parts, we obtain

$$\begin{aligned}
\frac{dN(\psi)}{dt} &= \int_{\mathbb{R}^d}\Bigg[-\frac{i+\lambda}{1+\lambda^2}\left(-\frac{1}{2}\nabla^2\psi + V(\mathbf{x},t)\psi + \beta_d|\psi|^2\psi - \Omega L_z\psi\right)\psi^* \\
&\qquad + \frac{i-\lambda}{1+\lambda^2}\left(-\frac{1}{2}\nabla^2\psi^* + V(\mathbf{x},t)\psi^* + \beta_d|\psi|^2\psi^* - \Omega(L_z)^*\psi^*\right)\psi\Bigg]d\mathbf{x} \\
&= \frac{\lambda}{1+\lambda^2}\int_{\mathbb{R}^d}\Bigg[\frac{1}{2}\left(\psi^*\nabla^2\psi + \psi\nabla^2\psi^*\right) - 2\left(V(\mathbf{x},t)|\psi|^2 + \beta_d|\psi|^4\right) \\
&\qquad\qquad + \Omega\left(\psi^* L_z\psi + \psi(L_z)^*\psi^*\right)\Bigg]d\mathbf{x} \\
&= \frac{-2\lambda}{1+\lambda^2}\int_{\mathbb{R}^d}\left[\frac{1}{2}|\nabla\psi|^2 + V(\mathbf{x},t)|\psi|^2 + \beta_d|\psi|^4 - \Omega\mathrm{Re}(\psi^* L_z\psi)\right]d\mathbf{x} \\
&= \frac{-2\lambda}{1+\lambda^2}\mu_{\beta,\Omega}(\psi). \qquad (2.49)
\end{aligned}$$

When $\gamma > 0$ and $|\Omega| < \gamma_{xy}$, by a completion of square [1, 13], we have

$$\frac{1}{2}|\nabla\psi|^2 + V(\mathbf{x},t)|\psi|^2 - \Omega\mathrm{Re}(\psi^* L_z\psi) = \frac{1}{2}|(\nabla - i\mathbf{A})\psi|^2 + \left[V(\mathbf{x},t) - \frac{|\Omega|^2}{2}(x^2+y^2)\right]|\psi|^2$$

for a vector potential $\mathbf{A} = \mathbf{A}(x,y) = (y,-x)\Omega$ in 2D and $\mathbf{A} = \mathbf{A}(x,y,z) = (y,-x,0)\Omega$ in 3D. Thus, $\mu_{\beta,\Omega}(\psi) > 0$. Consequently, we get

$$\frac{dN(\psi)}{dt} < 0, \qquad t\geq 0, \tag{2.50}$$

which immediately implies the decreasing of the total density. ∎

3. Numerical methods. In this section, we will present an efficient and accurate numerical method to solve the following GPE for the dynamics of rotating BEC.

Due to the trapping potential $V_d(\mathbf{x})$ given by (1.6), the solution $\psi(\mathbf{x},t)$ of (2.44)–(2.45) decays to zero exponentially fast when $|\mathbf{x}| \to \infty$. Thus in practical computation, we truncate the problem (2.44)–(2.45) into a bounded computational domain with the homogeneous Dirichlet boundary condition:

$$(i-\lambda)\partial_t\psi(\mathbf{x},t) = -\frac{1}{2}\nabla^2\psi + V(\mathbf{x},t)\psi + \beta_d|\psi|^2\psi - \Omega L_z\psi, \quad \mathbf{x}\in\Omega_{\mathbf{x}}, \quad t>0, \tag{3.1}$$

$$\psi(\mathbf{x},t) = 0, \quad \mathbf{x}\in\Gamma = \partial\Omega_{\mathbf{x}}, \quad t\geq 0, \tag{3.2}$$

$$\psi(\mathbf{x},0) = \psi_0(\mathbf{x}), \qquad \mathbf{x}\in\bar{\Omega}_{\mathbf{x}}, \tag{3.3}$$

where we choose $\Omega_{\mathbf{x}} = \{(x,y),\ r=\sqrt{x^2+y^2} < R\}$ in 2D and, respectively, $\Omega_{\mathbf{x}} = \{(x,y,z),\ r=\sqrt{x^2+y^2}<R,\ a<z<b\}$ in 3D with R, $|a|$, and b sufficiently large. The use of more sophisticated radiation boundary conditions is an interesting topic that remains to be examined in the future.

3.1. Time-splitting. We choose a time step size $\Delta t>0$. For $n=0,1,2,\ldots,$ from time $t=t_n=n\Delta t$ to $t=t_{n+1}=t_n+\Delta t$, the GPE (3.1) is solved in two splitting steps. One first solves

$$(i-\lambda)\ \partial_t\psi(\mathbf{x},t) = -\frac{1}{2}\nabla^2\psi - \Omega L_z\psi \tag{3.4}$$

for the time step of length Δt, followed by solving

$$(i-\lambda)\ \partial_t\psi(\mathbf{x},t) = V(\mathbf{x},t)\psi + \beta_d|\psi|^2\psi \tag{3.5}$$

for the same time step. Equation (3.4) will be discretized in detail in the next two subsections. For $t\in[t_n,t_{n+1}]$, after dividing (3.5) by $(i-\lambda)$, multiplying it by ψ^*, and adding with its complex conjugate, we obtain the following ODE for $\rho(\mathbf{x},t) = |\psi(\mathbf{x},t)|^2$:

$$\partial_t\rho(\mathbf{x},t) = -\frac{2\lambda}{1+\lambda^2}\left[V(\mathbf{x},t)\rho(\mathbf{x},t) + \beta_d\rho^2(\mathbf{x},t)\right], \quad \mathbf{x}\in\Omega_{\mathbf{x}}, \quad t_n\leq t\leq t_{n+1}. \tag{3.6}$$

The ODE for the phase angle $\phi(\mathbf{x},t)$ (determined as $\psi=\sqrt{\rho}e^{i\phi}$) is given by

$$\phi_t = -\frac{1}{1+\lambda^2}\left[V(\mathbf{x},t) + \beta_d\rho(\mathbf{x},t)\right], \quad \mathbf{x}\in\Omega_{\mathbf{x}}, \quad t_n\leq t\leq t_{n+1}. \tag{3.7}$$

For $\lambda\neq 0$, by (3.6), the above is equivalent to

$$\phi_t = \frac{1}{2\lambda}\partial_t\ln\rho, \quad \mathbf{x}\in\Omega_{\mathbf{x}}, \quad t_n\leq t\leq t_{n+1}. \tag{3.8}$$

Denoting $V_n(\mathbf{x},t) = \int_{t_n}^t V(\mathbf{x},\tau)d\tau$, we can solve (3.6) to get

$$\rho(\mathbf{x},t) = \frac{\rho(\mathbf{x},t_n)\ \exp[\frac{-2\lambda V_n(\mathbf{x},t)}{1+\lambda^2}]}{1+\rho(\mathbf{x},t_n)\frac{2\lambda\beta_d}{1+\lambda^2}\int_{t_n}^t \exp[\frac{-2\lambda V_n(\mathbf{x},\tau)}{1+\lambda^2}]\ d\tau}. \tag{3.9}$$

Consequently, in the special case $V(\mathbf{x},t) = V(\mathbf{x})$, we have some exact analytical solutions given by

$$\rho(\mathbf{x},t) = \begin{cases} \rho(\mathbf{x},t_n), & \lambda = 0, \\ \dfrac{(1+\lambda^2)\rho(\mathbf{x},t_n)}{(1+\lambda^2)+2\lambda\beta_d(t-t_n)\rho(\mathbf{x},t_n)}, & V(\mathbf{x}) = 0, \\ \dfrac{V(\mathbf{x})\rho(\mathbf{x},t_n)\ \exp[\frac{-2\lambda V(\mathbf{x})(t-t_n)}{1+\lambda^2}]}{V(\mathbf{x})+\left(1-\exp[\frac{-2\lambda V(\mathbf{x})(t-t_n)}{1+\lambda^2}]\right)\beta_d\rho(\mathbf{x},t_n)}, & V(\mathbf{x}) \neq 0. \end{cases} \tag{3.10}$$

Plugging (3.9) into (3.5), we get, for $t \in [t_n, t_{n+1}]$,

$$\psi(\mathbf{x},t) = \psi(\mathbf{x},t_n)\sqrt{U_n(\mathbf{x},t)}\exp\left[-\frac{i}{1+\lambda^2}\left(V_n(\mathbf{x},t)+\beta_d\int_{t_n}^{t}\rho(\mathbf{x},\tau)d\tau\right)\right], \tag{3.11}$$

where

$$U_n(\mathbf{x},t) = \frac{\exp[\frac{-2\lambda V_n(\mathbf{x},t)}{1+\lambda^2}]}{1+|\psi(\mathbf{x},t_n)|^2\frac{2\lambda\beta_d}{1+\lambda^2}\int_{t_n}^{t}\exp[\frac{-2\lambda V_n(\mathbf{x},\tau)}{1+\lambda^2}]\,d\tau}. \tag{3.12}$$

Again, with $V(\mathbf{x},t) = V(\mathbf{x})$, we can integrate exactly to get

$$\psi(\mathbf{x},t) = \psi(\mathbf{x},t_n)\begin{cases} \exp\left[-i(\beta_d|\psi(\mathbf{x},t_n)|^2+V(\mathbf{x}))(t-t_n)\right], & \lambda = 0, \\ \sqrt{\hat{U}_n(\mathbf{x},t)}\exp[\frac{i}{2\lambda}\ln\hat{U}_n(\mathbf{x},t)], & \lambda \neq 0, \end{cases} \tag{3.13}$$

where

$$\hat{U}_n(\mathbf{x},t) = \begin{cases} \dfrac{1+\lambda^2}{1+\lambda^2+2\lambda\beta_d(t-t_n)|\psi(\mathbf{x},t_n)|^2}, & V(\mathbf{x}) = 0, \\ \dfrac{V(\mathbf{x})\exp[-\frac{2\lambda(t-t_n)V(\mathbf{x})}{1+\lambda^2}]}{V(\mathbf{x})+\left(1-\exp[-\frac{2\lambda(t-t_n)V(\mathbf{x})}{1+\lambda^2}]\right)\beta_d|\psi(\mathbf{x},t_n)|^2}, & V(\mathbf{x}) \neq 0. \end{cases}$$

Remark 3.1. If the function $V_n(\mathbf{x},t)$ as well as other integrals in (3.9), (3.11), and (3.12) cannot be evaluated analytically, numerical quadrature can be used, e.g.,

$$V_n(\mathbf{x},t_{n+1}) = \int_{t_n}^{t_{n+1}} V(\mathbf{x},\tau)\,d\tau \approx \frac{\Delta t}{6}\left[V(\mathbf{x},t_n)+4V(\mathbf{x},t_n+\Delta t/2)+V(\mathbf{x},t_{n+1})\right].$$

3.2. Discretization in 2D. To solve (3.4), we try to formulate the equation in a variable separable form. When $d = 2$, we use the polar coordinate (r,θ) and discretize in the θ-direction by a Fourier pseudospectral method, in the r-direction by a finite element method (FEM), and in time by a Crank–Nicolson (C–N) scheme. Assume that

$$\psi(r,\theta,t) = \sum_{l=-L/2}^{L/2-1}\widehat{\psi}_l(r,t)\ e^{il\theta}, \tag{3.14}$$

where L is an even positive integer and $\widehat{\psi}_l(r,t)$ is the Fourier coefficient for the lth mode. Plugging (3.14) into (3.4) and noticing the orthogonality of the Fourier functions, we obtain, for $-\frac{L}{2} \le l \le \frac{L}{2} - 1$ and $0 < r < R$,

$$(i-\lambda)\,\partial_t \widehat{\psi}_l(r,t) = -\frac{1}{2r}\frac{\partial}{\partial r}\left(r\frac{\partial \widehat{\psi}_l(r,t)}{\partial r}\right) + \left(\frac{l^2}{2r^2} - l\Omega\right)\widehat{\psi}_l(r,t), \tag{3.15}$$

$$\widehat{\psi}_l(R,t) = 0 \quad (\text{for all } l), \qquad \widehat{\psi}_l(0,t) = 0 \quad (\text{for } l \neq 0). \tag{3.16}$$

Let P^k denote all polynomials with degree at most k, let $M > 0$ be a chosen integer, and $0 = r_0 < r_1 < r_2 < \cdots < r_M = R$ be a partition for the interval $[0, R]$ with a mesh size $h = \max_{0\le m<M}\{r_{m+1} - r_m\}$. Define a FEM subspace by

$$U^h = \left\{u^h \in C[0,R] \mid \ u^h\big|_{[r_m, r_{m+1}]} \in P^k,\ 0 \le m < M,\ u^h(R) = 0\right\}$$

for $l = 0$, and for $l \neq 0$,

$$U^h = \left\{u^h \in C[0,R] \mid \ u^h\big|_{[r_m, r_{m+1}]} \in P^k,\ 0 \le m < M,\ u^h(0) = u^h(R) = 0\right\};$$

then we obtain the FEM approximation for (3.15)–(3.16): Find $\hat{\psi}_l^h = \hat{\psi}_l^h(\cdot,t) \in U^h$ such that for all $\phi^h \in U^h$ and $t_n \le t \le t_{n+1}$,

$$(i-\lambda)\frac{d}{dt}A(\hat{\psi}_l^h(\cdot,t),\phi^h) = B(\hat{\psi}_l^h(\cdot,t),\phi^h) + l^2 C(\hat{\psi}_l^h,\phi^h) - l\Omega A(\hat{\psi}_l^h,\phi^h), \tag{3.17}$$

where

$$A(u^h,v^h) = \int_0^R r\ u^h(r)\ v^h(r)\ dr, \qquad B(u^h,v^h) = \int_0^R \frac{r}{2}\ \frac{du^h(r)}{dr}\ \frac{dv^h(r)}{dr}\ dr,$$
$$C(u^h,v^h) = \int_0^R \frac{1}{2r}\ u^h(r)\ v^h(r)\ dr, \qquad u^h,\ v^h \in U^h.$$

The ODE system (3.17) is then discretized by the standard C–N scheme in time. Although an implicit time discretization is applied for (3.17), the one-dimensional nature of the problem makes the coefficient matrix for the linear system band-limited. For example, if the piecewise linear polynomial is used, i.e., $k = 1$ in U^h, the matrix is tridiagonal. Fast algorithms can be applied to solve the resulting linear systems.

In practice, we always use the second-order Strang splitting [49]; i.e., from time $t = t_n$ to $t = t_{n+1}$ (i) evolve (3.5) for half time step $\Delta t/2$ with initial data given at $t = t_n$; (ii) evolve (3.4) for one time step Δt starting with the new data; (iii) evolve (3.5) for half time step $\Delta t/2$ with the newer data. For more general discussion on splitting methods, we refer the reader to [28, 38] for more details.

For the discretization considered here, the total memory requirement is $O(ML)$ and the total computational cost per time step is $O(ML \ln L)$. Furthermore, following the similar proofs in [6, 7, 14], the total density can be shown to be conserved in the discretized level when $\lambda = 0$ and to be decreased in the discretized level when $\lambda > 0$.

Remark 3.2. As noticed in [35, 34], another way for discretizing (3.15)–(3.16) is to use the finite difference in space on a mesh with a shifted grid and the C–N scheme in time. Choose an integer $M > 0$, a mesh size $\Delta r = 2R/(2M+1)$, and grid points

$r_m = (m-1/2)\Delta r$ for $0 \le m \le M+1$. Let $\hat{\psi}_{l,m}(t)$ be the approximation of $\hat{\psi}_l(r_m,t)$. A second-order finite difference discretization for (3.15)–(3.16) in space is

$$(i-\lambda)\frac{d\hat{\psi}_{l,m}(t)}{dt} = -\frac{r_{m+1/2}\hat{\psi}_{l,m+1}(t) - 2r_m\hat{\psi}_{l,m}(t) + r_{m-1/2}\hat{\psi}_{l,m-1}(t)}{2(\Delta r)^2 r_m} + \left(\frac{l^2}{2r_m^2} - l\Omega\right)\hat{\psi}_{l,m}(t), \quad m = 1,2,\ldots,M, \quad t_n \le t \le t_{n+1}, \tag{3.18}$$

with essential boundary conditions:

$$\hat{\psi}_{l,0}(t) = (-1)^l\hat{\psi}_{l,1}(t), \quad \hat{\psi}_{l,M+1}(t) = 0, \quad t_n \le t \le t_{n+1}. \tag{3.19}$$

The ODE system (3.18)–(3.19) may then be discretized in time by the C–N scheme so that only a tridiagonal linear system is to be solved with $O(M)$ arithmetic operations. We may further obtain a fourth-order finite difference discretization [35] for (3.15)–(3.16) on the interval $t \in [t_n, t_{n+1}]$:

$$\begin{aligned}(i-\lambda)\frac{d\hat{\psi}_{l,m}(t)}{dt} &= \left(\frac{l^2}{2r_m^2} - l\Omega\right)\hat{\psi}_{l,m}(t) \\ &\quad - \frac{-\hat{\psi}_{l,m+2}(t) + 16\hat{\psi}_{l,m+1}(t) - 30\hat{\psi}_{l,m}(t) + 16\hat{\psi}_{l,m-1}(t) - \hat{\psi}_{l,m-2}(t)}{24(\Delta r)^2} \\ &\quad - \frac{-\hat{\psi}_{l,m+2}(t) + 8\hat{\psi}_{l,m+1}(t) - 8\hat{\psi}_{l,m-1}(t) + \hat{\psi}_{l,m-2}(t)}{24\Delta r\ r_m}, \quad 1 \le m \le M, \end{aligned} \tag{3.20}$$

$$\begin{aligned}(i-\lambda)\frac{d\hat{\psi}_{l,M+1}(t)}{dt} &= \left(\frac{l^2}{2r_{M+1}^2} - l\Omega\right)\hat{\psi}_{l,M+1}(t) \\ &\quad - \frac{11\hat{\psi}_{l,M+2}(t) - 20\hat{\psi}_{l,M+1}(t) + 6\hat{\psi}_{l,M}(t) + 4\hat{\psi}_{l,M-1}(t) - \hat{\psi}_{l,M-2}(t)}{24(\Delta r)^2} \\ &\quad - \frac{3\hat{\psi}_{l,M+2}(t) + 10\hat{\psi}_{l,M+1}(t) - 18\hat{\psi}_{l,M}(t) + 6\hat{\psi}_{l,M-1}(t) - \hat{\psi}_{l,M-2}(t)}{24\Delta r\ r_{M+1}}, \end{aligned} \tag{3.21}$$

$$\hat{\psi}_{l,-1}(t) = (-1)^l\hat{\psi}_{l,2}(t), \quad \hat{\psi}_{l,0}(t) = (-1)^l\hat{\psi}_{l,1}(t), \quad \hat{\psi}_{l,M+1}(t) = 0. \tag{3.22}$$

Again the ODE system (3.20)–(3.22) may be discretized in time by the C–N scheme, and only a pentadiagonal linear system is to be solved, which can be done very efficiently too, i.e., via $O(M)$ arithmetic operations.

3.3. Discretization in 3D. When $d = 3$ in (3.4), we use the cylindrical coordinate (r,θ,z) and discretize in the θ-direction by the Fourier pseudospectral method, in the z-direction by the sine pseudospectral method, and in the r-direction by the finite element or finite difference method and in time by the C–N scheme. Assume that

$$\psi(r,\theta,z,t) = \sum_{l=-L/2}^{L/2-1}\sum_{k=1}^{K-1}\widehat{\psi}_{l,k}(r,t)\ e^{il\theta}\ \sin(\mu_k(z-a)), \tag{3.23}$$

where L and K are two even positive integers, $\mu_k = \frac{\pi k}{b-a}$ $(k = 1,\ldots,K-1)$, and $\widehat{\psi}_{l,k}(r,t)$ is the Fourier-sine coefficient for the (l,k)th mode. Plugging (3.23) into

(3.4) with $d = 3$ and noticing the orthogonality of the Fourier-sine modes, we obtain, for $-\frac{L}{2} \le l \le \frac{L}{2} - 1$, $1 \le k \le K - 1$, and $0 < r < R$, that

$$(i - \lambda)\,\partial_t \widehat{\psi}_{l,k}(r,t) = -\frac{1}{2r}\frac{\partial}{\partial r}\left(r\frac{\partial \widehat{\psi}_{l,k}(r,t)}{\partial r}\right) + \left(\frac{l^2}{2r^2} + \frac{\mu_k^2}{2} - l\Omega\right)\widehat{\psi}_{l,k}(r,t) \tag{3.24}$$

with essential boundary conditions

$$\widehat{\psi}_{l,k}(R,t) = 0 \text{ (for all } l), \qquad \widehat{\psi}_{l,k}(0,t) = 0 \text{ (for } l \neq 0). \tag{3.25}$$

The discretization of (3.24)–(3.25) is similar as that for (3.15)–(3.16) and is omitted here.

For the algorithm in 3D, the total memory requirement is $O(MLK)$ and the total computational cost per time step is $O(MLK \ln(LK))$.

4. Numerical simulations. In this section, we first test the accuracy of our numerical method. Then we apply it to study the dynamics of condensate width, a central vortex state with a shift in its center, and a quantized vortex lattice. Properties such as the conservation of energy and the angular momentum expectation and the stability of central vortices in rotating BEC are also discussed.

4.1. Numerical accuracy. To test the accuracy of our method, we take $d = 2$, $\lambda = 0$, $\gamma_x = \gamma_y = 1$, $\Omega = 0.8$, and $W(\mathbf{x}, t) \equiv 0$ in (2.44). The initial condition in (2.45) is taken as

$$\psi_0(\mathbf{x}) = \frac{2^{1/4}}{\pi^{1/2}} e^{-(x^2+2y^2)/2}, \qquad \mathbf{x} \in \mathbb{R}^2.$$

We take $R = 12$ for the bounded computational domain $\Omega_{\mathbf{x}}$ and the piecewise linear polynomial for U^h. Let ψ be the *exact* solution which is obtained numerically using our method with a very fine mesh and small time step, e.g., $\Delta r = \frac{1}{1024}$, $\Delta\theta = \frac{\pi}{128}$, and $\Delta t = 0.0001$, and let $\psi^{(\Delta r, \Delta\theta, \Delta t)}$ be the numerical solution obtained with mesh size $(\Delta r, \Delta\theta)$ and time step Δt.

First, we test the spectral accuracy in the θ-direction by choosing a very small mesh size in the r-direction $\Delta r = \frac{1}{1024}$ and the time step $\Delta t = 0.0001$ and by solving the problem for each fixed β_2 with different mesh size $\Delta\theta$ so that the discretization errors in the r-direction and in time can be neglected comparing to that in the θ-direction. The errors $||\psi(t) - \psi^{(\Delta r, \Delta\theta, \Delta t)}(t)||_{l^2}$ at $t = 2.0$ are shown in Table 4.1 for different values of β_2 and $\Delta\theta$.

Then we test the second-order accuracy in the r-direction by choosing a very fine mesh size $\Delta\theta = \frac{\pi}{128}$ and time step $\Delta t = 0.0001$ and by solving the problem with different values of β_2 and Δr. Table 4.2 shows the errors at $t = 2.0$ for different values of β_2 and Δr.

TABLE 4.1
Discretization error $\|\psi(t) - \psi^{(\Delta r, \Delta\theta, \Delta t)}(t)\|_{l^2}$ *at* $t = 2.0$ *in the* θ*-direction.*

Mesh size $\Delta\theta$	$\pi/2$	$\pi/4$	$\pi/8$	$\pi/16$	$\pi/32$
$\beta_2 = 0$	9.448E-2	1.203E-2	5.059E-4	4.981E-7	6.987E-13
$\beta_2 = 10$	0.3351	1.868E-2	4.408E-4	3.078E-7	7.597E-13
$\beta_2 = 50$	0.8577	8.609E-2	2.221E-3	1.527E-6	1.059E-12
$\beta_2 = 100$	1.1345	0.1994	9.415E-3	1.008E-5	3.553E-11

TABLE 4.2
Discretization error $\|\psi(t) - \psi^{(\Delta r, \Delta\theta, \Delta t)}(t)\|_{l^2}$ *at* $t = 2.0$ *in the* r*-direction.*

Mesh size Δr	1/32	1/64	1/128	1/256	1/512
$\beta_2 = 0$	2.716E-4	6.771E-5	1.673E-5	3.983E-6	7.968E-7
$\beta_2 = 10$	6.349E-3	1.586E-3	3.921E-4	9.337E-5	1.868E-5
$\beta_2 = 50$	0.1118	2.959E-2	7.358E-3	1.753E-3	3.507E-4
$\beta_2 = 100$	0.5203	0.1840	4.734E-2	1.131E-2	2.263E-3

TABLE 4.3
Discretization error $\|\psi(t) - \psi^{(\Delta r, \Delta\theta, \Delta t)}(t)\|_{l^2}$ *at* $t = 2.0$ *in time.*

Time step Δt	1/160	1/320	1/640	1/1280	1/2560
$\beta_2 = 0$	7.812E-5	1.952E-5	4.864E-6	1.201E-6	2.856E-7
$\beta_2 = 10$	2.236E-3	5.582E-4	1.391E-4	3.433E-5	8.155E-6
$\beta_2 = 50$	0.1111	3.581E-2	9.394E-3	2.328E-3	5.531E-4
$\beta_2 = 100$	0.5445	0.3044	0.1032	2.654E-2	6.319E-3

Next, we test the second-order accuracy in time. Table 4.3 lists the errors at $t = 2.0$ for different values of β_2 and time steps Δt with a very fine mesh in space, e.g., $\Delta r = \frac{1}{1024}$ and $\Delta\theta = \frac{\pi}{128}$.

From Tables 4.1–4.3, we can conclude that our method is of spectral-order accuracy in the θ-direction, second-order accuracy in time, and second-order accuracy in the r-direction when the piecewise linear FEM is used. Usually, for given parameter's setup and initial data, the bigger the β_2, the larger the errors. This implies that more grid points and a small time step should be used when β_2 is larger in order to get high accuracy. Furthermore, additional numerical experiments have been tested to verify the fourth-order accuracy in the r-direction when the continuous piecewise cubic element space is used. Such cubic elements are always used for the U^h in the following simulations.

4.2. Dynamics of a stationary state with a shifted center. To verify the analytical solution (2.29) and to study the dynamics of a stationary state with a shifted center through the direct simulation of the GPE for the rotating BEC, we take $d = 2$, $\lambda = 0$, $\gamma_x = \gamma_y = 1$, $\beta_2 = 100$, and $W(\mathbf{x}, t) \equiv 0$ in (2.44). The initial condition in (2.45) is taken as

$$\psi_0(\mathbf{x}) = \phi_e(\mathbf{x} - \mathbf{x}_0), \quad \phi_e(\mathbf{x}) = f(r)e^{i\theta}, \quad \mathbf{x} \in \mathbb{R}^2,$$

where $\phi_e(\mathbf{x})$ is a central vortex state with winding number $m = 1$ under the same parameter set; i.e., $f(r)$ is found numerically by the methods proposed in [5, 13]. This setup corresponds to a shift of the trap center from the origin to $-\mathbf{x}_0$. We take $\mathbf{x}_0 = (1, 1)^T$, $R = 12$ for $\Omega_{\mathbf{x}}$, mesh size $\Delta r = 0.004$, $\Delta\theta = \frac{\pi}{64}$, and time step $\Delta t = 0.0001$. The trajectory and position of the central vortex center with respect to time t are shown in Figure 1 for different rotation speed Ω. Notice that for the parameters chosen, the characteristic roots of (2.43) are given by $\pm(|\Omega| \pm 1)i$. When $\Omega = 0$ or $\Omega = \pm 1$, we get roots with higher multiplicities; otherwise, we have four distinct pure imaginary roots, and the periodicity of the orbits is thus implied for rational values of the frequency Ω. For $\Omega = 0$, (2.31)–(2.33) reduces to $\ddot{\mathbf{x}}(t) + \mathbf{x}(t) = 0$ with $\dot{\mathbf{x}}(0) = 0$. It is easy to see that the trajectory is a straight line. For $\Omega = \pm 1$, (2.31)–(2.33) reduces to $\ddot{\mathbf{x}}(t) \mp 2J\dot{\mathbf{x}}(t) = 0$, which leads to $\dot{\mathbf{x}}(t) = Q(\pm 2t)\dot{\mathbf{x}}(0)$ with $Q(2t)$ being a rotation of angle $\pm 2t$. The trajectory thus stays as a circle. In addition,

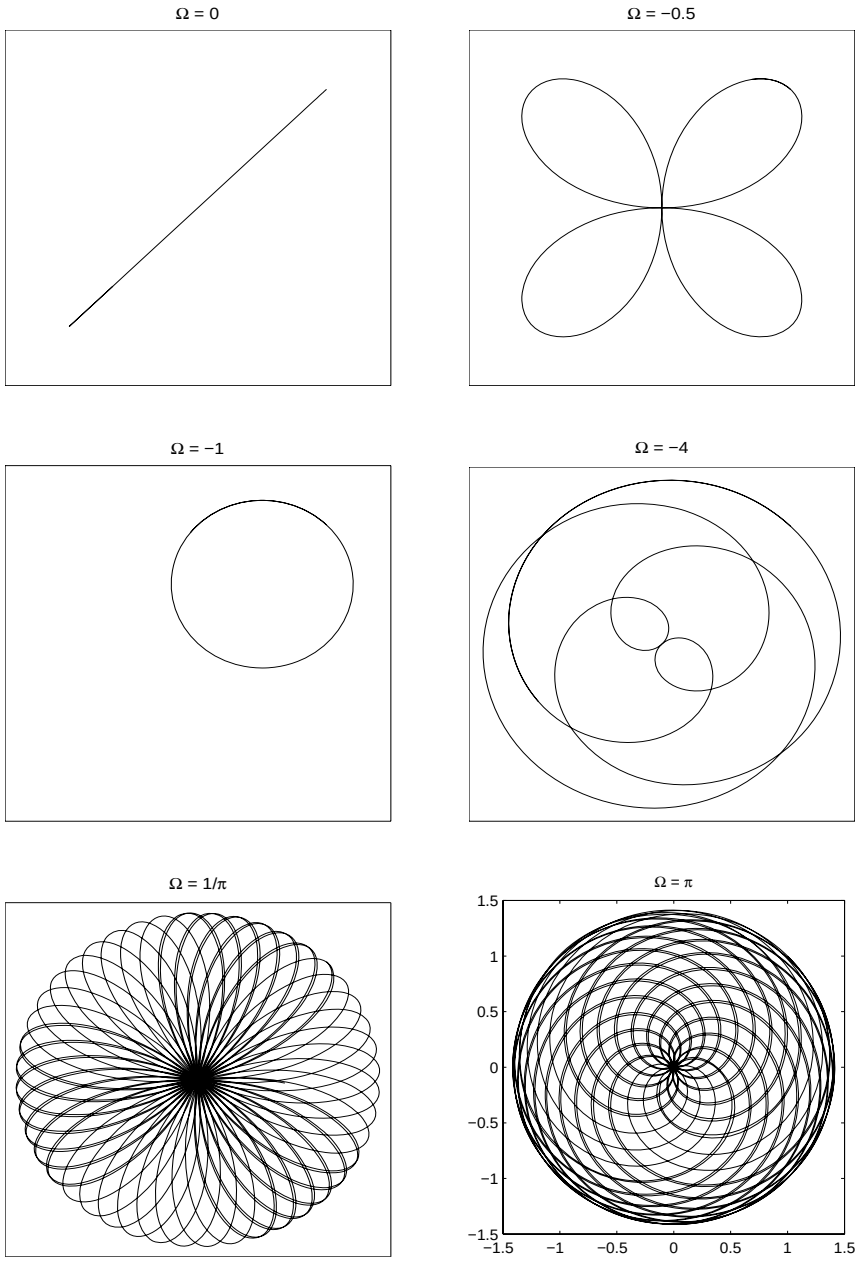


FIG. 1. *Trajectory of the central vortex center,* $\mathbf{x}(t) = (x(t), y(t))^T$ *for* $0 \le t \le 100$*, for different rotation speed* Ω*.*

for all other values of Ω, we can check that the equation is invariant under the rotation transformation, due to the fact that $Q(\theta)J = JQ(\theta)$ for any rotation matrix $Q(\theta)$. Thus, if $\Omega \neq 0, \pm 1$, and Ω is a rational number, there always exists a time t such that $e^{\pm i(\Omega \pm 1)t} = -1$; the trajectory thus always has the inversion symmetry (with respect to the origin). Other symmetries may also be explored for special values of Ω.

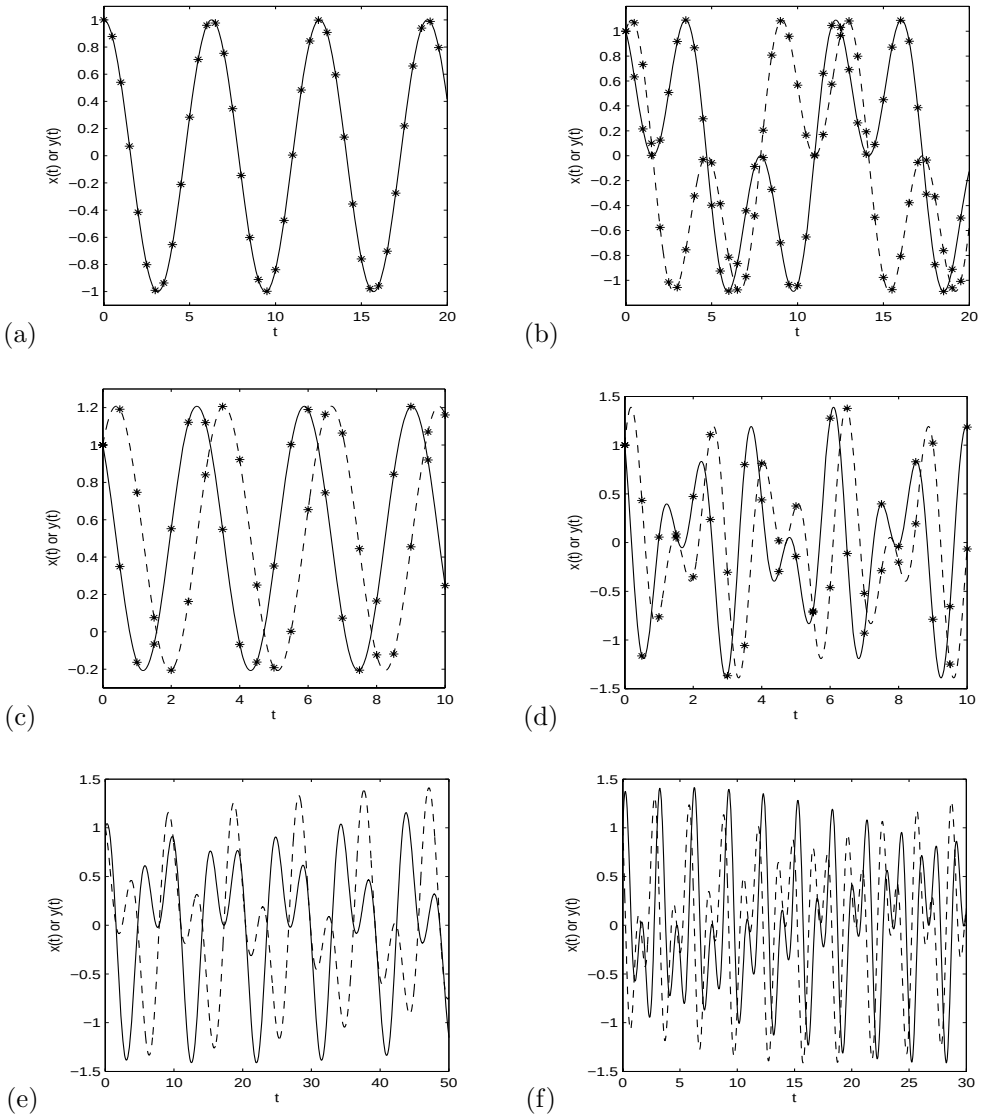


FIG. 1. (cont'd)*: Coordinates of the trajectory* $\mathbf{x}(t) = (x(t), y(t))^T$ *(solid line:* $x(t)$*; dashed line:* $y(t)$*).* (a): $\Omega = 0$*;* (b): $\Omega = -0.5$*;* (c): $\Omega = -1$*;* (d): $\Omega = -4$*;* (e): $\Omega = 1/\pi$*;* (f): $\Omega = \pi$*. In* (a)–(d)*, the solid and dashed lines are obtained from solving the GPE* (1.4)*, where "*" *is obtained from solving the ODE* (2.31)–(2.33)*.*

From Figure 1, we indeed see that when $\Omega = 0$, the center moves like a pendulum with period $T = 2\pi$; when $\Omega = -1$, it moves along a circle with period $T = \pi$. For other cases, the trajectory curve has inversion symmetry as predicted through the theoretical analysis. The solution trajectory and the coordinates in Figure 1 are obviously consistent with the above description of the solutions of the ODE system (2.31)–(2.33) for any given Ω. This provides a numerical verification of the exact solution constructed earlier for the GPE with an angular momentum rotation term and the reliability of our numerical scheme. Furthermore, based on Figure 1 and additional numerical experiments conducted, we find the following: (i) When Ω is a

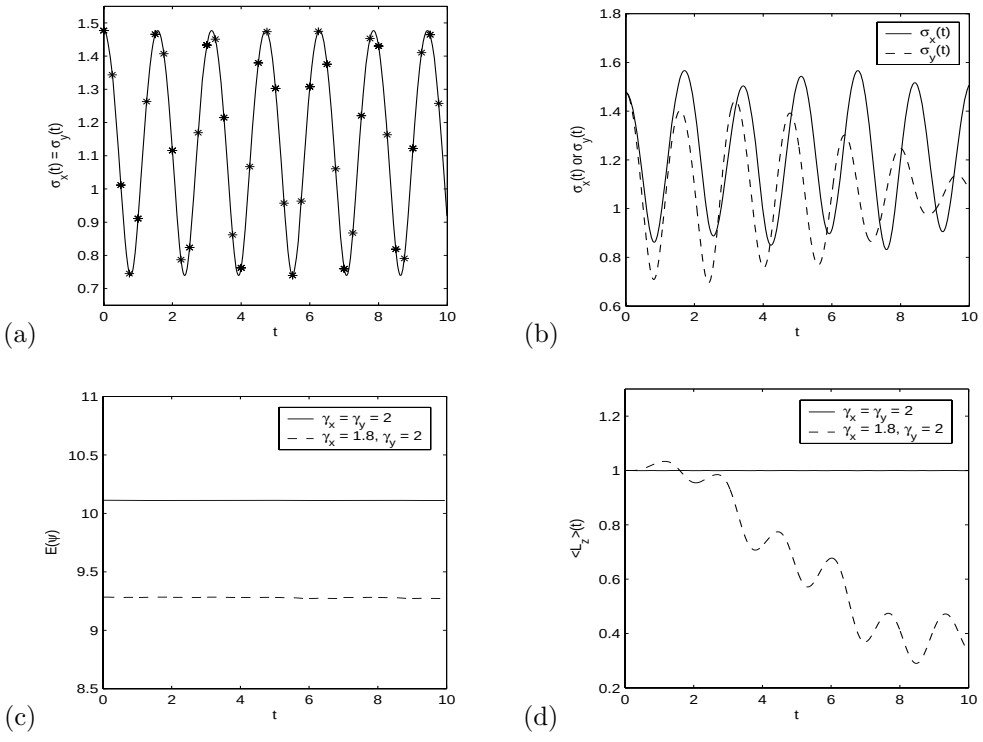


FIG. 2. *Time evolution of condensate widths (cf.* (a)*:* $\gamma_x = \gamma_y = 2$*;* (b)*:* $\gamma_x = 1.8$ *and* $\gamma_y = 2$*), energy* $E(\psi) := E_{\beta,\Omega}(\psi)$ *(cf.* (c)*) and angular momentum expectation (cf.* (d)*). In* (a)*, the solid line is obtained from solving the GPE* (1.4) *and "*" is obtained from the analytical solution* (2.15)*.*

rational number, i.e., $|\Omega| = q/p$ with q and p nonnegative integers and no common factor, then the trajectory of the vortex center moves periodically with period $p\pi$ if q and p are odd integers, and $2p\pi$ otherwise. (ii) When Ω is an irrational number, the trajectory of the vortex center moves chaotically, but the envelope of the trajectory is a circle centered at the origin with radius $r = |\mathbf{x}_0| = \sqrt{2}$ in our example (cf. Figure 1).

4.3. Dynamics of condensate width, energy, and angular momentum expectation. To verify the conservation of energy and angular momentum expectation as well as dynamics of condensate width, we take $d = 2$, $\lambda = 0$, $\beta_2 = 100$, $\Omega = 0.8$, and $W(\mathbf{x}, t) \equiv 0$ in (2.44). The initial condition in (2.45) is taken as the central vortex state with winding number $m = 1$ of the GPE with $\gamma_x = \gamma_y = 1$ [5, 14, 13], which is computed numerically by the method proposed in [5, 13]. Then at $t = 0$, we change the trap frequency by setting $\gamma_x = \gamma_y = 2$, or $\gamma_x = 1.8$ and $\gamma_y = 2$, respectively. Figure 2 shows the time evolution of condensate widths $\sigma_x(t)$ and $\sigma_y(t)$, the energy $E_{\beta,\Omega}(\psi)$, and the angular momentum expectation.

From Figure 2, we can see that (i) the condensate widths $\sigma_x(t)$ and $\sigma_y(t)$ are periodic functions of period $T = \pi/2$ when $\gamma_x = \gamma_y = 2$ (cf. Figure 2(a)) and periodic functions of period $T = \pi/2$ with a perturbation when $1.8 = \gamma_x \neq \gamma_y = 2$ (cf. Figure 2(b)), again confirming the analytical results (2.16) and (2.17), respectively; (ii) the energy $E_{\beta,\Omega}(\psi)$ is conserved in the discretized level (cf. Figure 2(c)); (iii) the angular momentum expectation is conserved when $\gamma_x = \gamma_y$ (cf. Figure 2(d) and the analytical result (2.3)). Furthermore, when $\gamma_x \neq \gamma_y$ and the initial condition is chosen as a

central vortex state with winding number $m = 1$, the angular momentum expectation is no longer conserved (cf. Figure 2(d)). We note that in the literature, there have been more studies both analytically and numerically on the thermodynamic stability of the central vortex state, though there is not much discussion available on the dynamic stability in real time. The experimental results shown here are thus of interest.

4.4. Dynamics of a quantized vortex lattice. Now we present the simulation results, via the algorithm discussed here, on the dynamics of a vortex lattice in rotating BEC under an anisotropic external perturber. We take $d = 2$, $\lambda = 0$, $\beta_2 = 1000$, $\gamma_x = \gamma_y = 1 := \gamma_r$, and $\Omega = 0.9$ in (2.44). The initial condition in (2.45) is taken as the ground state [13, 1] of the GPE with $W(\mathbf{x}, t) \equiv 0$, which is computed numerically by the normalized gradient flow with the backward Euler finite difference discretization proposed in [13]. For $t \geq 0$, an external perturber is introduced; i.e., $W(\mathbf{x}, t)$ in (2.44) is chosen as

$$W(\mathbf{x}, t) = \frac{\varepsilon}{2}\gamma_r^2 \left[(x^2 - y^2)\cos(2\tilde{\Omega}t) + 2xy\sin(2\tilde{\Omega}t)\right], \quad \mathbf{x} \in \mathbb{R}^2, \quad t \geq 0.$$

This implies the total potential $V(\mathbf{x}, t)$ in (2.44) is taken as

$$V(\mathbf{x}, t) = \frac{1}{2}\gamma_r^2 \left[(1+\epsilon)X(t)^2 + (1-\epsilon)Y(t)^2\right],$$

where $X(t) = x\cos(\tilde{\Omega}t) + y\sin(\tilde{\Omega}t)$, $Y(t) = y\cos(\tilde{\Omega}t) - x\sin(\tilde{\Omega}t)$.

This kind of time-dependent potential was used in [46] for studying the dynamics of nonrotating BEC. In our computation, we take $\epsilon = 0.35$, $\tilde{\Omega} = 0.75$, $R = 30$ for $\Omega_{\mathbf{x}}$, mesh size $\Delta r = 0.0075$ and $\Delta\theta = \frac{\pi}{128}$, and time step $\Delta t = 0.0001$. Figure 3 shows contour plots of the density function $|\psi(\mathbf{x}, t)|^2$ at different time steps.

For Figure 3, at $t = 0$, there are about 45 quantized vortices in the ground state. During the time evolution, the lattice is rotated due to the angular momentum term with different lattice patterns being formed due to the anisotropic external stirrer $W(\mathbf{x}, t)$. One may compare our numerical results with the experimental observations in [23], where the anisotropic compression of the vortex lattices was observed due to the dynamic distortion of the trap potentials.

4.5. Stability of central vortex states. Similarly as in [14, 16, 29, 30] for nonrotating BEC, we hereby also study numerically the stability of central vortex states in rotating BEC. We take $d = 2$, $\gamma_x = \gamma_y = 1$, $\beta_2 = 100$, $\Omega = -0.8$, and $\lambda = 0$ in (2.44). The initial condition in (2.45) is taken as a central vortex state [14, 5, 13] with winding number m of the GPE with $W(\mathbf{x}, t) \equiv 0$; i.e., $\psi_0(\mathbf{x}) = f_m(r)e^{im\theta}$, where $f_m(r)$ is computed numerically by the method proposed in [5, 14]. In order to study the stability, when $t \in [0, \pi/2]$, we introduce a far-blue detuned Gaussian laser beam stirrer (2.47), and when $t \geq \pi/2$, the perturber is removed. The parameters in (2.47) are chosen as

$$(x_s(t), y_s(t)) \equiv (3, 0), \quad \omega_s = 1, \quad W_s(t) = \begin{cases} 5\sin^2(2t), & t \in [0, \pi/2], \\ 0, & t \geq \pi/2. \end{cases}$$

In our computation, we take $R = 12$ for $\Omega_{\mathbf{x}}$, mesh size $\Delta r = 0.004$ and $\Delta\theta = \frac{\pi}{64}$, and time step $\Delta t = 0.0001$.

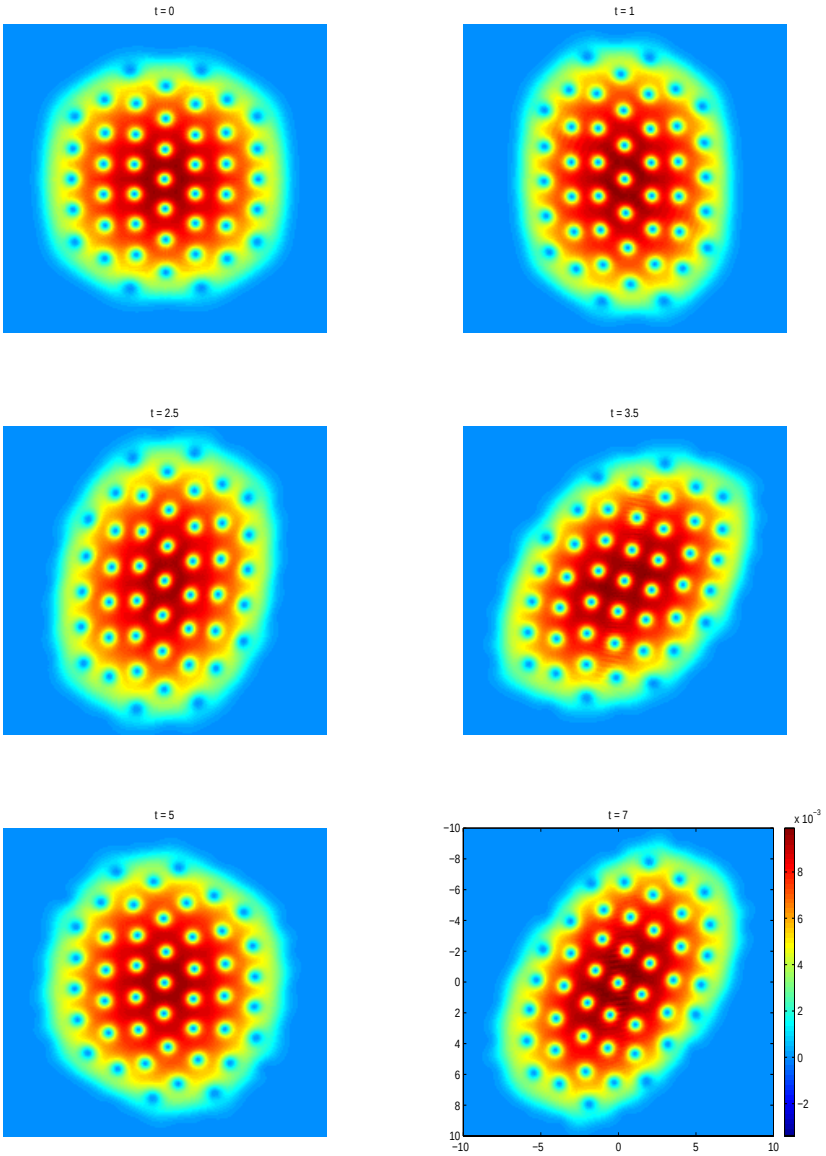


FIG. 3. *Contour plots of the density function* $|\psi(\mathbf{x}, t)|^2$ *for dynamics of a vortex lattice at different times.*

To quantitatively analyze the numerical results, we define the hydrodynamic velocity as

$$\mathbf{u} = (u, v) = \mathrm{Im}(\psi^* \nabla \psi)/|\psi|^2.$$

Figure 4 shows the velocity fields during the time evolution of the central vortex states with winding number $m = 1$ and $m = 2$, while the dynamic evolution of the energy and that of the angular momentum expectation are shown in Figure 5.

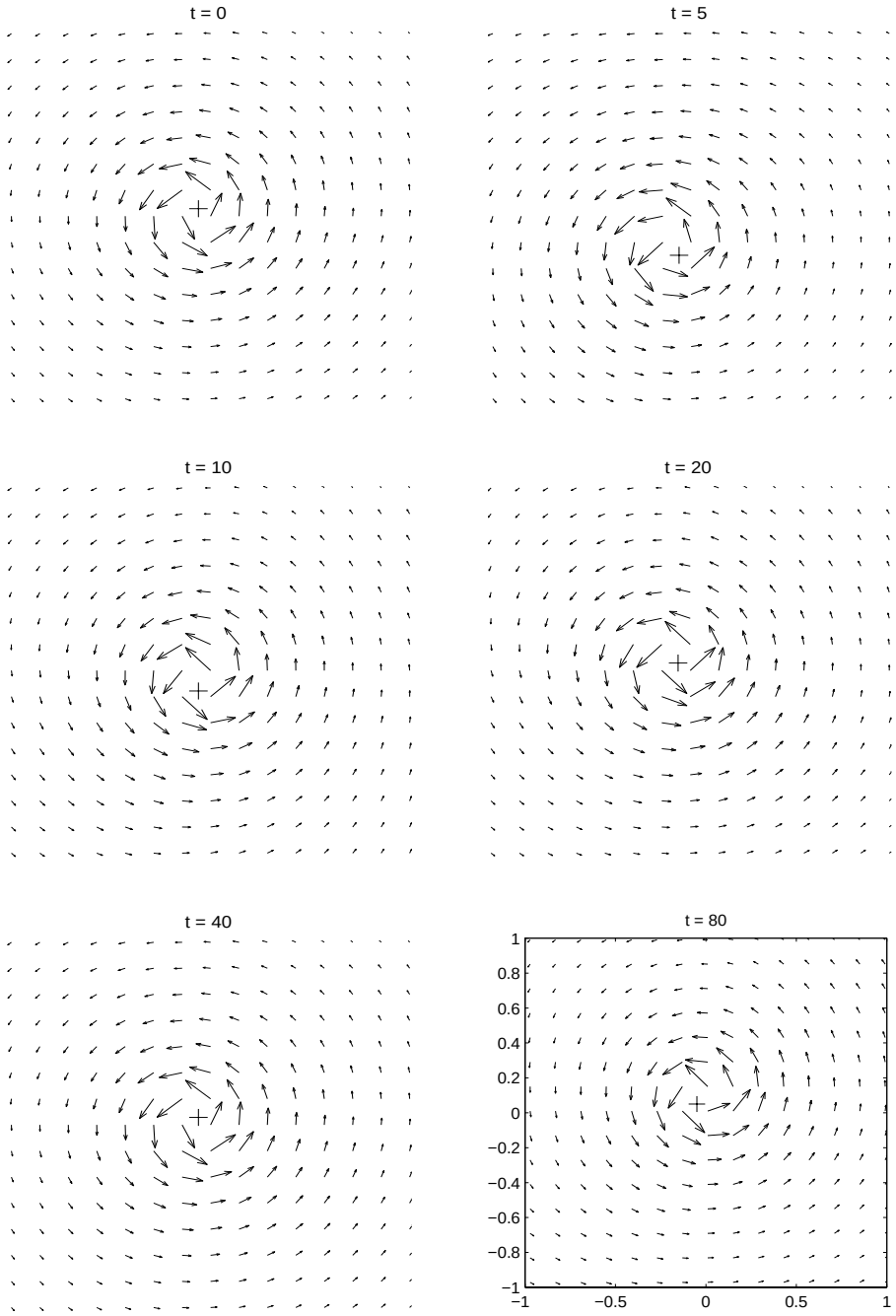


FIG. 4. *Velocity field at different times for stability of a central vortex state.* I. *For winding number* $m = 1$.

From Figure 4 and additional numerical experiments conducted, we find that the central vortex states with an index (or degree, winding number) $m = \pm 1$ are dynamically stable, but they are unstable when $|m| > 1$ in rotating BEC. Furthermore, Figure 5 depicts the increase in the energy and the decrease of the angular momentum

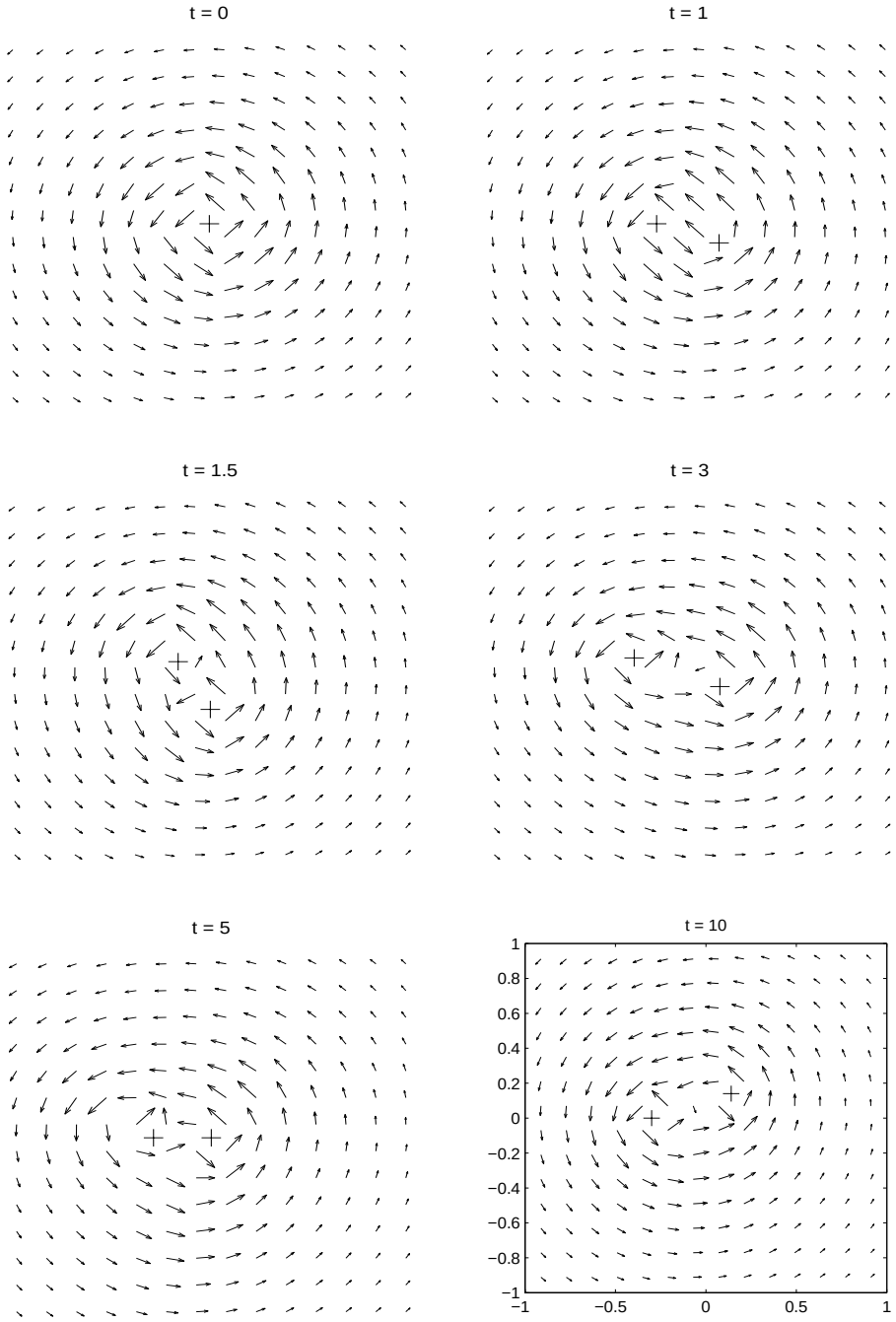
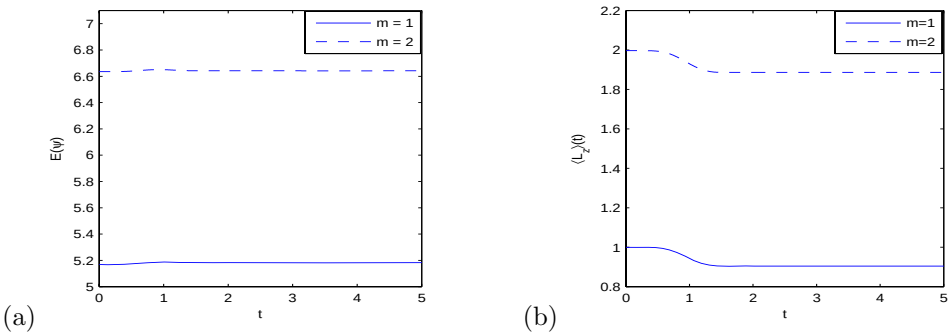


FIG. 4. (cont'd): II. *For* $m = 2$.

expectation when $t \in [0, \pi/2]$ due to the appearance of the perturber. After removing the perturber at $t = \pi/2$, they are conserved with time, which again confirm the conservation laws (1.8) and (2.3).



(a) (b)

FIG. 5. *Time evolution of energy* $E(\psi) := E_{\beta,\Omega}(\psi)$ *and angular momentum expectation* $\langle L_z \rangle$ *in studying the stability of central vortex states.*

4.6. Dissipation effect on the GPE. In order to study the effect of the damping mechanism in the GPE (2.44), we take $d = 2$, $\gamma_x = 1$, $\gamma_y = 1.1$, $\beta_2 = 500$, $\Omega = 0.9$, and $W(\mathbf{x}, t) \equiv 0$ in (2.44). The initial condition in (2.45) is taken as

$$\psi_0(\mathbf{x}) = \frac{(\gamma_x \gamma_y)^{1/4}}{\sqrt{\pi}} \, e^{-(\gamma_x x^2 + \gamma_y y^2)/2}, \qquad \mathbf{x} \in \mathbb{R}^2.$$

We take $R = 30$ for $\Omega_{\mathbf{x}}$, mesh size $\Delta r = 0.0075$ and $\Delta\theta = \frac{\pi}{64}$, and time step $\Delta t = 0.0001$. Figure 6 shows the normalized density $\frac{|\psi(\mathbf{x},t)|^2}{\|\psi(\cdot,t)\|^2}$ at different times for $\lambda = 0.03$, while Figure 7 illustrates the time evolution of the energy and angular momentum expectation per particle, i.e., $\frac{E_{\beta,\Omega}(\psi)}{\|\psi(\cdot,t)\|^2}$ and $\frac{\langle L_z \rangle}{\|\psi(\cdot,t)\|^2}$ for different $\lambda > 0$.

From Figures 6 and 7, we can see that when a dissipation term is applied to the GPE, a dent appears in the center of the density function during time evolution. The larger the damping parameter λ, the faster the energy per particle decreases and the slower the angular momentum expectation per particle increases. In fact, the change in the angular momentum expectation is due to the anisotropy of the external trapping potential, i.e., $\gamma_x \neq \gamma_y$.

5. Conclusion. We have studied the dynamics of the Gross–Pitaevskii equation with an angular momentum rotation term for rotating BEC both analytically and numerically. Along the analytical front, we proved the conservation of the angular momentum expectation when the external trapping potential is radially symmetric in 2D and, respectively, cylindrically symmetric in 3D. A second-order ODE was also derived to describe the time evolution of the condensate width as a periodic function with/without a perturbation, and the frequency of the periodic function doubles the trapping frequency. We also presented an ODE system with a complete initial data that governs the dynamics of a stationary state with a shifted center, and we also illustrated the decrease in the total density when a damping term is applied in the GPE. On the numerical side, we proposed an efficient, accurate, and unconditionally stable numerical method for simulating the rotating BEC with/without a time-dependent external perturber or a damping term. We also applied the new method to study numerically the dynamics of condensate including the condensate widths, energy, and angular momentum expectation as well as a quantized vortex lattice and a stationary state with a shifted center. We numerically found that, for the

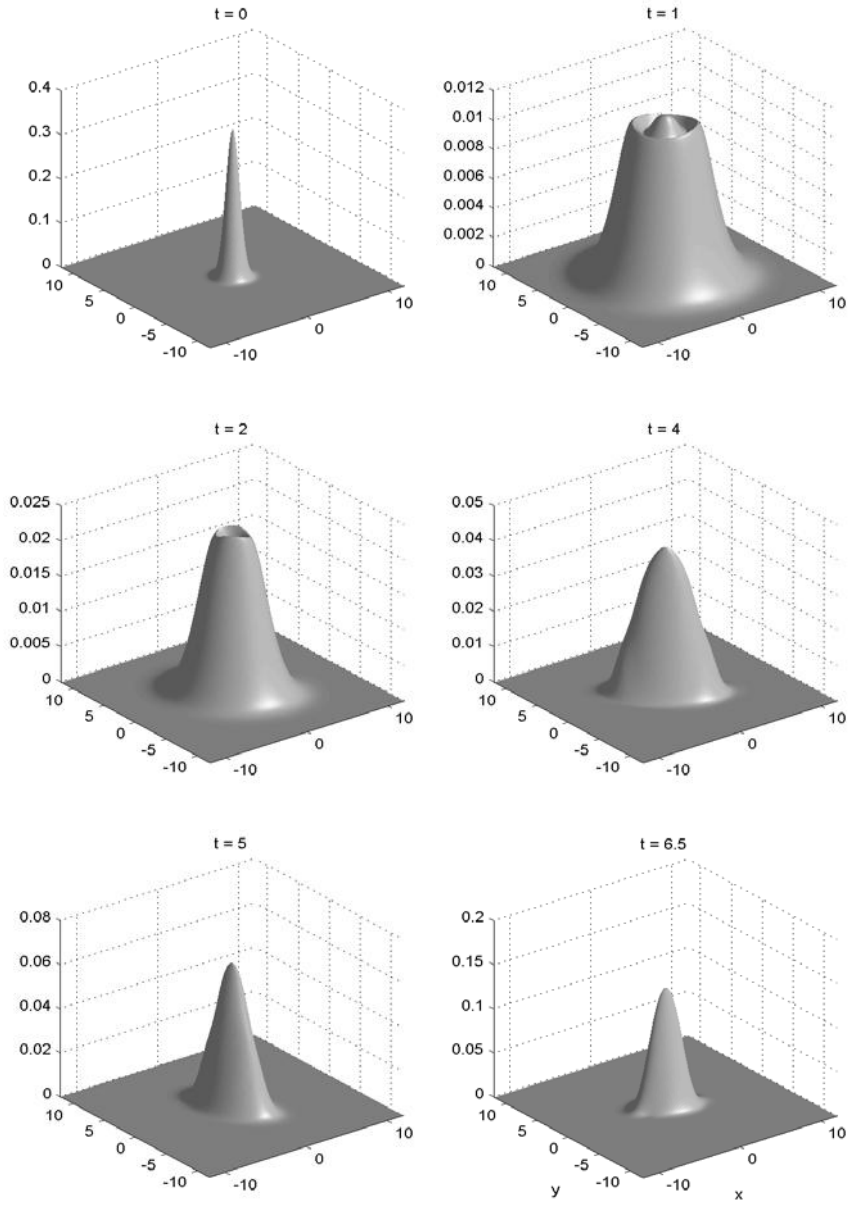


FIG. 6. *Surface plots of the normalized density function* $\frac{|\psi(\mathbf{x},t)|^2}{\|\psi(\cdot,t)\|^2}$ *in section* 4.6 *for GPE with a damping term at different times.*

real time dynamics, the central vortex states are dynamically stable only for the one with index (or winding number) $m = \pm 1$. In the future, this efficient and accurate numerical method can be used to study the dynamics and interaction of vortex line states in 3D for rotating BEC and to make more close comparisons with experimental findings [42, 47, 48].

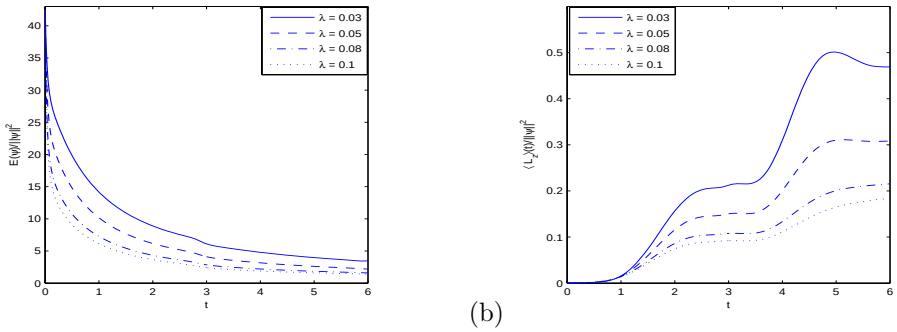


FIG. 7. *Time evolution of the energy per particle* $\frac{E_{\beta,\Omega}(\psi)}{\|\psi\|^2}$ *and angular momentum expectation* $\frac{\langle L_z\rangle}{\|\psi\|^2}$ *in section* 4.6 *for the GPE with a damping term.*

REFERENCES

[1] A. AFTALION AND Q. DU, *Vortices in a rotating Bose–Einstein condensate: Critical angular velocities and energy diagrams in the Thomas–Fermi regime*, Phys. Rev. A, 64 (2001), 063603.

[2] A. AFTALION, Q. DU, AND Y. POMEAU, *Dissipative flow and vortex shedding in the Painlevé boundary layer of a Bose–Einstein condensate*, Phy. Rev. Lett., 91 (2003), 090407.

[3] M. H. ANDERSON, J. R. ENSHER, M. R. MATTHEWA, C. E. WIEMAN, AND E. A. CORNELL, *Observation of Bose–Einstein condensation in a dilute atomic vapor*, Science, 269 (1995), pp. 198–201.

[4] W. BAO, *Ground states and dynamics of multicomponent Bose–Einstein condensates*, Multiscale Model. Simul., 2 (2004), pp. 210–236.

[5] W. BAO AND Q. DU, *Computing the ground state solution of Bose–Einstein condensates by a normalized gradient flow*, SIAM J. Sci. Comput., 25 (2004), pp. 1674–1697.

[6] W. BAO AND D. JAKSCH, *An explicit unconditionally stable numerical method for solving damped nonlinear Schrödinger equations with a focusing nonlinearity*, SIAM J. Numer. Anal., 41 (2003), pp. 1406–1426.

[7] W. BAO, D. JAKSCH, AND P. A. MARKOWICH, *Numerical solution of the Gross–Pitaevskii equation for Bose–Einstein condensation*, J. Comput. Phys., 187 (2003), pp. 318–342.

[8] W. BAO, D. JAKSCH, AND P. A. MARKOWICH, *Three dimensional simulation of jet formation in collapsing condensates*, J. Phys. B: At. Mol. Opt. Phys., 37 (2004), pp. 329–343.

[9] W. BAO, S. JIN, AND P. A. MARKOWICH, *On time-splitting spectral approximation for the Schrödinger equation in the semiclassical regime*, J. Comput. Phys., 175 (2002), pp. 487–524.

[10] W. BAO, S. JIN, AND P. A. MARKOWICH, *Numerical study of time-splitting spectral discretizations of nonlinear Schrödinger equations in the semiclassical regimes*, SIAM J. Sci. Comput., 25 (2003), pp. 27–64.

[11] W. BAO AND J. SHEN, *A fourth-order time-splitting Laguerre–Hermite pseudospectral method for Bose–Einstein condensates*, SIAM J. Sci. Comput., 26 (2005), pp. 2010–2028.

[12] W. BAO AND W. TANG, *Ground state solution of trapped interacting Bose–Einstein condensate by directly minimizing the energy functional*, J. Comput. Phys., 187 (2003), pp. 230–254.

[13] W. BAO, H. WANG, AND P. A. MARKOWICH, *Ground, symmetric and central vortex states in rotating Bose–Einstein condensates*, Commun. Math. Sci., 3 (2005), pp. 57–88.

[14] W. BAO AND Y. ZHANG, *Dynamics of the ground state and central vortex states in Bose–Einstein condensation*, Math. Models Methods Appl. Sci., 15 (2005), pp. 1863–1896.

[15] I. BIALYNICKI-BIRULA AND Z. BIALYNICKI-BIRULA, *Center-of-mass motion in the many-body theory of Bose–Einstein condensates*, Phys. Rev. A, 65 (2002), 063606.

[16] B. M. CARADOC-DAVIS, R. J. BALLAGH, AND K. BURNETT, *Coherent dynamics of vortex formation in trapped Bose–Einstein condensates*, Phys. Rev. Lett., 83 (1999), pp. 895–898.

[17] Y. CASTIN AND R. DUM, *Bose-Einstein condensates with vortices in rotating traps*, Eur. Phys. J. DAt. Mol. Opt. Phys., 7(1999), pp. 399–412.

[18] M. M. Cerimele, M. L. Chiofalo, F. Pistella, S. Succi, and M. P. Tosi, *Numerical solution of the Gross–Pitaevskii equation using an explicit finite-difference scheme: An application to trapped Bose–Einstein condensates*, Phys. Rev. E, 62 (2000), pp. 1382–1389.

[19] M. M. Cerimele, F. Pistella, and S. Succi, *Particle-inspired scheme for the Gross–Pitaevskii equation: An application to Bose–Einstein condensation*, Comput. Phys. Comm., 129 (2000), pp. 82–90.

[20] F. Dalfovo, S. Giorgini, L. P. Pitaevskii, and S. Stringari, *Theory of Bose–Einstein condensation in trapped gases*, Rev. Mod. Phys., 71 (1999), pp. 463–512.

[21] K. B. Davis, M. O. Mewes, M. R. Andrews, N. J. van Druten, D. S. Durfee, D. M. Kurn, and W. Ketterle, *Bose–Einstein condensation in a gas of sodium atoms*, Phys. Rev. Lett., 75 (1995), pp. 3969–3973.

[22] Q. Du, *Numerical computations of quantized vortices in Bose–Einstein condensate*, in Recent Progress in Computational and Applied PDEs, T. F. Chan, Y. Huang, T. Tang, J. Xu, and L.-A. Ying, eds., Kluwer Academic Publishers, Boston, Dordrecht, London, 2002, pp. 155–168.

[23] P. Engels, I. Coddington, P. Haijan, and E. Cornell, *Nonequilibrium effects of anisotropic compression applied to vortex lattices in Bose–Einstein condensates*, Phy. Rev. Lett, 89 (2002), 100403.

[24] D. L. Feder, C. W. Clark, and B. I. Schneider, *Nucleation of vortex arrays in rotating anisotropic Bose–Einstein condensates*, Phys. Rev. A, 61 (1999), 011601.

[25] D. L. Feder, C. W. Clark, and B. I. Schneider, *Vortex stability of interacting Bose–Einstein condensates confined in anisotropic harmonic traps*, Phys. Rev. Lett., 82 (1999), pp. 4956–4959.

[26] D. L. Feder, A. A. Svidzinsky, A. L. Fetter, and C. W. Clark, *Anomalous modes drive vortex dynamics in confined Bose–Einstein condensates*, Phys. Rev. Lett., 86 (2001), pp. 564–567.

[27] J. J. Garcia-Ripoll, V. M. Perez-Garcia, and V. Vekslerchik, *Construction of exact solutions by spatial translations in inhomogeneous nonlinear Schrödinger equations*, Phys. Rev. E, 64 (2001), 056602.

[28] R. Glowinski and P. Le Tallec, *Augmented Lagrangian and Operator Splitting Methods in Nonlinear Mechanics*, SIAM Stud. Appl. Math. 9, SIAM, Philadelphia, 1989.

[29] B. Jackson, J. F. McCann, and C. S. Adams, *Vortex formation in dilute inhomogeneous Bose–Einstein condensates*, Phys. Rev. Lett., 80 (1998), pp. 3903–3906.

[30] B. Jackson, J. F. McCann, and C. S. Adams, *Dissipation and vortex creation in Bose–Einstein condensate gases*, Phys. Rev. A, 61 (2000), 051603.

[31] D. Jaksch, C. Bruder, J. I. Cirac, C. W. Gardiner, and P. Zoller, *Cold bosonic atoms in optical lattices*, Phys. Rev. Lett., 81 (1998), pp. 3108–3111.

[32] A. M. Kamchatnov, A. Gammal, and R. A. Kraenkel, *Dissipationless shock waves in repulsive Bose–Einstein condensates*, Phys. Rev. A, 69 (2004), 063605.

[33] K. Kasamatsu, M. Tsubota, and M. Ueda, *Nonlinear dynamics of vortex lattice formation in a rotating Bose–Einstein condensate*, Phys. Rev. A, 67 (2003), 033610.

[34] M.-C. Lai, W.-W. Lin, and W. Wang, *A fast spectral/difference method without pole conditions for Poisson-type equations in cylindrical and spherical geometries*, IMA J. Numer. Anal., 22 (2002), pp. 537–548.

[35] M.-C. Lai and W.-C. Wang, *Fast direct solvers for Poisson equation on 2D polar and spherical geometries*, Numer. Methods Partial Differential Equations, 18 (2002), pp. 56–68.

[36] K. W. Madison, F. Chevy, W. Wohlleben, and J. Dalibard, *Vortex formation in a stirred Bose–Einstein condensate*, Phys. Rev. Lett., 84 (2000), pp. 806–809.

[37] K. W. Madison, F. Chevy, V. Bretin, and J. Dalibard, *Stationary states of a rotating Bose–Einstein condensate: Routes to vortex nucleation*, Phys. Rev. Lett., 86 (2001), pp. 4443–4446.

[38] G. Marchuk, *Splitting and alternating direction methods*, in Handbook of Numerical Analysis, North–Holland, Amsterdam, 1990.

[39] M. R. Matthews, B. P. Anderson, P. C. Haljan, D. S. Hall, C. E. Wiemann, and E. A. Cornell, *Vortices in a Bose–Einstein condensate*, Phys. Rev. Lett., 83 (1999), pp. 2498–2501.

[40] A. Minguzzi, S. Succi, F. Toschi, M. P. Tosi, and P. Vignolo, *Numerical methods for atomic quantum gases with applications to Bose–Einstein condensates and to ultracold fermions*, Phys. Rep., 395 (2004), pp. 223–355.

[41] P. Muruganandam and S. K. Adhikari, *Bose–Einstein condensation dynamics in three dimensions by pseudospectral and finite-difference methods*, J. Phys. B: At. Mol. Opt. Phys., 36 (2003), pp. 2501–2513.

[42] A. A. Penckwitt, R. J. Ballagh, and C. W. Gardiner, *Nucleation, growth, and stabilization of Bose–Einstein condensate vortex lattices*, Phys. Rev. Lett., 89 (2002), 260402.
[43] C. J. Pethick and H. Smith, *Bose–Einstein Condensation in Dilute Gases*, Cambridge University Press, Cambridge, UK, 2002.
[44] L. P. Pitaevskii and S. Stringari, *Bose–Einstein Condensation*, Clarendon Press, Oxford, UK, 2003.
[45] C. Raman, J. R. Abo-Shaeer, J. M. Vogels, K. Xu, and W. Ketterle, *Vortex nucleation in a stirred Bose–Einstein condensate*, Phys. Rev. Lett., 87 (2001), 210402.
[46] B. I. Schneider and D. L. Feder, *Numerical approach to the ground and excited states of a Bose–Einstein condensed gas confined in a completely anisotropic trap*, Phys. Rev. A, 59 (1999), pp. 2232–2242.
[47] T. P. Simula, P. Engels, I. Coddington, V. Schweikhard, E. A. Cornell, and R. J. Ballagh, *Observations on sound propagation in rapidly rotating Bose–Einstein condensates*, Phys. Rev. Lett., 94 (2005), 080404.
[48] T. P. Simula, A. A. Penckwitt, and R. J. Ballagh, *Giant vortex lattice deformations in rapidly rotating Bose–Einstein condensates*, Phys. Rev. Lett., 92 (2004), 060401.
[49] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), pp. 506–517.

SIAM J. APPL. MATH.
Vol. 66, No. 3, pp. 787–801

THERMOPHORETIC MOTION OF A SLIGHTLY DEFORMED SPHERE THROUGH A VISCOUS FLUID*

ARUNA MOHAN† AND HOWARD BRENNER†

Abstract. This paper provides a general approach to the solution of the problem of nonisothermal Stokes flow relative to a heat-conducting particle having the shape of a slightly deformed sphere, taking account of Maxwell's [J. C. Maxwell, *Philos. Trans. R. Soc. Lond.*, 170 (1879), pp. 231–256] thermal creep condition at the surface of the particle. The results, which are of interest in connection with the phenomenon of thermophoresis, have potential applications in aerosol technology, and in the nonisothermal transport and processing of particulate matter. For the specific case of thermally nonconducting particles, the results obtained herein accord with Morrison's [F. A. Morrison, *J. Colloid Interface Sci.*, 34 (1970), pp. 210–214] proof in the comparable electrophoretic case that the phoretic velocity of a nonconducting, force- and torque-free, nonspherical particle undergoing electrophoresis in a fluid that is otherwise at rest is independent of the size, shape and orientation of the particle, and is identical to that of a sphere.

Key words. Stokes flow, thermal creep, thermophoresis

AMS subject classifications. 35Q30, 76D07, 76N15

DOI. 10.1137/050632075

1. Introduction.

The no-slip boundary condition, conventionally applied to the velocity of a fluid at a solid surface, is known to fail when applied to a gas at its boundary with a nonuniformly heated solid. Under such nonisothermal conditions, the thermal slip (or thermal creep) condition first proposed by Maxwell [16] must instead be used. This boundary condition is given for a gas of Maxwellian molecules [7, 11] by the expression

$$\mathbf{v} - \mathbf{U} = \frac{3}{4}\frac{\mu}{\rho T}(\mathbf{I} - \hat{\mathbf{n}}\hat{\mathbf{n}}) \cdot \nabla T \tag{1}$$

on the solid boundary, where $\mathbf{v}$, $\mathbf{U}$, T, μ and ρ denote, respectively, the gas velocity, wall velocity, temperature, viscosity and density of the gas, $\mathbf{I}$, the idemfactor, and $\hat{\mathbf{n}}$ the unit normal to the surface. Because Maxwell's derivation of the thermal slip condition assumes the distribution function of the gas molecules in the Knudsen layer to be the same as that in the bulk gas, the 3/4 coefficient is subject to some uncertainty, as was pointed out by Maxwell himself. Note that (1) automatically satisfies the condition $\hat{\mathbf{n}} \cdot (\mathbf{v} - \mathbf{U}) = 0$ that the solid be impermeable to mass flow through its surface.

As discussed by Kogan [11], in order to find the correct boundary conditions to be imposed on the gas velocity at the solid surface, the Boltzmann equation must be solved both inside the Knudsen layer proximate to the surface as well as in the bulk gas. The subsequent matching of these inner and outer solutions at the outer limit of the Knudsen layer yields the tangential velocity boundary condition to be imposed on the continuum hydrodynamic equations governing the velocity in the bulk gas. The difference between the boundary condition on velocity thereby obtained

*Received by the editors May 21, 2005; accepted for publication (in revised form) October 17, 2005; published electronically February 3, 2006.
http://www.siam.org/journals/siap/66-3/63207.html
†Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 (aruna@mit.edu, hbrenner@mit.edu).

and the velocity of the wall arises due to thermal slip. Various such derivations of the thermal slip condition are summarized in [12], with the form of the resulting boundary condition being identical to Maxwell's slip condition, (1); however, the numerical coefficient appearing therein, namely, 3/4 in Maxwell's case, is found to depend on the potential of interactions between the molecules of the gas and the solid surface.

It is evident from (1) that the characteristic velocity of gas flow induced by thermal slip scales with the ratio of its kinematic viscosity $\nu = \mu/\rho$ to the length scale $L \equiv \|\nabla \ln T\|^{-1}$ of the externally imposed temperature variation, wherein the modulus bars denote an appropriate norm. Thus, when L is large compared with the characteristic size, say a, of a small particle present in the gas whose surface constitutes the solid boundary, the appropriate Reynolds number Re governing the gas flow scales as $a/L \ll 1$. As a result, the inertial term $\rho\mathbf{v} \cdot \nabla\mathbf{v}$ of the momentum equation proves to be negligible in comparison with the viscous term, $\mu\nabla^2\mathbf{v}$. Similarly, the Peclet number Pe = RePr, with the Prandtl number Pr being $O(1)$ for gases [7], scales as a/L, whence the convective term $\mathbf{v} \cdot \nabla T$ of the thermal energy equation proves to be negligible in comparison with the conduction term, $\alpha\nabla^2 T$. Although density is a function of temperature, it is readily proved, by making use of the resulting heat conduction equation, in conjunction with the fact that $\rho T = \text{const}$ for ideal gases when the pressure remains (approximately) constant throughout the gas, that the incompressible continuity equation, $\nabla \cdot \mathbf{v} = 0$, is valid. As a result, the equations governing $\mathbf{v}$ are the equations of incompressible Stokes flow, subject to the thermal creep boundary condition. The applicability of Stokes equations satisfying this boundary condition to slow, nonisothermal gas flow is discussed in [9].

In the presence of temperature gradients, the inclusion of the Burnett thermal stresses [7] alongside the Newtonian viscous stress tensor of the Navier–Stokes equations has also been found necessary [13]. However, as was first deduced by Maxwell, and elaborated in [9], when the imposed temperature gradient is sufficiently small, such that the inertial and convective terms of the respective momentum and energy equations are negligible in comparison with the viscous and conduction terms appearing therein, the contribution of such thermal stresses to the flow vanishes. In this approximation, these stresses do not contribute to the force or torque acting on the particle immersed in the gas [9].

The thermophoretic motion of a solid particle in a nonisothermal gas, arising from the thermally-induced slip of the gas at the solid surface, has applications in aerosol technology, as a method of microcontamination control in the semiconductor industry, and in the fabrication of optical fibers. It also has potential applications in microgravity manufacturing processes. All of the above applications are reviewed by Zheng [26].

The thermophoretic motion of a sphere in a gas, which occurs in the direction opposite to that of the imposed temperature gradient, was first calculated by Epstein [8] based upon Maxwell's thermal creep condition. Subsequent analyses of the problem are summarized by Zheng [26]. Most known solutions of the problem, however, deal only with the motion of spherical particles, whereas many particles encountered in practical applications are irregularly shaped. Existing theoretical studies of the thermophoresis of nonspherical particles are currently available only for axisymmetric flows, wherein the imposed temperature gradient lies parallel to the particle's axis of symmetry [24, 25]. Reference [14] provides a numerical solution for asymmetric thermophoretic flow around a two-sphere aggregate. Given this dearth of information on nonsymmetric thermophoretic particle motions, we were motivated to study a "simple" example of such motion.

The present investigation develops an asymptotic expansion of the equations of Stokes flow satisfying the thermal slip condition at the surface of a heat-conducting, arbitrarily deformed sphere, wherein the deviation from the spherical shape is small, while the imposed temperature gradient is arbitrarily oriented with respect to the particle's geometry. In principle, the asymptotic solution may be obtained correct to any order in the perturbation parameter, which measures the deviation from the spherical shape, for arbitrarily deformed particles. We here provide the explicit solution, correct to the first order, for the specific case of an ellipsoidal particle. For the special case of a force- and torque-free, thermally insulated (i.e., nonconducting) particle, it is found that the thermophoretic velocity of the particle reduces to that of a sphere moving under the same temperature gradient in a gas otherwise at rest. This result accords with the related findings of Morrison [18], namely that the phoretic velocity of an insulated particle is independent of its shape and orientation, as well as of its size. (Morrison's proof, though offered in the context of electrophoresis, is equally applicable to the thermophoretic case, provided that the particle is nonconducting, i.e., thermally insulated in the latter case.)

While the specific problem considered here is that of flow of gas around a deformed, conducting sphere under nonisothermal conditions, the results are also generally applicable to other situations for which the fluid is known to slip at the surface of the solid. Such slip conditions have been proposed in order to explain the thermally-induced motion of particles in liquids [21, 22, 23], using an approach analogous to the treatment of phoretic motion in liquids [1]. With the use of an appropriate slip coefficient, the results of the present paper may thus be applied to thermophoretic motion in liquids.

Furthermore, our results are applicable to other sources of slip occurring at solid surfaces, such as electrokinetic slip at an insulated surface, wherein the electric potential of the electrolytic liquid is governed by equations mathematically identical in form to those governing the temperature field in the present nonisothermal situation. In such circumstances, the liquid's electrokinetic slip velocity at the particle surface is proportional to the gradient of the electric potential, analogous to the corresponding thermal slip condition, (1).

The detailed solution for the case of isothermal Stokes flow around a slightly deformed sphere (subject to no slip at its surface) was presented by Brenner and his collaborators [3, 4, 10, 20]. That solution is here extended to allow for slip at the surface of the deformed sphere arising from temperature inhomogeneities in the fluid.

2. Problem formulation. Consider the incompressible Stokes flow around a slightly deformed sphere moving without rotation at a velocity of $\mathbf{U}$ through a fluid across which an otherwise uniform temperature gradient, here denoted by the space-fixed constant vector $\mathbf{G}$, has been imposed under undisturbed flow conditions, i.e., in the absence of the particle. It is assumed that the surface of the particle, S_p, is described geometrically in invariant form by the equation

$$r = a\left[1 + \epsilon \sum_n \mathbf{A}_n \boxed{n} \mathbf{P}_n(\hat{\mathbf{r}}) + O(\epsilon^2)\right], \tag{2}$$

where $\mathbf{A}_n$ is an $O(1)$ body-fixed polyadic of rank n describing the shape of the deformed body. The nth-rank polyadics

$$\mathbf{P}_n(\hat{\mathbf{r}}) = (-1)^n (n!)^{-1} r^{n+1} \overbrace{\nabla\nabla\ldots\nabla}^{n \text{ times}} (1/r)$$

are the polyadic surface harmonics of order n [4, 20], whose argument $\hat{\mathbf{r}}$ is the space-fixed unit position vector, $\hat{\mathbf{r}} = \mathbf{r}/r$, while $\epsilon \ll 1$ is a small, dimensionless quantity, inseparable from the term $\sum_n \mathbf{A}_n \boxed{n} \mathbf{P}_n(\hat{\mathbf{r}})$ describing the deformation of the surface. The $r = |\mathbf{r}|$ radial coordinate is measured from an origin situated at the center of the undeformed sphere of radius a. Since the surface spherical harmonics constitute a complete set of orthonormal polynomials in the surface coordinates (θ, ϕ), any function thereof, say $f(\hat{\mathbf{r}})$, may be expanded in invariant form in terms of the polyadics $\mathbf{P}_n(\hat{\mathbf{r}})$, analogous to the well-known scalar spherical harmonic expansion of an arbitrary function $f(\theta, \phi)$ [15]; hence, (2) is the general equation characterizing the surface of any arbitrarily shaped body whose deviation from a spherical shape is small. The symbol $\boxed{n}$ refers to n successive dot-product contraction operations performed in the order prescribed by the nesting convention of [4], wherein the n indices of the corresponding polyadics are contracted sequentially, beginning with the innermost indices. It follows that $\mathbf{A}_n \boxed{n} \mathbf{P}_n(\hat{\mathbf{r}}) = \mathbf{P}_n(\hat{\mathbf{r}}) \boxed{n} \mathbf{A}_n$, whence the surface S_p may also be equivalently described by the alternate expression

$$r = a\left[1 + \epsilon \mathbf{P}_n(\hat{\mathbf{r}}) \boxed{n} \mathbf{A}_n + O(\epsilon^2)\right], \tag{3}$$

wherein we have here invoked the summation convention, which will be used throughout, with the repeated index n implying a sum over that index.

The unit normal to S_p is given by the expression $\hat{\mathbf{n}} = \nabla f / |\nabla f|$, with f defined by the relation $f(\mathbf{r}) = r - a\left[1 + \epsilon \mathbf{P}_n \boxed{n} \mathbf{A}_n + O(\epsilon^2)\right] = 0$. Together with the use of the gradient operator in invariant spherical coordinates,

$$\nabla = \hat{\mathbf{r}} \left(\frac{\partial}{\partial r}\right)_{\hat{\mathbf{r}}} + \frac{1}{r}\left(\frac{\partial}{\partial \hat{\mathbf{r}}}\right)_r, \tag{4}$$

one thus obtains

$$\hat{\mathbf{n}} = \hat{\mathbf{r}} - \epsilon \hat{\nabla} \mathbf{P}_n \boxed{n} \mathbf{A}_n + O(\epsilon^2), \tag{5}$$

wherein $\hat{\nabla} = (\partial/\partial \hat{\mathbf{r}})_r$ is the surface gradient operator. In deriving the latter, we have used the identities $\hat{\mathbf{r}}\hat{\mathbf{r}} \cdot \hat{\nabla} = 0$ and $(\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}) \cdot \hat{\nabla} = \hat{\nabla}$, arising from the orthonormality of the coordinate axes.

The equations of Stokes flow, describing the flow around the deformed sphere, are

$$\nabla \cdot \mathbf{v} = 0, \tag{6}$$

$$\mu \nabla^2 \mathbf{v} = \nabla p, \tag{7}$$

where p denotes the pressure. The boundary conditions imposed upon the fluid's velocity require that the latter approach the particle-free, undisturbed fluid motion at infinity,

$$\mathbf{v} \to \mathbf{0} \text{ as } r \to \infty, \tag{8}$$

and that the thermal slip condition on S_p, (1), satisfy

$$\mathbf{v} = \mathbf{U} + C_s (\mathbf{I} - \hat{\mathbf{n}}\hat{\mathbf{n}}) \cdot \nabla T_s, \tag{9}$$

where T_s denotes the temperature within the solid (with $T_s = T$ on S_p), and C_s is the thermal slip coefficient, which in the case of gases takes the value $3/4\,(\mu/\rho T)$, according to Maxwell. In that case, C_s is a constant, owing to the inverse relationship

existing between ρ and T from the ideal gas law, with the transport coefficient μ assumed constant, at least for small temperature gradients. For liquids, Semenov [23] has provided a formula whereby C_s can be calculated from the properties of the liquid and the solid particle. Values of C_s for the thermophoretic motion of silica particles in several solvents are found to be of the order of 10^{-8}–10^{-7} cm^2 s^{-1} K^{-1} [23]. Values of similar order of magnitude for various particle-liquid systems have been found in the experimental study of [19]. Thermal creep in liquids has also been proposed to exist in [2], where C_s is given by the product of the liquid's coefficient of thermal expansion and its thermometric diffusivity.

Since the boundary condition on velocity, namely (9), may be expressed as the sum of the particle velocity $\mathbf{U}$ and the thermal slip velocity $C_s\,(\mathbf{I}-\hat{\mathbf{n}}\hat{\mathbf{n}})\cdot\nabla T_s$, we will restrict ourselves, initially, to solving for the flow caused by thermal slip alone. Later, the flow arising from the motion $\mathbf{U}$ of the particle will be linearly superposed.

The temperature fields within the fluid and the solid are governed by the respective equations

$$\nabla^2 T = 0,$$
$$\nabla^2 T_s = 0, \tag{10}$$

subject to the joint boundary conditions

$$T = T_s, \tag{11}$$
$$\hat{\mathbf{n}}\cdot\nabla T = \gamma\hat{\mathbf{n}}\cdot\nabla T_s, \tag{12}$$

on S_p, where γ denotes the ratio k_s/k of the thermal conductivity of the solid, k_s, to that of the fluid, k. In addition, we have the further conditions that the temperature within the particle remains finite at $r = 0$, while that within the fluid at infinity is given by the expression

$$T\,(r\to\infty) = \text{const} + \mathbf{G}\cdot\mathbf{r}. \tag{13}$$

The constant appearing in (13) is physically irrelevant, whence we may set it to zero. Perturbation solutions will now be developed for the flow and temperature fields in terms of the perturbation parameter, ϵ.

3. Temperature fields. Define *vector* temperature fields in the fluid and the solid, $\mathbf{T}$ and $\mathbf{T}_s$, respectively, through the relations

$$T = \mathbf{T}\cdot\mathbf{G},$$
$$T_s = \mathbf{T}_s\cdot\mathbf{G}. \tag{14}$$

Since the scalar temperature fields satisfy Laplace's equation, and whereas the gradient at infinity, $\mathbf{G}$, is an arbitrary constant, it follows from (10)–(12) that the vector temperature fields satisfy the respective equations

$$\nabla^2\mathbf{T} = \mathbf{0},$$
$$\nabla^2\mathbf{T}_s = \mathbf{0}, \tag{15}$$

subject to the boundary conditions

$$\mathbf{T} = \mathbf{T}_s \tag{16}$$

and

$$\hat{\mathbf{n}} \cdot \nabla \mathbf{T} = \gamma \hat{\mathbf{n}} \cdot \nabla \mathbf{T}_s, \tag{17}$$

on the surface S_p of the deformed sphere. We expand the vector temperature fields in powers of ϵ,

$$\begin{aligned} \mathbf{T}(\mathbf{r};\epsilon) &= \mathbf{T}^{(0)}(\mathbf{r}) + \epsilon \mathbf{T}^{(1)}(\mathbf{r}) + O(\epsilon^2), \\ \mathbf{T}_s(\mathbf{r};\epsilon) &= \mathbf{T}_s^{(0)}(\mathbf{r}) + \epsilon \mathbf{T}_s^{(1)}(\mathbf{r}) + O(\epsilon^2). \end{aligned} \tag{18}$$

The temperature fields inside and outside of the particle at various perturbation orders, subject to the boundary conditions on S_p given by (16) and (17), may be obtained from the solution of a sequence of problems satisfying appropriate boundary conditions on the surface of the *undeformed* sphere. These latter boundary conditions, to be imposed on the sphere surface, are obtained by a Taylor series expansion of (16) and (17) about $r = a$. Thus, expansion of (16) while making use of (2) furnishes the following conditions to be imposed at $r = a$ on the $O(1)$ and $O(\epsilon)$ vector temperature fields:

$$\mathbf{T}^{(0)}\Big|_a = \mathbf{T}_s^{(0)}\Big|_a, \tag{19}$$

$$\mathbf{T}^{(1)}\Big|_a + a\mathbf{A}_n \boxed{n} \mathbf{P}_n \frac{\partial \mathbf{T}^{(0)}}{\partial r}\Big|_a = \mathbf{T}_s^{(1)}\Big|_a + a\mathbf{A}_n \boxed{n} \mathbf{P}_n \frac{\partial \mathbf{T}_s^{(0)}}{\partial r}\Big|_a. \tag{20}$$

Similarly, upon expanding $\nabla \mathbf{T}|_{S_p}$ and $\nabla \mathbf{T}_s|_{S_p}$ in Taylor series about $r = a$ and making use of (5), it follows from (17) that

$$\hat{\mathbf{r}} \cdot \nabla \mathbf{T}^{(0)}\Big|_a = \gamma\, \hat{\mathbf{r}} \cdot \nabla \mathbf{T}_s^{(0)}\Big|_a \tag{21}$$

and

$$\begin{aligned} &\hat{\mathbf{r}} \cdot \nabla \mathbf{T}^{(1)}\Big|_a + a\mathbf{A}_n \boxed{n} \mathbf{P}_n \frac{\partial}{\partial r} \hat{\mathbf{r}} \cdot \nabla \mathbf{T}^{(0)}\Big|_a - \hat{\nabla}\left(\mathbf{P}_n \boxed{n} \mathbf{A}_n\right) \cdot \nabla \mathbf{T}^{(0)}\Big|_a \\ &= \gamma\, \hat{\mathbf{r}} \cdot \nabla \mathbf{T}_s^{(1)}\Big|_a + a\gamma \mathbf{A}_n \boxed{n} \mathbf{P}_n \frac{\partial}{\partial r} \hat{\mathbf{r}} \cdot \nabla \mathbf{T}_s^{(0)}\Big|_a - \gamma \hat{\nabla}\left(\mathbf{P}_n \boxed{n} \mathbf{A}_n\right) \cdot \nabla \mathbf{T}_s^{(0)}\Big|_a. \end{aligned} \tag{22}$$

The leading-order, undeformed sphere problem is described by the set of equations

$$\begin{aligned} \nabla^2 \mathbf{T}^{(0)} &= \mathbf{0}, \\ \nabla^2 \mathbf{T}_s^{(0)} &= \mathbf{0}, \end{aligned} \tag{23}$$

subject to boundary conditions (19) and (21). Accordingly, the leading-order vector temperature fields are readily found to be

$$\begin{aligned} \mathbf{T}^{(0)} &= \left[1 + \left(\frac{1-\gamma}{2+\gamma}\right)\left(\frac{a}{r}\right)^3\right]\mathbf{r}, \\ \mathbf{T}_s^{(0)} &= \frac{3}{2+\gamma}\mathbf{r}. \end{aligned} \tag{24}$$

Upon applying the gradient operator, given by (4), to (24), and noting that $(\partial\hat{\mathbf{r}}/\partial\hat{\mathbf{r}})_r = \mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}$, we obtain

$$\nabla\mathbf{T}^{(0)} = \left[1 - 2\left(\frac{1-\gamma}{2+\gamma}\right)\frac{a^3}{r^3}\right]\hat{\mathbf{r}}\hat{\mathbf{r}} + \left[1 + \left(\frac{1-\gamma}{2+\gamma}\right)\left(\frac{a}{r}\right)^3\right](\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}),$$

$$\nabla\mathbf{T}_s^{(0)} = \frac{3}{2+\gamma}\mathbf{I}. \tag{25}$$

Making use of (25) furnishes the following $O(\epsilon)$ boundary conditions from (20) and (22):

$$\mathbf{T}^{(1)}\Big|_a = \mathbf{T}_s^{(1)}\Big|_a + 3\left(\frac{1-\gamma}{2+\gamma}\right)a\mathbf{A}_n \boxed{n} \mathbf{P}_n\mathbf{P}_1 \tag{26}$$

[wherein we have substituted $\mathbf{P}_1(\hat{\mathbf{r}})$ for $\hat{\mathbf{r}}$], and

$$\hat{\mathbf{r}}\cdot\nabla\mathbf{T}^{(1)}\Big|_a + 6\left(\frac{1-\gamma}{2+\gamma}\right)\mathbf{A}_n \boxed{n} \mathbf{P}_n\mathbf{P}_1 - 3\left(\frac{1-\gamma}{2+\gamma}\right)\hat{\nabla}\mathbf{P}_n \boxed{n} \mathbf{A}_n$$

$$= \gamma\,\hat{\mathbf{r}}\cdot\nabla\mathbf{T}_s^{(1)}\Big|_a. \tag{27}$$

The $O(\epsilon)$ vector temperature fields are governed by the equations

$$\nabla^2\mathbf{T}^{(1)} = \mathbf{0},$$

$$\nabla^2\mathbf{T}_s^{(1)} = \mathbf{0}, \tag{28}$$

subject to the boundary conditions given by (26) and (27). Since it is convenient to expand $\mathbf{T}^{(1)}$ and $\mathbf{T}_s^{(1)}$ as linear combinations of solid harmonics, these boundary conditions must be expressed as linear combinations of the polyadic surface harmonics.

4. Flow field. Next, expand the velocity and pressure fields as perturbation expansions in ϵ:

$$\mathbf{v}(\mathbf{r};\epsilon) = \mathbf{v}^{(0)}(\mathbf{r}) + \epsilon\mathbf{v}^{(1)}(\mathbf{r}) + O(\epsilon^2),$$

$$p(\mathbf{r};\epsilon) = p^{(0)}(\mathbf{r}) + \epsilon p^{(1)}(\mathbf{r}) + O(\epsilon^2). \tag{29}$$

At each order, the perturbation fields $\mathbf{v}^{(i)}$ and $p^{(i)}$ obey the Stokes equations,

$$\nabla\cdot\mathbf{v}^{(i)} = 0, \tag{30}$$

$$\mu\nabla^2\mathbf{v}^{(i)} = \nabla p^{(i)}. \tag{31}$$

The thermal slip boundary condition on the surface S_p of a stationary particle is, upon setting $\mathbf{U} = \mathbf{0}$ in (9),

$$\mathbf{v} = C_s\,(\mathbf{I} - \hat{\mathbf{n}}\hat{\mathbf{n}})\cdot\nabla T_s. \tag{32}$$

In addition, the velocity fields at all orders vanish at infinity. Expansion of the velocity at the surface S_p of the deformed sphere in a Taylor series about $r = a$ gives

$$\mathbf{v}|_{S_p} = \mathbf{v}^{(0)}\Big|_{r=a} + \epsilon\left(\mathbf{v}^{(1)}\Big|_{r=a} + a\mathbf{A}_n \boxed{n} \mathbf{P}_n\,\frac{\partial\mathbf{v}^{(0)}}{\partial r}\Big|_{r=a}\right) + O(\epsilon^2). \tag{33}$$

A similar expansion of the temperature gradient ∇T_s on S_p about $r = a$ yields

$$\nabla T_s|_{S_p} = \nabla T_s^{(0)}\Big|_{r=a} + \epsilon \left(\nabla T_s^{(1)}\Big|_{r=a} + a\mathbf{A}_n \boxed{n} \mathbf{P}_n \frac{\partial}{\partial r} \nabla T_s^{(0)}\Big|_{r=a} \right) + O(\epsilon^2). \tag{34}$$

Substitution of (5), (33) and (34) into (32) furnishes the following $O(1)$ and $O(\epsilon)$ boundary conditions at $r = a$:

$$\mathbf{v}^{(0)}\Big|_{r=a} = C_s \left(\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}\right) \cdot \nabla T_s^{(0)}\Big|_{r=a} \tag{35}$$

and

$$\begin{aligned} \mathbf{v}^{(1)}\Big|_{r=a} = &- a\mathbf{A}_n \boxed{n} \mathbf{P}_n \frac{\partial \mathbf{v}^{(0)}}{\partial r}\Big|_{r=a} + C_s \left\{ \hat{\mathbf{r}} \hat{\nabla} \left(\mathbf{P}_n \boxed{n} \mathbf{A}_n \right) \right. \\ &\left. + \hat{\nabla} \left(\mathbf{P}_n \boxed{n} \mathbf{A}_n \right) \hat{\mathbf{r}} \right\} \cdot \nabla T_s^{(0)}\Big|_{r=a} + C_s \left(\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}\right) \cdot \nabla T_s^{(1)}\Big|_{r=a} \\ &+ C_s a \left(\mathbf{A}_n \boxed{n} \mathbf{P}_n \right) \left(\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}\right) \cdot \frac{\partial}{\partial r} \nabla T_s^{(0)}\Big|_{r=a}. \end{aligned} \tag{36}$$

The leading-order velocity field, which vanishes at infinity and is subject to the thermal slip condition, given by (35) taken in conjunction with the leading-order temperature gradient on S_p from (25), corresponds to the flow around an undeformed sphere. It is found, using the method of [5], to be

$$\mathbf{v}^{(0)} = \frac{3C_s}{2+\gamma} \left[\left(\frac{a}{2r} + \frac{a^3}{2r^3} \right) \mathbf{I} + \left(\frac{a}{2r} - \frac{3a^3}{2r^3} \right) \hat{\mathbf{r}}\hat{\mathbf{r}} \right] \cdot \mathbf{G}. \tag{37}$$

Equations (25) and (37) in combination with (36) yield

$$\begin{aligned} \mathbf{v}^{(1)}\Big|_{r=a} = &\frac{6C_s}{2+\gamma} \left(\mathbf{A}_n \boxed{n} \mathbf{P}_n \right) \left(\mathbf{I} - 2\hat{\mathbf{r}}\hat{\mathbf{r}}\right) \cdot \mathbf{G} \\ &+ \frac{3C_s}{2+\gamma} \left[\hat{\mathbf{r}} \hat{\nabla} \left(\mathbf{P}_n \boxed{n} \mathbf{A}_n \right) + \hat{\nabla} \left(\mathbf{P}_n \boxed{n} \mathbf{A}_n \right) \hat{\mathbf{r}} \right] \cdot \mathbf{G} \\ &+ C_s \left(\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}\right) \cdot \nabla T_s^{(1)}\Big|_{r=a}. \end{aligned} \tag{38}$$

We have now obtained the governing equations and concomitant boundary conditions for the $O(\epsilon)$ temperature and velocity problems. However, the completely general solution of these problems, requiring use of general recurrence relations relating the various polyadic surface harmonics and their gradients, is extremely complicated algebraically. Accordingly, in lieu of attempting a completely general calculation, we instead provide the solution only for the specific case of a general triaxial ellipsoid by way of illustrating the general scheme.

5. Nonisothermal flow around an ellipsoid. Consider the ellipsoid,

$$\left(\frac{x_1}{a_1}\right)^2 + \left(\frac{x_2}{a_2}\right)^2 + \left(\frac{x_3}{a_3}\right)^2 = 1, \tag{39}$$

where (x_1, x_2, x_3) are Cartesian coordinates fixed in the ellipsoid, with the coordinate axes pointing along the principal axes of the ellipsoid, and with (a_1, a_2, a_3) the semi-lengths of these axes. Oblate and prolate spheroids, for which two of the three

principal axes are equal, are obtained as special cases of the ellipsoid, while a circular disk and a needle-shaped object may be approximated by an oblate and a prolate spheroid, respectively, one of whose semi-axes shrinks to zero.

In invariant form, the equation of the ellipsoid may be expressed as [6]

$$\mathbf{r}\cdot\mathbf{D}\cdot\mathbf{r}=1, \tag{40}$$

where $\mathbf{D}$ is the body-fixed dyadic

$$\mathbf{D}=\frac{\mathbf{i}_1\mathbf{i}_1}{a_1^2}+\frac{\mathbf{i}_2\mathbf{i}_2}{a_2^2}+\frac{\mathbf{i}_3\mathbf{i}_3}{a_3^2}, \tag{41}$$

in which $(\mathbf{i}_1,\mathbf{i}_2,\mathbf{i}_3)$ are body-fixed Cartesian unit vectors directed along the principal axes of the ellipsoid. Consider the case where the ellipsoid is a slightly deformed sphere, with the semi-lengths of the principal axes given by $a_1=a(1+\epsilon\alpha_1)$, $a_2=a(1+\epsilon\alpha_2)$ and $a_3=a(1+\epsilon\alpha_3)$, where $\epsilon\ll 1$. Since $\mathbf{i}_1\mathbf{i}_1+\mathbf{i}_2\mathbf{i}_2+\mathbf{i}_3\mathbf{i}_3=\mathbf{I}$, it follows that

$$\mathbf{D}=\frac{\mathbf{I}}{a^2}-\frac{2\epsilon}{a^2}\mathbf{B}+O(\epsilon^2), \tag{42}$$

where $\mathbf{B}$ is the body-fixed dyadic

$$\mathbf{B}=\alpha_1\mathbf{i}_1\mathbf{i}_1+\alpha_2\mathbf{i}_2\mathbf{i}_2+\alpha_3\mathbf{i}_3\mathbf{i}_3. \tag{43}$$

For a given ellipsoid, whose volume is $V=4\pi a_1a_2a_3/3$, it is convenient to choose the radius a of the sphere such that its volume, $4\pi a^3/3$, is equal to that of the ellipsoid; that is, $a^3=a_1a_2a_3$. Since $a_i=a(1+\epsilon\alpha_i)$, this requires that $\alpha_1+\alpha_2+\alpha_3=0$, i.e., $\mathbf{I}:\mathbf{B}=0$, or, alternatively,

$$\mathrm{Tr}(\mathbf{B})=0, \tag{44}$$

where, for any dyadic $\mathbf{B}$, $\mathrm{Tr}(\mathbf{B})$ denotes the trace, $\mathbf{B}:\mathbf{I}$.

Upon combining (40) and (42) and making use of the symmetry of $\mathbf{B}$ together with the identity $\hat{\mathbf{r}}\hat{\mathbf{r}}=2\mathbf{P}_2(\hat{\mathbf{r}})/3+\mathbf{I}/3$, we obtain

$$r=a\left[1+\frac{2}{3}\epsilon\mathbf{B}:\mathbf{P}_2(\hat{\mathbf{r}})+O(\epsilon^2)\right]. \tag{45}$$

$\mathbf{B}$ is diagonal, and, hence, symmetric in the body-fixed coordinate system. As a result, it is symmetric in any basis.

The detailed derivation of the $O(\epsilon)$ temperature fields outside and within the ellipsoid is presented in Appendix A. The resulting solutions are

$$\mathbf{T}^{(1)}=-\frac{6}{5}\left(\frac{1-\gamma}{2+\gamma}\right)^2\frac{a^3}{r^2}\mathbf{P}_1\cdot\mathbf{B}+\frac{6}{5}\left(\frac{1-\gamma}{2+\gamma}\right)\frac{a^5}{r^4}\mathbf{P}_3:\mathbf{B} \tag{46}$$

and

$$\mathbf{T}_s^{(1)}=-\frac{18}{5}\frac{1-\gamma}{(2+\gamma)^2}\mathbf{r}\cdot\mathbf{B}. \tag{47}$$

The $O(\epsilon)$ boundary condition on velocity, (38), thus reduces to

$$\begin{aligned}\mathbf{v}^{(1)}\Big|_{r=a}=\frac{2C_s}{2+\gamma}\Big[&2\mathbf{B}:\mathbf{P}_2\left(\mathbf{I}-2\hat{\mathbf{r}}\hat{\mathbf{r}}\right)+\hat{\mathbf{r}}\hat{\nabla}\mathbf{P}_2:\mathbf{B}+\hat{\nabla}\left(\mathbf{P}_2:\mathbf{B}\right)\hat{\mathbf{r}}\\&-\frac{9}{5}\left(\frac{1-\gamma}{2+\gamma}\right)\left(\mathbf{I}-\hat{\mathbf{r}}\hat{\mathbf{r}}\right)\cdot\mathbf{B}\Big]\cdot\mathbf{G}.\end{aligned} \tag{48}$$

The force on the ellipsoid can be obtained without solving the detailed boundary value problem by making use of the following expression for the force on a body of arbitrary shape [4]:

$$\mathbf{F} = -\frac{3}{2}\mu a \int_{S_1} \mathbf{v}|_{r=a}\, d\Omega, \tag{49}$$

where the integration is to be carried out over the surface of a unit sphere, S_1, and wherein $d\Omega = d^2\hat{\mathbf{r}}$ is a differential element of solid angle. The analogous expression for the torque on the body about the origin is [4]:

$$\mathbf{T} = 3\mu a^2 \boldsymbol{\varepsilon} : \int_{S_1} \hat{\mathbf{r}}\ \mathbf{v}|_{r=a}\, d\Omega, \tag{50}$$

where $\boldsymbol{\varepsilon} = -\mathbf{I} \times \mathbf{I}$ is the alternating unit triadic. Following [10], we introduce the translational hydrodynamic resistance dyadic $\mathbf{K}$ via the expression

$$\mathbf{F} = -\mu \mathbf{K} \cdot \mathbf{G}. \tag{51}$$

Expansion of $\mathbf{F}$ and $\mathbf{K}$ in powers of ϵ gives, for the leading-order terms,

$$\mathbf{F}^{(0)} = -\frac{3}{2}\mu a \int_{S_1} \mathbf{v}^{(0)}\Big|_{r=a}\, d\Omega = -\mu \mathbf{K}^{(0)} \cdot \mathbf{G}, \tag{52}$$

whence, from (35) [17],

$$\mathbf{K}^{(0)} = \frac{9C_s a}{2(2+\gamma)} \int_{S_1} (\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}})\, d\Omega = \frac{12\pi C_s a}{2+\gamma}\mathbf{I}, \tag{53}$$

in which we have used the identity $\int_{S_1} (\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}})\, d\Omega = 8\pi\mathbf{I}/3$. Similarly, for the first-order terms, we find that

$$\mathbf{F}^{(1)} = -\frac{3}{2}\mu a \int_{S_1} \mathbf{v}^{(1)}\Big|_{r=a}\, d\Omega = -\mu \mathbf{K}^{(1)} \cdot \mathbf{G}, \tag{54}$$

where

$$\begin{aligned}\mathbf{K}^{(1)} = \frac{3C_s a}{2+\gamma} \Bigg[& 2\int_{S_1} \mathbf{P}_2\, (\mathbf{I} - 2\hat{\mathbf{r}}\hat{\mathbf{r}}) : \mathbf{B} d\Omega + \int_{S_1} \hat{\mathbf{r}}\hat{\nabla}\mathbf{P}_2 : \mathbf{B} d\Omega + \int_{S_1} \hat{\nabla}\mathbf{P}_2 : \mathbf{B}\hat{\mathbf{r}} d\Omega \\ & -\frac{9}{5}\left(\frac{1-\gamma}{2+\gamma}\right) \int_{S_1} (\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}) \cdot \mathbf{B} d\Omega \Bigg].\end{aligned} \tag{55}$$

Details pertaining to the calculation of $\mathbf{K}^{(1)}$ are given in Appendix B, with the result being

$$\mathbf{K}^{(1)} = -\frac{24\pi}{5} C_s a \frac{1-4\gamma}{(2+\gamma)^2}\mathbf{B}. \tag{56}$$

We have now obtained the force acting on the stationary ellipsoid due to the thermal creep-induced flow. When the ellipsoid moves through the fluid at a velocity of $\mathbf{U}$, the preceding solution is superposed on the additional Stokes flow $(\mathbf{v}', p')$ satisfying the boundary conditions

$$\begin{aligned}\mathbf{v}' &= \mathbf{U} \text{ on } S_p, \\ \mathbf{v}'(r \to \infty) &\to \mathbf{0}.\end{aligned} \tag{57}$$

Analogous to (35) and (48), we require that

$$\mathbf{v}'^{(0)}\Big|_{r=a} = \mathbf{U},$$

$$\mathbf{v}'^{(1)}\Big|_{r=a} = -a\mathbf{A}_n \boxed{n} \mathbf{P}_n \frac{\partial \mathbf{v}'^{(0)}}{\partial r}\Big|_{r=a}. \tag{58}$$

The leading-order velocity field $\mathbf{v}'^{(0)}$ in invariant form is given by the expression [5, 10]

$$\mathbf{v}'^{(0)} = \left[\frac{3a}{4r}\left(\mathbf{I} + \hat{\mathbf{r}}\hat{\mathbf{r}}\right) + \frac{a^3}{4r^3}\left(\mathbf{I} - 3\hat{\mathbf{r}}\hat{\mathbf{r}}\right)\right] \cdot \mathbf{U}, \tag{59}$$

which, in combination with (58), yields

$$\mathbf{v}'^{(1)}\Big|_{r=a} = \frac{3}{2}\left(\mathbf{A}_n \boxed{n} \mathbf{P}_n\right)\left(\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}\right) \cdot \mathbf{U}. \tag{60}$$

At leading order, the force on the ellipsoid is given by Stokes' law, $\mathbf{F}'^{(0)} = -6\pi\mu a\mathbf{U}$, whereas at $O(\epsilon)$, we obtain

$$\mathbf{F}'^{(1)} = -\frac{3\mu a}{2}\int_{S_1} \mathbf{v}'^{(1)}\Big|_{r=a} d\Omega = -\mu\mathbf{K}'^{(1)} \cdot \mathbf{U}, \tag{61}$$

where

$$\begin{aligned}\mathbf{K}'^{(1)} &= \frac{3}{2}a\mathbf{B} : \int_{S_1} \mathbf{P}_2\left(\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}\right) d\Omega \\ &= -\frac{6\pi}{5}a\mathbf{B}.\end{aligned} \tag{62}$$

Equation (62) is in exact agreement with the corresponding result of [4] satisfying (57), wherein the surface of the deformed sphere was expressed as an expansion in scalar spherical harmonics.

The net force on the ellipsoid is obtained from the superposition of the force acting on a stationary ellipsoid under an externally imposed temperature gradient together with the force acting on an ellipsoid translating at a velocity of $\mathbf{U}$ under isothermal conditions:

$$\mathbf{F}_{\text{net}} = -\mu\left(\mathbf{K} \cdot \mathbf{G} + \mathbf{K}' \cdot \mathbf{U}\right). \tag{63}$$

It can be seen from (35) and (48) that the torque $\mathbf{T}$ acting on the ellipsoid about its center, given by (50), vanishes. Under isothermal conditions, the centroid of the ellipsoid is a center of hydrodynamic stress. The existence of such a center for a nonskew body such as an ellipsoid implies that no torque acts about its centroid as the ellipsoid translates without slip [10].

Accordingly, the force- and torque-free ellipsoid will move, without rotation, under the influence of the externally imposed temperature gradient $\mathbf{G}$ at a velocity $\mathbf{U} = -\left(\mathbf{K}'\right)^{-1} \cdot \mathbf{K} \cdot \mathbf{G}$, where

$$\mathbf{K}' = 6\pi a\left[\mathbf{I} - \frac{\epsilon}{5}\mathbf{B} + O(\epsilon^2)\right], \tag{64}$$

and

$$\mathbf{K} = \frac{12\pi C_s a}{2+\gamma}\left[\mathbf{I} - \epsilon\frac{2}{5}\left(\frac{1-4\gamma}{2+\gamma}\right)\mathbf{B} + O(\epsilon^2)\right]. \tag{65}$$

In view of the identity $\left[\mathbf{I}+\epsilon\mathbf{C}+O(\epsilon^2)\right]^{-1}=\mathbf{I}-\epsilon\mathbf{C}+O(\epsilon^2)$, valid for any dyadic $\mathbf{C}$, this makes

$$\mathbf{U}=-\frac{2C_s}{2+\gamma}\left[\mathbf{I}+\epsilon\frac{9}{5}\frac{\gamma}{2+\gamma}\mathbf{B}+O(\epsilon^2)\right]\cdot\mathbf{G}. \tag{66}$$

This constitutes the principal result of our ellipsoid calculation.

In the special case where the ellipsoid is nonconducting, $\gamma=0$, whereupon the above reduces simply to

$$\mathbf{U}=-C_s\mathbf{G}\left[1+O(\epsilon^2)\right]. \tag{67}$$

As such, to at least terms of first order in the deformation, the ellipsoid's thermophoretic velocity is independent of its shape, size and orientation, and hence is identical to that of a sphere. While we have formally demonstrated the latter result only to the first order, according to Morrison's theory this nonconducting result should hold to all orders in the ellipsoid deformation ϵ.

6. Discussion. We have developed an asymptotic solution of flow around a heat-conducting, slightly deformed sphere under an imposed temperature gradient, wherein the fluid slips at the surface of the deformed sphere. Although we have provided an explicit solution only for the case wherein the deformed sphere is an ellipsoid, to first order in the perturbation scheme, the technique may be readily applied to arbitrarily deformed bodies whose deviation from a spherical shape is small. In principle, the perturbation solution can be carried out to any order. For the specific case of nonconducting particles, we confirm Morrison's generic deduction [18] that the phoretic velocity of a particle is independent of its size, shape and orientation.

Appendix A. We here solve in detail the $O(\epsilon)$ temperature problem for the flow around the ellipsoid. From (26) and (27), the boundary conditions for the $O(\epsilon)$ temperature problem appropriate to this geometry reduce to

$$\mathbf{T}^{(1)}\Big|_a=\mathbf{T}_s^{(1)}\Big|_a+2\left(\frac{1-\gamma}{2+\gamma}\right)a\mathbf{B}:\mathbf{P}_2\mathbf{P}_1 \tag{A1}$$

and

$$\hat{\mathbf{r}}\cdot\nabla\mathbf{T}^{(1)}\Big|_a+2\left(\frac{1-\gamma}{2+\gamma}\right)\left[2\mathbf{B}:\mathbf{P}_2\mathbf{P}_1-\hat{\nabla}\mathbf{P}_2:\mathbf{B}\right]=\gamma\,\hat{\mathbf{r}}\cdot\nabla\mathbf{T}_s^{(1)}\Big|_a. \tag{A2}$$

Since $\mathbf{P}_1(\hat{\mathbf{r}})=\hat{\mathbf{r}}$ and $\mathbf{P}_2(\hat{\mathbf{r}})=\left(3\hat{\mathbf{r}}\hat{\mathbf{r}}-\mathbf{I}\right)/2$, we obtain

$$\mathbf{P}_2\mathbf{P}_1=\frac{3}{2}\hat{\mathbf{r}}\hat{\mathbf{r}}\hat{\mathbf{r}}-\frac{1}{2}\mathbf{I}\hat{\mathbf{r}}, \tag{A3}$$

which, in combination with the expression [4]

$$\hat{\mathbf{r}}\hat{\mathbf{r}}\hat{\mathbf{r}}=\frac{2}{5}\mathbf{P}_3+\frac{1}{5}\left[(\mathbf{I}\mathbf{P}_1)+(\mathbf{I}\mathbf{P}_1)^{\dagger}+(\mathbf{P}_1\mathbf{I})\right], \tag{A4}$$

yields

$$\mathbf{P}_2\mathbf{P}_1=\frac{3}{5}\mathbf{P}_3-\frac{1}{5}(\mathbf{I}\mathbf{P}_1)+\frac{3}{10}(\mathbf{I}\mathbf{P}_1)^{\dagger}+\frac{3}{10}(\mathbf{P}_1\mathbf{I}). \tag{A5}$$

Straightforward differentiation of $\mathbf{P}_2$ yields

$$\hat{\nabla}\mathbf{P}_2 = \frac{3}{2}\left[\mathbf{I}\hat{\mathbf{r}} + (\mathbf{I}\hat{\mathbf{r}})^\dagger - 2\hat{\mathbf{r}}\hat{\mathbf{r}}\hat{\mathbf{r}}\right], \tag{A6}$$

where the transposition symbol † entails the interchange of the two indices immediately preceding or succeeding it, according as it follows or precedes the argument to which it is affixed. Therefore, upon making use of (A4), we obtain

$$\hat{\nabla}\mathbf{P}_2 = \frac{9}{10}\left[\mathbf{I}\mathbf{P}_1 + (\mathbf{I}\mathbf{P}_1)^\dagger\right] - \frac{6}{5}\mathbf{P}_3 - \frac{3}{5}\mathbf{P}_1\mathbf{I}. \tag{A7}$$

Equations (A1) and (A2) now simplify to

$$\mathbf{T}^{(1)}\Big|_a = \mathbf{T}_s^{(1)}\Big|_a + 2\left(\frac{1-\gamma}{2+\gamma}\right) a\mathbf{B} : \left[\frac{3}{5}\mathbf{P}_3 - \frac{1}{5}\mathbf{I}\mathbf{P}_1 + \frac{3}{10}(\mathbf{I}\mathbf{P}_1)^\dagger + \frac{3}{10}\mathbf{P}_1\mathbf{I}\right] \tag{A8}$$

and

$$\begin{aligned}&\frac{\partial\mathbf{T}^{(1)}}{\partial r}\Big|_a + 2\left(\frac{1-\gamma}{2+\gamma}\right)\left\{2\mathbf{B} : \left[\frac{3}{5}\mathbf{P}_3 - \frac{1}{5}\mathbf{I}\mathbf{P}_1 + \frac{3}{10}(\mathbf{I}\mathbf{P}_1)^\dagger + \frac{3}{10}\mathbf{P}_1\mathbf{I}\right]\right.\\ &\left. - \left[\frac{9}{10}\mathbf{I}\mathbf{P}_1 + \frac{9}{10}(\mathbf{I}\mathbf{P}_1)^\dagger - \frac{6}{5}\mathbf{P}_3 - \frac{3}{5}\mathbf{P}_1\mathbf{I}\right] : \mathbf{B}\right\} = \gamma\,\frac{\partial\mathbf{T}_s^{(1)}}{\partial r}\Big|_a.\end{aligned} \tag{A9}$$

Owing to the symmetry of $\mathbf{B}$, the boundary conditions, (A8) and (A9), may be rewritten in the forms

$$\mathbf{T}^{(1)}\Big|_a = \mathbf{T}_s^{(1)}\Big|_a + \frac{6}{5}\left(\frac{1-\gamma}{2+\gamma}\right) a\left(\mathbf{P}_1\cdot\mathbf{B} + \mathbf{P}_3 : \mathbf{B}\right) \tag{A10}$$

and

$$\frac{\partial\mathbf{T}^{(1)}}{\partial r}\Big|_a - \frac{6}{5}\left(\frac{1-\gamma}{2+\gamma}\right)\left[\mathbf{P}_1\cdot\mathbf{B} - 4\mathbf{P}_3 : \mathbf{B}\right] = \gamma\,\frac{\partial\mathbf{T}_s^{(1)}}{\partial r}\Big|_a. \tag{A11}$$

Upon expanding $\mathbf{T}^{(1)}$ and $\mathbf{T}_s^{(1)}$ in solid harmonics and imposing the boundary conditions, (A10) and (A11), in conjunction with the vanishing of $\mathbf{T}^{(1)}$ at infinity and the finiteness of $\mathbf{T}_s^{(1)}$ at $r = 0$, we eventually obtain the respective expressions given in (46) and (47).

Appendix B. Here, we present the detailed computation of $\mathbf{K}^{(1)}$, which is given by (55). Since $P_0 = 1$, and $\hat{\mathbf{r}}\hat{\mathbf{r}} = 2\mathbf{P}_2/3 + \mathbf{I}/3$, it follows that

$$\int_{S_1}\mathbf{P}_2\left(\mathbf{I} - 2\hat{\mathbf{r}}\hat{\mathbf{r}}\right)d\Omega = -\frac{4}{3}\int_{S_1}\mathbf{P}_2\mathbf{P}_2 d\Omega, \tag{B1}$$

where we have used the orthogonality of the polyadics $\mathbf{P}_n$. We make use of (A7) to write

$$\begin{aligned}\int_{S_1}\hat{\mathbf{r}}\hat{\nabla}\mathbf{P}_2 d\Omega = {}& \frac{9}{10}\int_{S_1}\mathbf{P}_1\mathbf{I}\mathbf{P}_1 d\Omega + \frac{9}{10}\int_{S_1}\mathbf{P}_1(\mathbf{I}\mathbf{P}_1)^\dagger d\Omega\\ & -\frac{6}{5}\int_{S_1}\mathbf{P}_1\mathbf{P}_3 d\Omega - \frac{3}{5}\int_{S_1}\mathbf{P}_1\mathbf{P}_1 d\Omega\mathbf{I}.\end{aligned} \tag{B2}$$

We know that $\int_{S_1} \mathbf{P}_1\mathbf{P}_1 d\Omega = 4\pi\mathbf{I}/3$, and $\int_{S_1} \mathbf{P}_1\mathbf{P}_3 d\Omega = \mathbf{0}$. Reverting to Cartesian tensor notation in order to calculate the remaining integrals in (B1) and (B2) yields

$$\int_{S_1} (\mathbf{P}_2\mathbf{P}_2)_{ijkl}\, d\Omega = \frac{\pi}{5}\left(-2\delta_{ij}\delta_{kl} + 3\delta_{ik}\delta_{jl} + 3\delta_{il}\delta_{jk}\right), \tag{B3}$$

$$\int_{S_1} (\mathbf{P}_1\mathbf{I}\mathbf{P}_1)_{ijkl}\, d\Omega = \frac{4\pi}{3}\delta_{il}\delta_{jk}, \tag{B4}$$

$$\int_{S_1} \left[\mathbf{P}_1(\mathbf{I}\mathbf{P}_1)^\dagger\right]_{ijkl} d\Omega = \frac{4\pi}{3}\delta_{ik}\delta_{jl} \tag{B5}$$

and, hence,

$$\int_{S_1} \left(\hat{\mathbf{r}}\hat{\nabla}\mathbf{P}_2\right)_{ijkl} d\Omega = \frac{2\pi}{5}\left(3\delta_{il}\delta_{jk} + 3\delta_{ik}\delta_{jl} - 2\delta_{ij}\delta_{kl}\right). \tag{B6}$$

Since $\mathbf{B}$ is a traceless, symmetric dyadic, $\delta_{il}\delta_{jk} : \mathbf{B} = \delta_{ik}\delta_{jl} : \mathbf{B} = \mathbf{B}$, and $\delta_{ij}\delta_{kl} : \mathbf{B} = \mathrm{Tr}(\mathbf{B})\mathbf{I} = \mathbf{0}$.

Similarly,

$$\begin{aligned}\int_{S_1} \hat{\nabla}\mathbf{P}_2 : \mathbf{B}\hat{\mathbf{r}} d\Omega = {} & \frac{9}{10}\int_{S_1} \mathbf{I}\mathbf{P}_1 : \mathbf{B}\mathbf{P}_1 d\Omega + \frac{9}{10}\int_{S_1} (\mathbf{I}\mathbf{P}_1)^\dagger : \mathbf{B}\mathbf{P}_1 d\Omega \\ & -\frac{6}{5}\int_{S_1} \mathbf{P}_3 : \mathbf{B}\mathbf{P}_1 d\Omega - \frac{3}{5}\int_{S_1} \mathbf{P}_1\mathbf{I} : \mathbf{B}\mathbf{P}_1 d\Omega.\end{aligned} \tag{B7}$$

Again, we evaluate (B7) in Cartesian tensor notation to obtain

$$\int_{S_1} \hat{\nabla}\mathbf{P}_2 : \mathbf{B}\hat{\mathbf{r}} d\Omega = \frac{12\pi}{5}\mathbf{B}. \tag{B8}$$

Finally,

$$\int_{S_1} (\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}})\, d\Omega = \frac{8\pi}{3}\mathbf{I}. \tag{B9}$$

In invariant notation, we finally obtain, upon substitution of (B1), (B3), (B6), (B8) and (B9) in (55), the result for $\mathbf{K}^{(1)}$ given in (56).

REFERENCES

[1] J. L. Anderson, *Colloid transport by interfacial forces*, Annu. Rev. Fluid Mech., 21 (1989), pp. 61–99.

[2] J. R. Bielenberg and H. Brenner, *A hydrodynamic/Brownian motion model of thermal diffusion in liquids*, Physica A, 356 (2005), pp. 279–293.

[3] H. Brenner, *The Stokes resistance of a slightly deformed sphere*, Chem. Engrg. Sci., 19 (1964), pp. 519–539.

[4] H. Brenner, *The Stokes resistance of an arbitrary particle—*IV. *Arbitrary fields of flow*, Chem. Engrg. Sci., 19 (1964), pp. 703–727.

[5] H. Brenner, *The translational and rotational motions of an n-dimensional hypersphere through a viscous fluid at small Reynolds numbers*, J. Fluid Mech., 111 (1981), pp. 197–215.

[6] H. Brenner and L. J. Gajdos, *London–van der Waals forces and torques exerted on an ellipsoidal particle by a nearby semi-infinite slab*, Canad. J. Chem., 59 (1981), pp. 2004–2018.

[7] S. Chapman and T. G. Cowling, *The Mathematical Theory of Non-uniform Gases: An Account of the Kinetic Theory of Viscosity, Thermal Conduction and Diffusion in Gases*, Cambridge University Press, Cambridge, UK, 1970.
[8] P. S. Epstein, *Zur theorie des radiometers*, Z. Phys., 54 (1929), pp. 537–563.
[9] V. S. Galkin, M. N. Kogan, and O. G. Fridlender, *Some kinetic effects in continuum flows*, Fluid Dynam., 5 (1973), pp. 364–371.
[10] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics*, Prentice Hall, Englewood Cliffs, NJ, 1965.
[11] M. N. Kogan, *Rarefied Gas Dynamics,* Plenum Press, New York, 1969.
[12] M. N. Kogan, *Molecular gas dynamics*, Annu. Rev. Fluid Mech., 5 (1973), pp. 383–404.
[13] M. N. Kogan, V. S. Galkin, and O. G. Fridlender, *Stresses produced in gases by temperature and concentration inhomogeneities. New types of free convection*, Sov. Phys. Usp., 19 (1976), pp. 420–428.
[14] D. W. Mackowski, *Phoretic behavior of asymmetric particles in thermal nonequilibrium with the gas: Two-sphere aggregates*, J. Colloid Interface Sci., 140 (1990), pp. 138–157.
[15] T. M. MacRobert, *Spherical Harmonics: An Elementary Treatise on Harmonic Functions With Applications*, Dover Publications, New York, 1948.
[16] J. C. Maxwell, *On stresses in rarefied gases arising from inequalities of temperature*, Philos. Trans. R. Soc. Lond., 170 (1879), pp. 231–256.
[17] A. Mohan and H. Brenner, *An extension of Faxen's laws for nonisothermal flow around a sphere*, Phys. Fluids, 17 (2005), 038107.
[18] F. A. Morrison, *Electrophoresis of a particle of arbitrary shape*, J. Colloid Interface Sci., 34 (1970), pp. 210–214.
[19] A. Regazzerri, M. Hoyos, and M. Martin, *Experimental evidence of thermophoresis of non-Brownian particles in pure liquids and estimation of their thermophoretic mobility*, J. Phys. Chem. B, 108 (2004), pp. 15285–15292.
[20] D. L. Ripps and H. Brenner, *The Stokes resistance of a slightly deformed sphere*—II. *Intrinsic resistance operators for an arbitrary initial flow*, Chem. Engrg. Sci., 22 (1967), pp. 375–387.
[21] M. E. Schimpf and S. N. Semenov, *Mechanism of polymer thermophoresis in nonaqueous solvents*, J. Phys. Chem. B, 104 (2000), pp. 9935–9942.
[22] M. E. Schimpf and S. N. Semenov, *Latex particle thermophoresis in polar solvents*, J. Phys. Chem. B, 105 (2001), pp. 2285–2290.
[23] S. N. Semenov, *Mechanism of particle thermophoresis in pure solvents*, Philos. Mag., 83 (2003), pp. 2199–2208.
[24] M. M. R. Williams, *Thermophoretic forces acting on a spheroid*, J. Phys. D: Appl. Phys., 19 (1986), pp. 1631–1642.
[25] M. M. R. Williams, *The thermophoretic forces acting on a bispherical system*, J. Phys. D: Appl. Phys., 20 (1987), pp. 354–359.
[26] F. Zheng, *Thermophoresis of spherical and non-spherical particles: A review of theories and experiments*, Adv. Colloid Interface Sci., 97 (2002), pp. 255–278.

SIAM J. APPL. MATH.
Vol. 66, No. 3, pp. 802–819

MULTIPLE FOCUS AND HOPF BIFURCATIONS IN A PREDATOR-PREY SYSTEM WITH NONMONOTONIC FUNCTIONAL RESPONSE*

DONGMEI XIAO† AND HUAIPING ZHU‡

Abstract. In this paper, we develop a criterion to calculate the multiplicity of a multiple focus for general predator-prey systems. As applications of this criterion, we calculate the largest multiplicity of a multiple focus in a predator-prey system with nonmonotonic functional response $p(x) = \frac{x}{ax^2+bx+1}$ studied by Zhu, Campbell, and Wolkowicz [*SIAM J. Appl. Math.*, 63 (2002), pp. 636–682] and prove that the degenerate Hopf bifurcation is of codimension two. Furthermore, we show that there exist parameter values for which this system has a unique positive hyperbolic stable equilibrium and exactly two limit cycles, the inner one unstable and outer one stable. Numerical simulations for the existence of the two limit cycles bifurcated from the multiple focus are also given in support of the criterion.

Key words. predator-prey, Liénard system, multiple focus, Hopf bifurcation, codimension two, limit cycles

AMS subject classifications. Primary, 34C25, 92D25; Secondary, 58F14

DOI. 10.1137/050623449

1. Introduction.

The existence and number of limit cycles are important topics in the study of most applied mathematical models. Such study has made possible a better understanding of many real world oscillatory phenomena in nature [1, 11, 17]. For predator-prey systems, it is well known that the existence of limit cycles is related to the existence, stability, and bifurcation of a positive equilibrium. In a positively invariant region, if there exists a unique positive equilibrium which is unstable, then there must exist at least one limit cycle according to the theory of Poincaré-Bendixson. On the other hand, if the unique positive equilibrium of a predator-prey system is locally stable but not hyperbolic, there might be more than one limit cycle created via Hopf bifurcation(s). Numerical simulations of Hofbauer and So [8] indicated that this is indeed the case: there can exist at least two limit cycles for some two-dimensional predator-prey systems. It was proved by Zhu, Campbell, and Wolkowicz [26] that a predator-prey system with nonmonotonic functional response can undergo a degenerate Hopf bifurcation which produces two limit cycles, and the system can also have two limit cycles through a saddle node bifurcation of limit cycles. These two limit cycles can disappear through either supercritical Hopf or/and homoclinic bifurcations. Kuang [9] and Wrzosek [22] also observed that some predator-prey systems can even have more than two limit cycles. We shall point out that in all the models studied in [8, 9, 22] the death rate of the predator is nonlinear, but ours in [19, 26] is linear.

Hopf bifurcation theory is a powerful tool for studying the existence, number, and properties of the limit cycles in mathematical biology. However, the largest number

*Received by the editors January 28, 2005; accepted for publication (in revised form) August 31, 2005; published electronically February 15, 2006.

http://www.siam.org/journals/siap/66-3/62344.html

†Department of Mathematics, Shanghai Jiao Tong University, Shanghai 200030, China (xiaodm@sjtu.edu.cn). The research of this author was supported by the National Natural Science Foundations of China (10231020).

‡Corresponding author. Department of Mathematics and Statistics, York University, Toronto, ON, Canada M3J 1P3 (huaiping@mathstat.yorku.ca). The research of this author was supported by NSERC, CFI, and MITACS of Canada.

of limit cycles which can be created via a Hopf bifurcation is determined by the multiplicity of a multiple focus [2, 14] or correspondingly the codimension of the Hopf bifurcation. Therefore, the resolution of the multiplicity of a multiple focus plays a key role in determining the number of limit cycles for predator-prey systems.

In this paper, we first consider a general predator-prey system, taken to have the form

$$\begin{cases} \dot{x} = a(x) - b(y)p(x), \\ \dot{y} = c(y)q(x), \end{cases} \tag{1.1}$$

which has a positive nondegenerate equilibrium in the first quadrant and is a center-type equilibrium. A predator-prey system in the form of (1.1) often involves the Hopf bifurcation(s), and it is essential to identify the multiplicity of a multiple focus in order to determine the codimension of the Hopf bifurcation. There are formulas available for calculating the Lyapunov coefficients [2, 3, 10, 14], which can be used to decide the multiplicity, yet such calculations usually become very challenging if the multiple focus has multiplicity greater than or equal to two. In this paper, we shall first transform the general predator-prey system (1.1) into a generalized Liénard-type system. Then a criterion will be established for calculating the multiplicity of a focus for the general predator-prey system.

As applications of this criterion, following [26] we continue our study on the predator-prey systems with nonmonotonic functional response of the form

$$\begin{cases} \dot{x} = rx\Big(1 - \dfrac{x}{K}\Big) - yp(x), \\ \dot{y} = y\big(-d + cp(x)\big), \end{cases} \tag{1.2}$$

where x and y are functions of time representing population densities of prey and predator, respectively; $r > 0$ is the maximum growth rate, and $K > 0$ is the carrying capacity of the prey; $d > 0$ is the death rate of the predator. The functional response is of Holling type IV, and

$$p(x) = \frac{mx}{ax^2 + bx + 1}. \tag{1.3}$$

Here we write $p(x)$ as in [26] with $b > -2\sqrt{a}$ such that functional response $p(x)$ remains nonnegative for all $x \geq 0$. Note that if $b = 0$, $p(x)$ is reduced to the function used in [19]. For detailed biological interpretation and motivation of the model, interested readers may consult the work [19, 26] and the references therein.

Global qualitative and bifurcation analysis by Rothe and Shafer [18] and Zhu, Campbell, and Wolkowicz [26] show that system (1.2) with the functional response (1.3) has very interesting and rich dynamics. More precisely, by allowing b to be negative but $b > -2\sqrt{a}$, $p(x)$ is concave up for small values of $x > 0$ as it is for the sigmoidal functional response. It was shown in [26] that there exists a Bogdanov–Takens bifurcation point of codimension 3, which acts as an organizing center for the bifurcation diagram described in (b, d, K) space with $d, K > 0$ and $b > -2\sqrt{a}$. Then the bifurcation sequences were given for parameters (b, K) in each meaningful subregion of the (b, K) plane and the death rate of the predator, d, was varied. The bifurcation sequences involving Hopf bifurcations and homoclinic bifurcations as well as the saddle node bifurcations of limit cycles are determined using information from the study of the Bogdanov–Takens bifurcation of codimension 3 and an exhaustive

utilization of the geometry of isolines of the system. In particular, the Hopf bifurcation of codimension one was completely described in the parameter space (b, d, K) with $d, K > 0$ and $b > -2\sqrt{a}$. Also, a curve segment, or the so-called degenerate Hopf bifurcation curve, was defined in [26]. But there is one thing missing: when the parameters took the value along the degenerate Hopf bifurcation curve, the order of the Hopf bifurcation or the multiplicity of the multiple focus was left unknown, and for the parameters (b, K) in some subset, the bifurcation sequences were not complete (cf. Theorem 6.22 in [26]).

In this paper, we shall fill the gap of [26] and complete the study of the multiple focus and the degenerate Hopf bifurcation. We focus on the case when system (1.2) has a positive equilibrium. We shall prove that the equilibrium is a multiple focus of multiplicity at most two and can be exactly two. By using the Hopf bifurcation theorem, and combining the existing results regarding the Hopf bifurcations in [26], we shall give a complete description of a bifurcation diagram for Hopf bifurcation of codimension two. Furthermore, we investigate the existence of exactly two limit cycles of system (1.2). To the best of our knowledge, there are no references in which the existence of exactly two limit cycles in any predator-prey system has been rigorously proved. The difficulty is that, though there are theorems on the existence of exactly two limit cycles for certain Liénard-type systems (Zhang [24] and Zhou [25]), they are not applicable to the predator-prey systems that may have two limit cycles.

This paper is organized as follows. In section 2 we develop a criterion of the multiplicity of a multiple focus for the general system (1.1). System (1.2) with the functional response (1.3) is a typical example of the general predator-prey system with nonmontonic functional response. As applications, we will study (1.2) by focusing on the calculation of the multiplicity of the multiple focus and the description of the degenerate Hopf bifurcation. We shall prove that the multiplicity of the unique multiple focus is at most two and that there exist parameter values such that system (1.2) with the functional response (1.3) has a multiple focus whose multiplicity is exactly two. We shall also prove that system (1.2) with the functional response (1.3) has exactly two limit cycles for some parameter values. A Hopf bifurcation diagram involving the Hopf bifurcation of codimension two will also be given based on the results related to the Hopf bifurcation of codimension one from [26]. We end section 3 by giving numerical simulations of the two limit cycles. The paper ends with a brief discussion in section 4.

2. The multiplicity of a multiple focus in a general predator-prey system. In [21] Wolkowicz proposed a general form of the predator-prey model for studying the impact of the group defense. And Kazarinoff and van den Driessche in [13] studied a predator-prey system by incorporating a fairly general functional response. Here, we consider a general predator-prey system of the form

$$\begin{cases} \dot{x} = a(x) - b(y)p(x), & x(0) \geq 0, \\ \dot{y} = c(y)q(x), & y(0) \geq 0, \end{cases} \tag{2.1}$$

where x and y are functions of t representing the density of prey and predator populations, respectively, at a given time $t \geq 0$. Based on the biological meaning, we will restrict ourselves to the first quadrant, and in particular we make the following assumptions:

(A1) the growth function of the prey $a(x)$ is C^1 for $x \in [0, +\infty)$ and $a(0) = 0$;

(A2) $b(y)$, $c(y) \in C^1[0, +\infty)$, $b(0) = c(0) = 0$ and $b'(y) > 0$, $c'(y) > 0$ for all $y \geq 0$;

(A3) $p(x) \in C^1[0, +\infty)$, $p(0) = 0$, and $p(x) > 0$ for all $x \geq 0$;
(A4) $q(x) \in C^1[0, +\infty)$, and there exists $x_0 > 0$ such that

$$q(x_0) = 0, \quad \frac{a(x_0)}{p(x_0)} > 0, \quad \text{and } q'(x_0) > 0.$$

With the above assumptions, one can verify that both the x- and y-axes are invariant. Therefore, the first quadrant is positively invariant. Furthermore, a straightforward computation can verify that if the assumptions (A1)–(A4) hold, then system (2.1) has a positive equilibrium (x_0, y_0), which is nondegenerate, where $y_0 = b^{-1}(a(x_0)/p(x_0))$ and b^{-1} is the inverse function of $b(y)$. Thus, there exists a neighborhood Ω of (x_0, y_0) in the first quadrant such that system (2.1) has no other equilibria except (x_0, y_0) in Ω. It is clear that in the generic case any predator-prey system can have a positive equilibrium, which is a focus or a center or a node or a saddle point. Note that by (A2) we have $c(y_0) > 0$; hence in order for the positive equilibrium (x_0, y_0) to be a center-type focus, we have to further assume the following:

(A5) $a'(x_0) = b(y_0)p'(x_0)$.

Then, under assumptions (A1)–(A5), system (2.1) has an isolated positive equilibrium (x_0, y_0) in Ω, which is a center-type equilibrium.

Recall that many of the classical predator-prey systems can be written in the form of (2.1) with assumptions (A1)–(A5) satisfied. For example, the predator-prey system with response function of Holling type falls into this category [6, 19, 13, 21, 26]. There have been extensive studies on stability and bifurcations for predator-prey systems (see [8, 9, 18, 19, 21, 22, 23, 26] and references therein), but it is not an easy task to study the number of limit cycles which can be born through the bifurcation of a center-type focus, and the question remains unanswered for the predator-prey systems (2.1). It is well known that there are formulas available to calculate the first Lyapunov coefficient and the higher order Lyapunov coefficients (e.g., cf. [2, 14]); however, in general it is technically very challenging to draw conclusions from the formula directly due to its complexities. Pilyugin and Waltman [15, 16] develop a divergence criterion for a generic planar system, and the applications were made successfully to the study of multiple limit cycles in the chemostat with variable yield and other planar systems.

In this paper, we are going to develop some criteria to determine the multiplicity and stability of the multiple focus for the general predator-prey system (2.1). We can do this thanks to the magic of the Liénard-type system, which has been playing an increasingly important role in current research on the existence and uniqueness of limit cycles for predator prey systems (cf. [23] and reference therein). Hence it is not surprising that the equivalent form of (1.1), a Liénard-type system, can be simpler in calculating the multiplicity for a focus.

Consider the general Liénard-type system

$$\begin{cases} \dfrac{dx}{dt} = \phi(y) - F(x), \\ \dfrac{dy}{dt} = -g(x), \end{cases} \tag{2.2}$$

where $xg(x) > 0$ for $x \neq 0$. We first introduce a very useful lemma of [7] by Han.

LEMMA 2.1. *Suppose that $\phi(y)$, $F(x)$, and $g(x)$ are C^∞ smooth functions in a neighborhood of the origin, and that*

$$\phi(0) = g(0) = F(0) = F'(0) = 0, \quad \phi'(0) > 0, \quad \textit{and} \quad g'(0) > 0.$$

Let $G(x) = \int_0^x g(s)ds$. *If there exists a* C^∞ *smooth function* $\alpha(x)$, $\alpha(x) = -x+O(x^2)$, *such that* $G(\alpha(x)) \equiv G(x)$ *and*

$$F(\alpha(x)) - F(x) = \sum_{i\geq 1} B_i x^i,$$

then the equilibrium $(0,0)$ *of* (2.2) *is a multiple focus of multiplicity* k *if* $B_j = 0$, $j = 1, 2, \ldots, 2k$, *and* $B_{2k+1} \neq 0$. *Furthermore, it is locally stable (unstable) if* $B_{2k+1} < 0$ $(B_{2k+1} > 0$, *respectively).*

The key to the proof of Lemma 2.1 is to transform (2.2) into

$$\begin{cases} \dfrac{du}{d\tau} &= v - K(v)F^*(u), \\ \dfrac{dv}{d\tau} &= -u \end{cases} \tag{2.3}$$

by a C^∞ transformation of variables (x, y) near the origin and time t. Then let

$$F_e(u) = \frac{1}{2}(F^*(u) + F^*(-u)), \qquad F_o(u) = \frac{1}{2}(F^*(u) - F^*(-u)).$$

By the principle of symmetry, the orbits of the system

$$\begin{cases} \dfrac{du}{d\tau} &= v - K(v)F_e(u), \\ \dfrac{dv}{d\tau} &= -u \end{cases} \tag{2.4}$$

near the origin are symmetric with respect to the v-axis. In a small neighborhood of the origin, introducing the polar coordinates $u = r\cos\theta$ and $v = r\sin\theta$ to (2.3) and (2.4), one can get two equations:

$$\frac{dr}{d\theta} = \frac{\cos\theta K(r\sin\theta)F^*(r\cos\theta)}{1 - \sin\theta K(r\sin\theta)F^*(r\cos\theta)/r} \tag{2.5}$$

and

$$\frac{dr}{d\theta} = \frac{\cos\theta K(r\sin\theta)F_e(r\cos\theta)}{1 - \sin\theta K(r\sin\theta)F_e(r\cos\theta)/r}, \tag{2.6}$$

respectively. According to the classical method of Lyapunov, we can define the displacement map $d(r_0)$ of (2.5) as

$$d(r_0) = r\left(-\frac{3\pi}{2}, r_0\right) - r\left(\frac{\pi}{2}, r_0\right)$$

for $0 < r_0 \ll 1$. From the technical analysis of $d(r_0)$, the conclusions of the lemma can be obtained. For more details on the proof of the lemma and its applications, we refer the reader to [7] by Han.

We next utilize the technique developed in [23] to transform the general predator-prey system (2.1) into a Liénard-type system, and use Lemma 2.1 to establish conditions to determine the multiplicity of the center-type focus. We focus our attention on system (2.1) with assumptions (A1)–(A5) in the open set Ω.

Since $p(x) > 0$, rescaling the time t of system (2.1) by

$$\tau = \int_0^t p(x(s))ds, \tag{2.7}$$

we obtain

$$\begin{cases} \dfrac{dx}{d\tau} &= \dfrac{a(x)}{p(x)} - b(y), \\ \dfrac{dy}{d\tau} &= c(y)\dfrac{q(x)}{p(x)}. \end{cases} \tag{2.8}$$

System (2.8) has an equilibrium at (x_0, y_0). Let

$$x = -u + x_0, \qquad y = h(v) + y_0; \tag{2.9}$$

we translate the equilibrium (x_0, y_0) of system (2.8) to the origin, where $h(v)$ is a solution of the following initial problem:

$$\frac{dh(v)}{dv} = c(h(v) + y_0), \qquad h(0) = 0. \tag{2.10}$$

Since $c'(y) > 0$, $h(v)$, the solution to the initial value problem (2.10) exists for $v > 0$ and is unique. Thereby the inverse function of $h(v)$ exists. Let us denote the inverse function by h^{-1}. Thus, there exists an inverse transformation of (2.9) such that system (2.8) in Ω can be transformed to

$$\begin{cases} \dfrac{du}{d\tau} &= [b(h(v) + y_0) - b(y_0)] - \left[\dfrac{a(-u + x_0)}{p(-u + x_0)} - b(y_0)\right], \\ \dfrac{dv}{d\tau} &= -\left[-\dfrac{q(-u + x_0)}{p(-u + x_0)}\right], \end{cases} \tag{2.11}$$

in the neighborhood of the origin. If we define

$$\begin{aligned} \phi(v) &= b(h(v) + y_0) - b(y_0), \\ F(u) &= \frac{a(-u + x_0)}{p(-u + x_0)} - b(y_0), \\ g(u) &= -\frac{q(-u + x_0)}{p(-u + x_0)}, \end{aligned}$$

then system (2.11) becomes

$$\begin{cases} \dfrac{du}{d\tau} &= \phi(v) - F(u), \\ \dfrac{dv}{d\tau} &= -g(u). \end{cases} \tag{2.12}$$

System (2.12) is a Liénard-type system in the neighborhood of the origin, and it can be observed that

$$\phi(0) = g(0) = F(0) = F'(0) = 0, \quad \phi'(0) > 0, \quad \text{and } g'(0) > 0.$$

Applying Lemma 2.1 to system (2.12), we obtain the following result.

THEOREM 2.2. *Assume that* (A1)–(A5) *hold for system* (2.1). *Suppose that* $b(y)$ *and* $c(y)$ *are* C^∞ *functions of* y *in a neighborhood of* y_0, *and that* $a(x)/p(x)$ *and* $q(x)/p(x)$ *are* C^∞ *functions of* x *in a neighborhood of* x_0. *Let*

$$G(x) = \int_0^x -\frac{q(-u+x_0)}{p(-u+x_0)}du. \tag{2.13}$$

If there exists a C^∞ *function* $\alpha(x)$, $\alpha(x) = -x + O(x^2)$, *such that* $G(\alpha(x)) \equiv G(x)$ *and*

$$F(\alpha(x)) - F(x) = \frac{a(-\alpha(x)+x_0)}{p(-\alpha(x)+x_0)} - \frac{a(-x+x_0)}{p(-x+x_0)} = \sum_{i\geq 1} B_i x^i, \tag{2.14}$$

and if $B_j = 0$, $j = 1, 2, \ldots, 2k$, *and* $B_{2k+1} \neq 0$, *then the equilibrium* (x_0, y_0) *of* (2.1) *is a multiple focus of multiplicity* k, *which is locally stable (unstable) if* $B_{2k+1} < 0$ *(*$B_{2k+1} > 0$, *respectively).*

3. Multiple focus and degenerate Hopf bifurcation. Now we consider system (1.2). From the work of [18, 4] to [19] and [26] and the references therein, there have been extensive studies of various bifurcations for system (1.2). It follows from [26] that system (1.2) undergoes bifurcations including saddle node bifurcation, Hopf bifurcation(s), homoclinic bifurcation, and saddle node bifurcation of limit cycles. Note that system (1.2) with the functional response (1.3) involves an extra parameter b. As indicated in [26], by varying this parameter b one can connect all the bifurcation branches to the organizing center, which is a Bogdanov–Takens bifurcation of codimension three. Although the bifurcation study for (1.2) in [26] is almost exhaustive, the codimension of the degenerate Hopf bifurcation was left untouched, and the multiplicity of the multiple focus is unknown.

It follows from the discussion in [26] that by rescaling the state variables and time, one can eliminate three parameters from the system (1.2). For the purpose of combining the results from [19] and [26] regarding the bifurcations, here we eliminate a, m, and c by using

$$\begin{aligned} &(t,\ x,\ y) \longrightarrow \left(\frac{\sqrt{a}}{mc}t,\ \frac{1}{\sqrt{a}}x,\ \frac{c}{\sqrt{a}}y\right), \\ &(r,\ K,\ b,\ d) \longrightarrow \left(\frac{mc}{\sqrt{a}}r,\ \frac{1}{\sqrt{a}}K,\ \sqrt{a}b,\ \frac{1}{\sqrt{a}}d\right). \end{aligned} \tag{3.1}$$

Then system (1.2) with the functional response (1.3) becomes

$$\begin{cases} \dot{x} = rx\Big(1 - \dfrac{x}{K}\Big) - \dfrac{xy}{x^2+bx+1}, \\ \dot{y} = y\Big(-d + \dfrac{x}{x^2+bx+1}\Big). \end{cases} \tag{3.2}$$

System (3.2) involves four parameters b, d, K, and r, where $r > 0$ is not a bifurcation parameter for the equilibria. It follows from the results in [26] that we need three parameters to unfold the Bogdanov–Takens bifurcation of codimension three, and hence it is natural to describe the degenerate Hopf bifurcation curve in the parameter space (b, d, K). We start first by summarizing the results from [26] regarding the Hopf bifurcations for system (3.2), then we apply Theorem 2.2 to determine the multiplicity of the multiple focus and complete the diagram of Hopf bifurcation.

In [26], the geometry of the isoclines plays an important role in understanding both the equilibria and their bifurcations. Denote the prey isoline of (3.2) as (still using the notation in [26] but with a hat)

$$\hat{F}(x) = r\left(1 - \frac{x}{K}\right)(x^2 + bx + 1). \tag{3.3}$$

It follows from the linear stability analysis in [26] that besides the equilibria at the origin and $(K, 0)$, if $d < \frac{1}{b+2}$, system (3.2) may have up to two possible positive equilibria $(x_i, y_i = \hat{F}(x_i))$ $(i = 1, 2, x_1 < x_2)$ if x_1 and x_2 are positive and smaller than K, where

$$\begin{aligned} x_1 &= \frac{1 - bd - \sqrt{(b^2-4)d^2 - 2bd + 1}}{2d}, \\ x_2 &= \frac{1 - bd + \sqrt{(b^2-4)d^2 - 2bd + 1}}{2d}. \end{aligned} \tag{3.4}$$

Note that x_1 and x_2 are the real roots of the quadratic equation

$$\hat{g}(x) = dx^2 - (1 - bd)x + d = 0, \tag{3.5}$$

and we also have

$$x_1 + x_2 = \frac{1 - bd}{d}, \qquad x_1 x_2 = 1. \tag{3.6}$$

The equilibrium (x_2, y_2) is always a hyperbolic saddle if $0 < x_1 < x_2 < K$. The stability of (x_1, y_1) as a hyperbolic focus (or node) can be found in [26]. Here we are interested only in the equilibrium (x_1, y_1) and related Hopf bifurcations. Recall from [26] that (x_1, y_1) may undergo a Hopf bifurcation if $\hat{F}'(x_1) = 0$. The prey isocline can have two humps $(H_m, \hat{F}(H_m))$ and $(H_M, \hat{F}(H_M))$, where

$$\begin{aligned} H_m &= \frac{1}{3a}\left[aK - b - \sqrt{a^2K^2 + abK + b^2 - 3a}\,\right], \\ H_M &= \frac{1}{3a}\left[aK - b + \sqrt{a^2K^2 + abK + b^2 - 3a}\,\right]. \end{aligned} \tag{3.7}$$

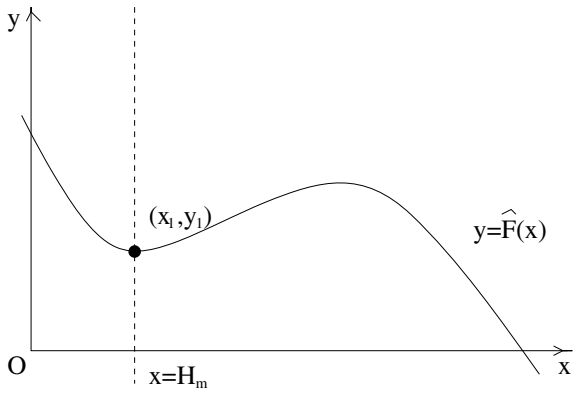
As shown in Figure 1, when a Hopf bifurcation occurs, it occurs at either the left hump $(H_m, \hat{F}(H_m))$ or the right hump $(H_M, \hat{F}(H_M))$. By (3.6) we know that $x_1 \le 1$. Hence for the Hopf bifurcation to occur at a hump of $\hat{F}(x)$, the hump cannot be at the right of the vertical line $x = 1$ (Corollary 4.2 in [26]).

If the Hopf bifurcation occurs, it occurs at one of the humps of $\hat{F}(x)$ or where $\hat{F}'(x) = 0$. Hence the Hopf bifurcation can only occur at either $x_1 = H_m$ or $x_1 = H_M$. Note that we can also solve $\hat{F}'(x_1) = 0$ in terms of K; then when the Hopf bifurcation occurs at $x = x_1$, we should also have

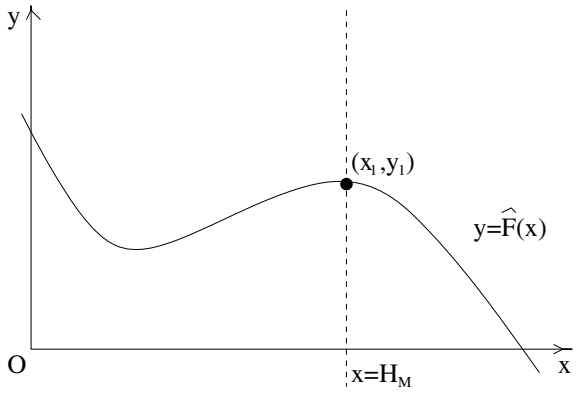
$$K = \frac{1 + 2bx_1 + 3x_1^2}{b + 2x_1}. \tag{3.8}$$

Substituting x_1 into (3.8) or eliminating x from $\hat{g}(x) = 0$ and $\hat{F}'(x) = 0$, one gets the Hopf bifurcation surface (equation (4.2) in [26]) in the parameter space (b, K, d):

$$\Sigma_H : (4 - b^2)(K^2 + bK + 1)d^2 + 2(bK^2 + 2(b^2 - 2)K + b)d + 3(1 - bK) = 0. \tag{3.9}$$



(a) At the left hump, the Hopf bifurcation can be both subcritical and supercritical



(b) At the right hump, the Hopf bifurcation is always supercritical

FIG. 1. *Hopf bifurcation occurs at the humps of the prey isocline.*

Solving for d, we obtain

$$d_{\pm}(b,K) = \frac{-(bK^2+2(b^2-2)K+b) \pm (2+bK)\sqrt{K^2+bK+b^2-3}}{(4-b^2)(K^2+bK+1)}. \tag{3.10}$$

Hence for any (b,K) such that $\hat{F}(x)$ has an hump to the left of the vertical line $x=1$, if $d=d_-$ or $d=d_+$, the Hopf bifurcation may occur. By using the formula given by Wolkowicz [21], the first Lyapunov coefficient was calculated in [26], and

$$\sigma(x) = -\frac{p(x)\hat{F}''(x)p''(x)}{p'(x)} + p(x)\hat{F}'''(x) + 2p'(x)\hat{F}''(x). \tag{3.11}$$

It was then proved that for the Hopf bifurcation occurring at the right hump, i.e., $x_1 = H_M$, we have $\sigma(H_M) < 0$. Hence if the Hopf bifurcation occurs at the right

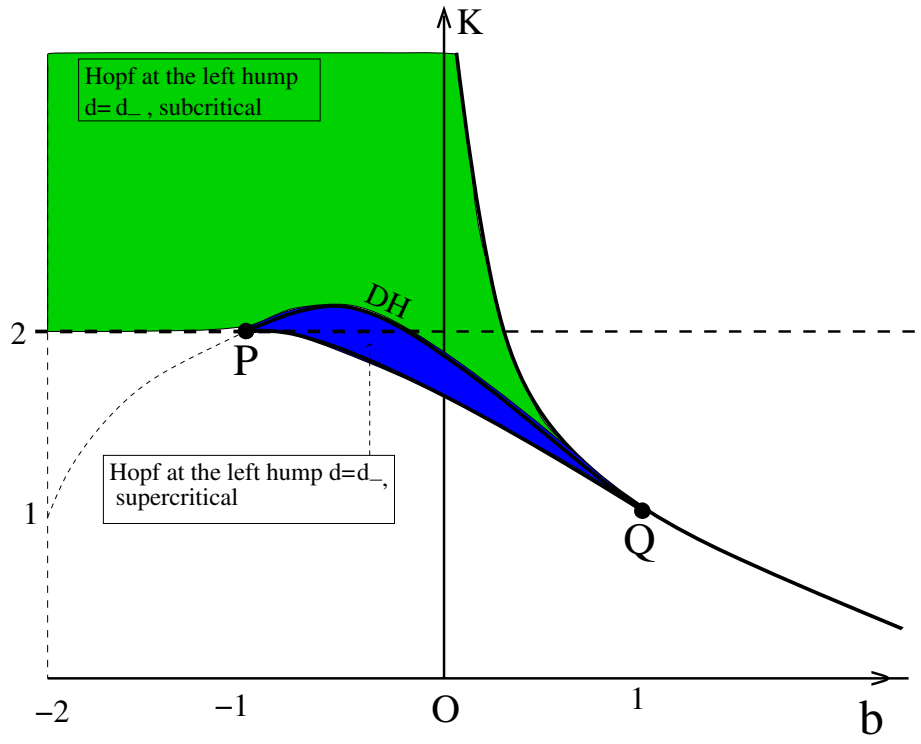


FIG. 2. *Hopf bifurcation occurs at the left hump when* $d = d_-(b, K)$. *The curve segment* DH, *the projection of the curve segment* C_{DH} *into the* bK *plane, is plotted. It separates the colored region into two: for* (b, k) *in the blue (darkest) region and* $d = d_-$, *the Hopf bifurcation at the left hump is supercritical; for* (b, k) *in the green (lighter shaded) region and* $d = d_-$, *the Hopf bifurcation at the left hump is subcritical* [26]. *Here* C_1: $K = \frac{1}{2}[\sqrt{3(4-b^2)} - b]$, $-2 \le b \le 1$; C_2: $K = \frac{1}{b}$, $b > 0$. *For* (b, K) *below* C_1, $\hat{F}(x)$ *does not have any humps; to the right of* C_2, *there is only the right hump in the positive cone. Above the line* $K = 2$, *only one hump is to the left of* $x = 1$.

hump, it is always supercritical. For the Hopf bifurcation occurring at the left hump, when $d = d_-$, it can be both subcritical and supercritical.

Plugging x_1 from (3.4) into $\sigma(x_1) = 0$, in [26] a degenerate Hopf bifurcation curve was defined. Let

(3.12)
$$C_{DH} = \left\{ (b, d, K) \;\middle|\; \begin{array}{l} d = d_-(b, K) \ \text{ and} \\ \quad 16K^4 + b(8 - 3b^2)K^3 - (144 - 15b^2)K^2 \\ \quad -8b(9 - b^2)K + 16b^4 - 144b^2 + 300 = 0, \end{array} \quad -1 < b < 1 \right\}.$$

Then it follows from the results in [26] that C_{DH} is a curve segment in the Hopf bifurcation surface Σ_H, and the projection of the curve onto the (b, K) plane was plotted in Figure 2; it is a curve segment with endpoints at $P(-1, 2)$ and $Q(1, 1)$. In the next proposition, we summarize the result regarding the Hopf bifurcation(s) of codimension one from [26].

PROPOSITION 3.1. *If the Hopf bifurcation occurs at the right hump when* $d = d_+$, *it is always supercritical. For the Hopf bifurcation occurring at the left hump when* $d = d_-$, *as shown in Figure* 2, *it is subcritical if* (b, K) *is above* DH *and is supercritical if* (b, K) *is below* DH. *For* $(b, K) \in DH$ *and* $d = d_-$, *a degenerate Hopf bifurcation*

occurs.

By an easy computation from (3.8) and (3.4), when a Hopf bifurication occurs, we have

$$\begin{aligned} K &= K_0(b,d) = \frac{2-\sqrt{(1-bd)^2-4d^2}}{1-\sqrt{(1-bd)^2-4d^2}}x_1 \\ &= \left(1+\frac{1}{1-\sqrt{(1-bd)^2-4d^2}}\right)x_1. \end{aligned} \tag{3.13}$$

Since the equilibrium (x_1, y_1) is isolated in the interior of the first quadrant, there exists a neighborhood Ω of (x_1, y_1) such that system (3.2) has no other equilibria in Ω except (x_1, y_1). Next we investigate the multiplicity of the multiple focus (x_1, y_1) of system (3.2) when $(b, d, K) \in C_{DH}$ or, equivalently, when $K = K_0(b, d)$ and $d = d_-$. Note that in manipulating the expressions for conclusion, the relation (3.6) of x_1 and x_2 with the coefficients of the quadratic equation $\hat{g}(x) = 0$ will be repeatedly used.

Taking $h(v) = y_1(e^v-1)$, we transform system (3.2) in Ω to a Liénard-type system by transformations (2.7) and (2.9) when $K = K_0(b, d)$ and $d = d_-$. For simplicity, we still denote u, v, and τ by x, y, and t, respectively. Then system (1.2) can be written as

$$\begin{cases} \dfrac{dx}{dt} = \phi(y) - F(x), \\ \dfrac{dy}{dt} = -g(x), \end{cases} \tag{3.14}$$

where

$$\begin{aligned} &\phi(y) = y_1(e^y - 1), \\ &F(x) = \hat{F}(-x + x_1) - y_1 = r\left(\left(1 - \frac{b+3x_1}{K_0}\right)x^2 + \frac{x^3}{K_0}\right), \\ &g(x) = \frac{d}{p(-x+x_1)} - 1 = \frac{dx(x - x_1 + x_2)}{x_1 - x}, \end{aligned}$$

where $(x, y) \in \Omega_1$, a neighborhood of the origin which comes from Ω under the transformation (2.9).

It is clear that $\phi(y)$, $F(x)$, and $g(x)$ are C^∞ smooth functions in Ω_1, and

- $\phi(0) = g(0) = F(0) = F'(0) = 0$,
- $\phi'(0) = y_1 > 0$,
- $g'(0) = \frac{p'(x_1)}{p^2(x_1)} = d(\frac{x_2}{x_1} - 1) > 0$ and

$$G(x) = \int_0^x \left[\frac{d}{p(-u+x_1)} - 1\right] du = -d\left(\frac{1}{2}x^2 + x_2x + x_1x_2 \ln\frac{x_1 - x}{x_1}\right). \tag{3.15}$$

A straightforward computation can verify that there exists a C^∞ smooth function

$$\alpha(x) = -x + a_2x^2 + a_3x^3 + a_4x^4 + O(x^5) \tag{3.16}$$

such that $G(\alpha(x)) \equiv G(x)$, where

$$\begin{aligned} a_2 &= -\frac{G'''(0)}{3G''(0)} = -\frac{2(p'(x_1))^2 - p(x_1)p''(x_1)}{3p(x_1)p'(x_1)} = -\frac{2x_2}{3x_1(x_2-x_1)}, \\ a_3 &= -a_2^2, \\ a_4 &= a_2\left(2a_2^2 + \frac{3a_2}{2x_1} + \frac{3}{5x_1^2}\right). \end{aligned} \tag{3.17}$$

Performing a Taylor expansion of function $F(\alpha(x)) - F(x)$ at $x = 0$, we obtain

$$F_2(\alpha(x)) - F_2(x) = r\left(A_3x^3 + A_4x^4 + A_5x^5 + O(x^6)\right),$$

where

$$\begin{aligned} A_3 &= \frac{1}{3}[F'''(x_1) - 3a_2F''(x_1)] = -2a_2\left(1 - \frac{b+3x_1}{K_0}\right) - \frac{2}{K_0}, \\ A_4 &= \left(1 - \frac{b+3x_1}{K_0}\right)(a_2^2 - 2a_3) + \frac{3a_2}{K_0}, \\ A_5 &= \left(1 - \frac{b+3x_1}{K_0}\right)(2a_2a_3 - 2a_4) + \frac{3a_3 - 3a_2^2}{K_0}. \end{aligned} \tag{3.18}$$

It follows from (3.11) and (3.17) that

$$\begin{aligned} A_3 &= \frac{1}{3p(x_1)}\left[pF'''(x_1) + \frac{2(p'(x_1))^2 - p(x_1)p''(x_1)}{p'(x_1)}F''(x_1)\right] \\ &= \frac{1}{3p(x_1)}\sigma(x_1). \end{aligned} \tag{3.19}$$

Hence if $\sigma(x_1) \neq 0$, $A_3 \neq 0$, the equilibrium $(0,0)$ of system (3.14) is a multiple focus of multiplicity one.

Along the curve C_{DH} where $\sigma(x_1) = 0$, we have $A_3 = 0$. Then the equilibrium $(0,0)$ is a multiple focus of multiplicity at least two and we have to calculate $A_4, A_5, \ldots$ in order to determine the multiplicity of equilibrium $(0,0)$. For the purpose of identifying the sign for A_5, we need the fact that along DH, $A_3 = 0$, and it is equivalent to

$$x_2 = \frac{(3b+6x_1)x_1^2}{12x_1^2 + 9bx_1 + 2(b^2-1)}. \tag{3.20}$$

Assuming that $A_3 = 0$ and using the expressions of a_i $(i = 2,3,4)$ and condition (3.20), we compute

$$\begin{aligned} A_4 &= -\frac{1}{a_2K_0}\left(a_2^2 - 2a_3\right) + \frac{3a_2}{K_0} = 0, \\ A_5 &= -\frac{1}{a_2K_0}(2a_2a_3 - 2a_4) + \frac{3a_3 - 3a_2^2}{K_0} \\ &= \frac{3}{5K_0x_1}\left(5a_2 + \frac{2}{x_1}\right) \\ &= -\frac{2(3x_1 + 2x_2)}{5K_0x_1^2(x_2 - x_1)} < 0. \end{aligned} \tag{3.21}$$

Thus, equilibrium $(0,0)$ is a stable multiple focus of multiplicity two by Theorem 2.2. Summarizing the above arguments, we obtain the following.

THEOREM 3.2. *For $(b,d,K) \in C_{DH}$, or equivalently, if $(b,K) \in DH$ with $-1 < b < 1$ and $d = d_-$, the origin of system (3.14) (i.e., equilibrium (x_1,y_1) of system (3.2)) is a multiple focus of multiplicity two, which is locally asymptotically stable.*

From Theorem 3.2, we immediately have the next result regarding the degenerate Hopf bifurcation.

THEOREM 3.3. *For $(b,d,K) \in C_{DH}$, system (3.2) undergoes a Hopf bifurcation of codimension two, and the diagram for the Hopf bifurcation occurring at the left hump, Figure 2, is complete.*

Remark 3.1. It follows from the expression (3.21) that A_5 goes to infinity if x_1 and x_2 are getting close enough. This is consistent with the fact that the degenerate Hopf curve C_{DH} is connected to the Bogdanov–Takens bifurcation point of codimension three, where x_1 and x_2 coincide at the infection point of the prey isocline $y = \hat{F}(x)$.

Using the standard Hopf bifurcation theorem in [2] and [14], we can see that system (3.2) undergoes a *supercritical Hopf bifurcation of codimension one* and a *subcritical Hopf bifurcation of codimension one* in succession when $d = d_-$ and as (b,K) moves from below to above of the curve DH. Therefore, exactly two limit cycles may appear.

To discuss the exact number of limit cycles of system (3.2) via Hopf bifurcations, by restricting our analysis to the case $b = 0$, we can prove that system (3.2) has no limit cycles if system (3.2) has a multiple focus of multiplicity two. More precisely, we have the following theorem.

THEOREM 3.4. *Consider system (3.2) with $b = 0$. Assume that $0 < x_1 < K \le K_0$ and $\sqrt{\frac{3}{18+2\sqrt{6}}} \le d < \frac{\sqrt{3}}{4}$; then the equilibrium (x_1,y_1) of system (3.2) is globally stable in the interior of R^2_+. (Note that (x_1,y_1) of system (3.2) is a multiple focus of multiplicity two when $b = 0$, $K = K_0$, and $d = \sqrt{\frac{3}{18+2\sqrt{6}}}$.)*

Proof. It is clear that system (3.2) has a unique positive equilibrium (x_1,y_1) if $b = 0$, $0 < x_1 < K \le K_0$, and $\sqrt{\frac{3}{18+2\sqrt{6}}} \le d < \frac{\sqrt{3}}{4}$. Since the solutions of (3.2) in the interior of R^2_+ are positive and eventually bounded (i.e., there exists a positive number T such that $0 < x(t) < K$ for $t > T$), we claim that the unique positive equilibrium (x_1,y_1) of system (3.2) is globally stable in the interior of R^2_+. We prove the result by showing that system (3.2) has no closed orbits in the domain Ω_2, where

$$\Omega_2 = \{(x,y) :\ 0 < x < K,\ \ 0 < y < +\infty\}.$$

Taking $h(v) = y_1(e^v - 1)$, we transform system (3.2) in Ω_2 to the following Liénard-type system by transformations (2.7) and (2.9) when $b = 0$, $0 < x_1 < K \le K_0$, and $\sqrt{\frac{3}{18+2\sqrt{6}}} \le d < \frac{\sqrt{3}}{4}$ (we still denote u, v, and τ by x, y, and t, respectively):

$$\begin{aligned} \frac{dx}{dt} &= \phi_1(y) - F_1(x), \\ \frac{dy}{dt} &= -g_1(x) \end{aligned} \tag{3.22}$$

in the domain Ω_3, where $\phi_1(y) = y_1(e^y - 1)$, $F_1(x) = \frac{rx}{K}(x^2 + (K - 3x_1)x + x_1(x_2 + 3x_1 - K))$, $g_1(x) = \frac{dx(x-x_1+x_2)}{x_1-x}$, and

$$\Omega_3 = \{(x,y) :\ x_1 - K < x < x_1,\ \ -\infty < y < +\infty\}.$$

Note that the problem of existence of closed orbits of system (3.2) in the domain Ω_2 is equivalent to that of system (3.22) in the domain Ω_3.

By the Filippov transformation, we can see that system (3.22) has no closed orbits if $\frac{f_1(x)}{g_1(x)} \le \frac{f_1(u)}{g_1(u)}$ holds for any (u,x) satisfying $G_1(x) = G_1(u)$ with $x_1 - K < u < 0$ and $0 < x < x_1$ (cf. Theorem 2.5 in [23]), where $G_1(x) = \int_0^x g_1(s)ds$ and

$$f_1(x) \overset{\text{def}}{=} F_1'(x) = \frac{r}{k}\big(3x^2 + 2(K - 3x_1)x + x_1(x_2 + 3x_1 - K)\big).$$

Next we claim that $u + x < 0$ if $G_1(u) = G_1(x)$ with $x_1 - K < u < 0$ and $0 < x < x_1$. In fact, if

$$G_1(u) = G_1(x)$$

as $x_1 - K < u < 0$ and $0 < x < x_1$, then

$$\frac{1}{2}(x-u)(x+u) + x_2(x-u) - x_1x_2 \ln\frac{x_1-u}{x_1-x} = 0.$$

Note that $2\frac{x-y}{x+y} \le \ln\frac{x}{y}$ as $0 < y \le x$, and the equality holds if and only if $x = y$. Hence, we have

$$(x+u)\left(x_1 - x_2 - \frac{1}{2}(x+u)\right) > 0,$$

which implies that $2(x_1 - x_2) < u + x < 0$.

As $G_1(u) = G_1(x)$ with $x_1 - K < u < 0$ and $0 < x < x_1$, we consider

$$\begin{aligned}
\frac{f_1(x)}{g_1(x)} - \frac{f_1(u)}{g_1(u)} &= \frac{r}{Kdxu(x - x_1 + x_2)(u - x_1 + x_2)}\Big([3x^2 + 2(K - 3x_1)x \\
&\quad + x_1(x_2 + 3x_1 - 2K)](x_1 - x)u(u - x_1 + x_2) - [3u^2 + 2(K - 3x_1)u \\
&\quad + x_1(x_2 + 3x_1 - 2K)](x_1 - u)x(x - x_1 + x_2)\Big) \\
&= \frac{r(x-u)}{Kdxu(x - x_1 + x_2)(u - x_1 + x_2)}\Big(- 3u^2x^2 - 3(x_2 - x_1)xu(x+u) \\
&\quad - (2Kx_2 + 2Kx_1 - 10x_1x_2)xu + (2K - x_2 - 3x_1)x_1^2(x+u) \\
&\quad + x_1^2(2K - x_2 - 3x_1)(x_2 - x_1)\Big) \\
&= \frac{r(x-u)}{Kdxu(x - x_1 + x_2)(u - x_1 + x_2)}\Big(- 3u^2x^2 - 3(x_2 - x_1)xu(x+u) \\
&\quad + (10x_1x_2 - 2Kx_2 - 2Kx_1)xu \\
&\quad - (x_2 + 3x_1 - 2K)x_1^2(x + u + x_2 - x_1)\Big).
\end{aligned}$$

Because $\sqrt{\frac{3}{18+2\sqrt{6}}} \le d < \frac{\sqrt{3}}{4}$ and $0 < x_1 < K \le K_0 = (3x_1 + x_2)/2$, $10x_1x_2 - 2Kx_2 - 2Kx_1 \ge 0$ and $x_2 + 3x_1 - 2K > 0$. On the other hand, $x + u + x_2 - x_1 > x_2 - K > 0$. Therefore,

$$\frac{f_1(x)}{g_1(x)} - \frac{f_1(u)}{g_1(u)} > 0.$$

TABLE 1
Parameter values for the simulation to verify the existence of two limit cycles.

	d	K	x_1	x_2
Multiple focus	$\sqrt{\frac{3}{18+2\sqrt{6}}}$	$K_0 = 1.809766597$	.4283729906	2.334414218
Perturbed	$\sqrt{\frac{3}{18+2\sqrt{6}}} + 0.01$	$K_0 = 1.809766597 + .01$	.4459113886	2.242598026

From above arguments and Theorem 2.5 in [23], we know that system (3.22) does not have any closed orbits in the range Ω_3, which implies that (x_1, y_1) is globally stable. □

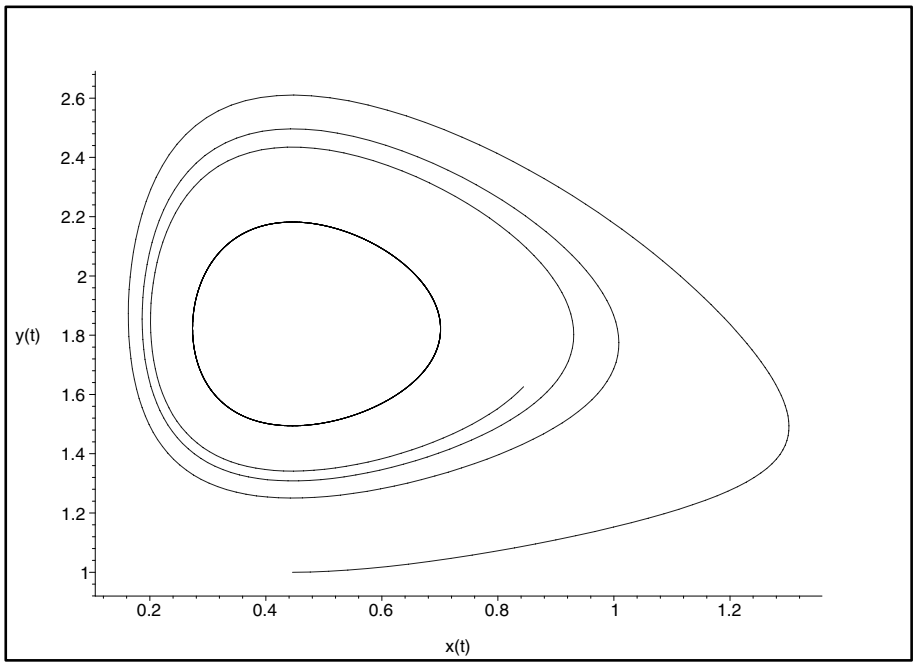
In Figure 3, a numerical simulation was carried out using MAPLE [20] to verify the existence of two limit cycles. For the simulation, we take $r = 2$ and $b = 0$; the other parameters are given in Table 1, where the system has a unique positive equilibrium which is a multiple focus of multiplicity two. We then perturb the parameters d and K; the new perturbed system has two limit cycles, as illustrated in Figure 3(a) and (b).

Remark 3.2. In [26], the saddle node bifurcation of limit cycles was studied, and the existence of two limit cycles was proved. In fact, it follows from the results in [26] and Theorems 3.2 and 3.4 that the bifurcation of the multiple focus of multiplicity two produces two limit cycles, and these two cycles can either disappear through the Hopf bifurcations or one Hopf bifurcation and one supercritical homoclinic bifurcation, or disappear through the saddle node bifurcation of limit cycles.

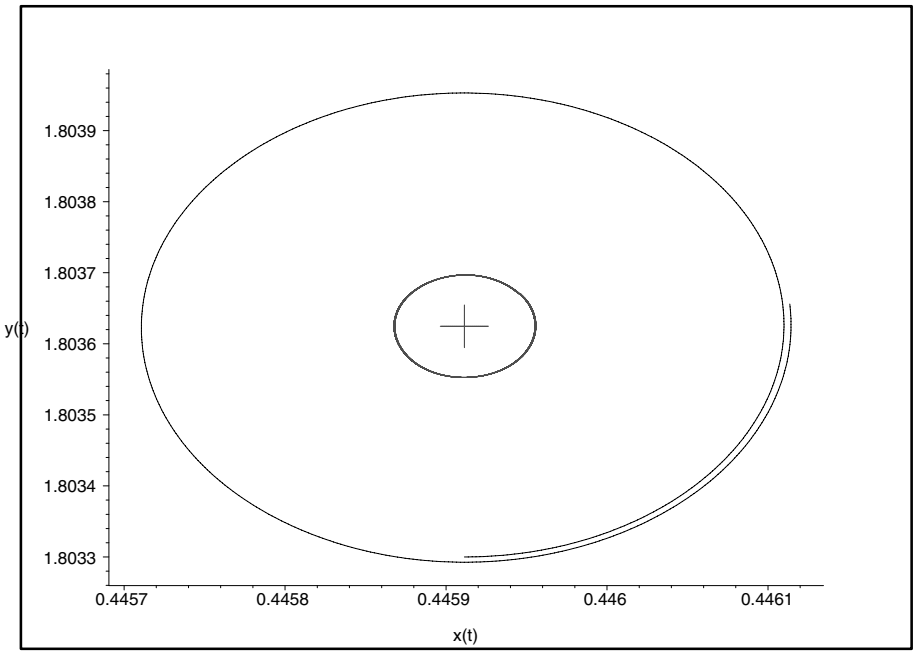
Remark 3.3. In Theorem 6.22 of [26], for parameters (b, K) in certain region near the curve segment DH, the bifurcation sequences were given but not complete. There we could not exclude the saddle node bifurcations of limit cycles and supercritical homoclinic bifurcations. By the above Theorem 3.4, we can now conclude that the sequences given in Theorem 6.22 are complete for (b, K) in the neighborhood of $b = 0$ and near the curve segment DH.

4. Discussion. The existence of limit cycles in predator-prey systems can be used to explain many real-world oscillatory phenomena, such as the Canadian lynx-snowshoe hare 10-year cycles. However, the classical Lotka–Volterra predator-prey system does not exhibit periodic fluctuations since the positive equilibrium is globally stable. One can perturb the classical Lotka–Volterra predator-prey system to obtain a unique nonconstant periodic orbit (see Freedman [5]). The second method for modifying the Lotka–Volterra predator-prey model so that it exhibits periodic solutions is to introduce a functional response function. Various forms of functional response functions have been proposed. In general, most of these functional response functions are monotone and have self-saturation effect. The dynamics of the generalized Gause-type predator-prey systems with general monotonic functional response have been extensively studied and very well understood. Multiple limit cycles were observed in predator-prey systems for cases when the predator death rate is both linear [26] and nonlinear (Hofbauer and So [8], Kuang [9], Wrzosek [22], etc.). However, the existence of exactly two limit cycles in predator-prey systems has not been studied yet.

Nonmonotonic functional response has been used to model the group defense phenomenon in population dynamics and inhibition in microbial dynamics (Freedman and Wolkowicz [6], Mischaikow and Wolkowicz [12], Rosenzweig [17], and Wolkowicz [21]), and very rich and interesting dynamics have been observed in Gause-type predator-



(a) The outer stable limit cycle



(b) The equilibrium (x_1, y_1) is stable, and the trajectory nearby spirals outwards, which indicates the existence of an unstable limit cycle

FIG. 3. *Two limit cycles bifurcated from the multiple focus of multiplicity two. The figures were produced using MAPLE* [20].

prey models with nonmonotonic functional response and only linear predator death rate ([18, 19, 26]). However, it is difficult to discuss the degenerate Hopf bifurcation for these models since it is hard to determine the multiplicity of a multiple focus by usual methods in [2, 14]. In this paper, we introduce a new method for solving the problem successfully. For the Gause-type predator-prey model with nonmonotonic functional response as in (1.2), we have shown that it has a multiple focus with multiplicity exactly two, and there exist some parameter values for which system (1.2) with the functional response (1.3) has exactly two limit cycles. These results also extend the bifurcation analysis of (3.2) by Ruan and Xiao [19] and Zhu, Campbell, and Wolkowicz [26].

It is natural that the predator-prey interaction has the tendency or potential to produce periodic oscillations. Hence one always needs to study the multiplicity of multiple focus and Hopf bifurcations in such predator-prey systems. The machinery developed in this paper can also be applied to the study of the multiplicity of multiple focus of other predator-prey systems with linear death rate for the predator, even with harvesting.

Acknowledgments. The authors are grateful to the anonymous referees and the handling editor for their valuable comments and suggestions. The authors also thank one referee for pointing us towards the reference [16].

REFERENCES

[1] F. Albrecht, H. Gatzke, and N. Wax, *Stable limit cycles in pre-predator populations*, Science, 181 (1973), pp. 1073–1074.
[2] A. Andronov, E. A. Leontovich, I. I. Gordon, and A. G. Maier, *Theory of Bifurcations of Dynamical Systems on a Plane*, Israel Program for Scientific Translations, Jerusalem, 1971.
[3] S. N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.
[4] A. D. Bazykin, *Nonlinear Dynamics of Interacting Populations*, World Sci. Ser. Nonlinear Sci. Ser. A Monogr. Treatises 11, World Scientific Publishing, River Edge, NJ, 1998.
[5] H. I. Freedman, *Deterministic Mathematical Models in Population Ecology*, Marcel Dekker, New York, 1980.
[6] H. I. Freedman and G. S. K. Wolkowicz, *Predator-prey systems with group defence: The paradox of enrichment revisited*, Bull. Math. Biol., 48 (1986), pp. 493–508.
[7] M. Han, *Liapunov constants and Hopf cyclicity of Lienard systems*, Ann. Differential Equations, 15 (1999), pp. 113–126.
[8] J. Hofbauer and. J. W.-H. So, *Multiple limit cycles for predator-prey models*, Math. Biosci., 99 (1990), pp. 71–75.
[9] Y. Kuang, *Nonuniqueness of limit cycles of Gause-type predator-prey systems*, Appl. Anal., 29 (1988), pp. 269–287.
[10] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Appl. Math. Sci. 112, Springer-Verlag, New York, 1995.
[11] R. M. May, *Limit cycles in predator-prey communities*, Science, 177 (1972), pp. 900–902.
[12] K. Mischaikow and G. S. K. Wolkowicz, *A predator-prey system involving group defense: A connection matrix approach*, Nonlinear Anal., 14 (1990), pp. 955–969.
[13] N. Kazarinoff and P. van der Driessche, *A model of predator-prey system with functional response*, Math. Biosci., 39 (1978), pp. 124–134.
[14] L. Perko, *Differential Equations and Dynamical Systems*, Springer-Verlag, New York, 1996.
[15] S. S. Pilyugin and P. Waltman, *Divergence criterion for generic planar systems,* SIAM J. Appl. Math., 64 (2003), pp. 81–93.
[16] S. Pilyugin and P. Waltman, *Multiple limit cycles in the chemostat with variable yield*, Math. Biosci., 182 (2003), pp. 151–166.
[17] M. L. Rosenzweig, *Paradox of enrichment: Destabilization of exploitation ecosystems in ecological time*, Science, 171 (1971), pp. 385–387.
[18] F. Rothe and D. S. Shafer, *Multiple bifurcation in a predator-prey system with nonmonotonic predator response*, Proc. Roy. Soc. Edinburgh Sect. A, 120 (1992), pp. 313–347.

[19] S. Ruan and D. Xiao, *Global analysis in a predator-prey system with nonmonotonic functional response*, SIAM J. Appl. Math., 61 (2001), pp. 1445–1472.
[20] Waterloo Maple Software, *MAPLE* VII, University of Waterloo, Waterloo, Canada, 2001.
[21] G. S. K. Wolkowicz, *Bifurcation analysis of a predator-prey system involving group defence*, SIAM J. Appl. Math., 48 (1988), pp. 592–606.
[22] D. M. Wrzosek, *Limit cycles in predator-prey models*, Math. Biosci., 98 (1990), pp. 1–12.
[23] D. Xiao and Z. Zhang, *On the uniqueness and nonexistence of limit cycles for predator-prey systems*, Nonlinearity, 16 (2003), pp. 1185–1201.
[24] Z. Zhang, *On the existence of exactly two limit cycles for the Liénard equation*, Acta Math. Sinica, 24 (1981), pp. 710–716 (in Chinese).
[25] Y. R. Zhou, *Existence and uniqueness of a limit cycle for the system of equations* $\dot{x} = \varphi(y) - F(x)$, $\dot{y} = -g(x)$, *and the existence of exactly two limit cycles*, Chinese Ann. Math., 3 (1982), pp. 89–102 (in Chinese).
[26] H. Zhu, S. A. Campbell, and G. S. K. Wolkowicz, *Bifurcation analysis of a predator-prey system with nonmonotonic function response*, SIAM J. Appl. Math., 63 (2002), pp. 636–682.

SIAM J. APPL. MATH.
Vol. 66, No. 3, pp. 820–842

EFFECTIVE MASS THEOREMS FOR NONLINEAR SCHRÖDINGER EQUATIONS*

CHRISTOF SPARBER†

Abstract. We consider time-dependent nonlinear Schrödinger equations subject to smooth lattice-periodic potentials plus additional confining potentials, slowly varying on the lattice scale. After an appropriate scaling we study the homogenization limit for vanishing lattice spacing. Assuming well-prepared initial data, the resulting effective dynamics is governed by a homogenized nonlinear Schrödinger equation with an effective mass tensor depending on the initially chosen Bloch eigenvalue. The given results rigorously justify the use of the effective mass formalism for the description of Bose–Einstein condensates on optical lattices.

Key words. nonlinear Schrödinger equation, effective mass theorems, Bloch waves, homogenization limit, Bose–Einstein condensate

AMS subject classifications. 35Q55, 35B27, 35B25, 74Q10

DOI. 10.1137/050623759

1. Introduction. Recent experiments on *Bose–Einstein condensates* (BECs) study the influence of *optical lattices* (or superlattices) on the dynamics of the condensate; cf. [12, 14, 28, 32]. The theoretical description of such systems is usually based on the famous *Gross–Pitaevskii equation*, i.e.,

$$i\hbar\partial_t\psi = -\frac{\hbar^2}{2m}\Delta\psi + V(x)\psi + U_0(x)\psi + N\alpha_0|\psi|^2\psi, \quad x \in \mathbb{R}^3,\, t \in \mathbb{R}, \tag{1.1}$$

where m is the atomic mass, $\hbar$ is the Planck constant, N is the number of atoms in the condensate, and

$$\alpha_0 = \frac{4\pi\hbar^2 a}{m}, \tag{1.2}$$

with $a \in \mathbb{R}$ being the s-wave scattering length derived from the corresponding N-particle theory; cf. [36]. Depending on the sign of a, the condensate is said to be either *repulsive* (stable) or *attractive* (unstable). In the nonlinear Schrödinger equation (NLS) (1.1), the potential $U_0(x)$ models some given external *confinement*, whereas $V(x)$ represents the *lattice-potential*, satisfying

$$V(x+\gamma) = V(x) \quad \forall x \in \mathbb{R}^3,\ \gamma \in \underline{\Gamma}, \tag{1.3}$$

where $\underline{\Gamma} \simeq \mathbb{Z}^3$ denotes some given *regular lattice*, generated through a basis $\{\underline{\zeta}_1, \underline{\zeta}_2, \underline{\zeta}_3\}$, $\underline{\zeta}_l \in \mathbb{R}^3$, i.e.,

$$\underline{\Gamma} = \left\{ \gamma \in \mathbb{R}^3 :\ \gamma = \sum_{l=1}^{3} \gamma_l \underline{\zeta}_l,\ \gamma_l \in \mathbb{Z} \right\}. \tag{1.4}$$

*Received by the editors February 2, 2005; accepted for publication (in revised form) October 10, 2005; published electronically February 15, 2006. This work has been partially supported by the APART grant of the author (funded by the Austrian Academy of Science), the Wittgenstein Award 2000 of Peter Markowich (funded by the Austrian research fund FWF), and the Warwick EPSRC symposium on the "Mathematics of quantum systems."

http://www.siam.org/journals/siap/66-3/62375.html

†Wolfgang Pauli Institute & Faculty of Mathematics, Vienna University, Nordbergstraße 15, A-1090 Vienna, Austria (christof.sparber@univie.ac.at).

Of course the nonlinear dynamics described by (1.1) can be highly involved. In the physics literature it is therefore frequently proposed (cf. [27, 38, 41]) that one consider the following simplifications: First it is assumed that the matter wave-field $\psi(t)$ can be characterized by a (fixed) central wave vector $k_0 \in \mathbb{R}^3$, and second one tries to capture the rapid oscillations in the wave function $\psi(t)$ by performing an asymptotic expansion in terms of *Bloch waves* $\chi_n(y, k_0)$ (see (2.4) below for their precise definition). The center of mass of the wave function is then described by a slowly varying envelope function $f(t, x)$, the dynamics of which is *formally* found to be governed by an *effective-mass NLS*. These types of approximations are well known in solid-state physics [8], though mostly in a time-independent setting [35], where one considers the motion of electrons in a crystal. It is the purpose of this work to *rigorously justify* the described approach specifically within the considered nonlinear context. To this end we shall consider a more general NLS than that originally proposed in (1.1), namely,

$$\begin{cases} i\hbar\partial_t\psi = -\dfrac{\hbar^2}{2m}\Delta\psi + V(x)\psi + U_0(t,x)\psi + \alpha|\psi|^{2\sigma}\psi, \quad x \in \mathbb{R}^d,\ t \in \mathbb{R}, \\ \psi\big|_{t=0} = \psi_I(x), \end{cases} \tag{1.5}$$

where $\alpha \in \mathbb{R}$ and $\sigma \in \mathbb{N}$. To motivate the choice $\sigma \geq 1$, we note that for $d < 3$ higher order nonlinearities are frequently used in the description of BECs; cf. [26, 29]. Moreover, different NLS-type models of the form (1.5) also appear in nonlinear optics and laser physics; cf. [42] (see also [24] for a rigorous derivation). We assume $\psi_I \in L^2(\mathbb{R}^d)$ to be normalized such that

$$\int_{\mathbb{R}^d} |\psi_I(x)|^2 dx = 1. \tag{1.6}$$

This normalization condition is henceforth conserved during the time-evolution. Again V is assumed to be periodic w.r.t. some regular lattice $\underline{\Gamma} \simeq \mathbb{Z}^d$, and U_0 denotes some, in general time-dependent, smooth external potential. Now, we rescale (1.5) in order to precisely identify the asymptotic regime we shall be dealing with in what follows. We have in mind a situation where the potential U_0 is slowly varying on the lattice-scale corresponding to V. Hence, there are essentially two scales in this problem. First, the *macroscopic length-* and *time-scales*, denoted L and T, respectively, which are introduced via U_0 by rewriting it in the following dimensionless form:

$$U_0(t,x) = \frac{mL^2}{T^2}\, U\left(\frac{t}{T}, \frac{x}{L}\right). \tag{1.7}$$

In other words, the scaled potential U is such that a free particle of mass m under the influence of U_0 will travel the distance L in the time-unit T. On the other hand, we can also introduce a couple of *microscopic scales*, λ and τ, via a rescaling of the periodic potential V such that

$$V(x) = \frac{m\lambda^2}{\tau^2}\, V_\Gamma\left(\frac{x}{\lambda}\right). \tag{1.8}$$

The *rescaled lattice* Γ is henceforth generated through a basis $\{\zeta_l\}_{l=1}^3$, where $\zeta_l = \underline{\zeta}_l/\lambda$, and the microscopic time-unit τ is then given by

$$\tau = \frac{m\lambda^2}{\hbar}. \tag{1.9}$$

We consequently define two *small dimensionless parameters*, ε and δ, as being the length- and time-ratios, respectively, i.e.,

$$\varepsilon = \frac{\lambda}{L}, \quad \delta = \frac{\tau}{T}. \tag{1.10}$$

In the following, both of these are assumed to be *small*, i.e., $\varepsilon \ll 1$, $\delta \ll 1$, but in general *not necessarily equal*. Next we introduce new space- and time-variables $\tilde{x}$ and $\tilde{t}$ via

$$\tilde{x} = \frac{x}{L}, \quad \tilde{t} = \frac{t}{T}, \tag{1.11}$$

and rescale the NLS (1.5) in dimensionless form. Having in mind the normalization condition (1.6), we also need to rescale the wave function ψ by

$$\tilde{\psi}(\tilde{t}, \tilde{x}) = L^{d/2}\psi(t, x). \tag{1.12}$$

After multiplying (1.5) by $T^2/(mL^2)$, we consequently arrive at the following dimensionless two-parameter model (where we again omit all "˜" for simplicity):

$$\begin{cases} ih\partial_t\psi = -\dfrac{h^2}{2}\Delta\psi + \dfrac{h^2}{\varepsilon^2}V_\Gamma\left(\dfrac{x}{\varepsilon}\right)\psi + U(t,x)\psi + \kappa\,|\psi|^{2\sigma}\psi, \\ \psi\big|_{t=0} = \psi_I(x). \end{cases} \tag{1.13}$$

Here we introduced two additional dimensionless parameters,

$$h := \frac{\hbar T}{mL^2}, \quad \kappa := \frac{\alpha T^2}{mL^{d\sigma+2}}, \tag{1.14}$$

the former of which can be considered to be Planck's constant in the macroscopic variables. Note that the following important relation holds,

$$\varepsilon^2 = h\delta, \tag{1.15}$$

connecting the ratio of the length-scales ε with the corresponding time-scale ratio δ. Finally, since we are aiming for nonlinearities of order $\mathcal{O}(1)$, we shall impose from now on that

$$|\kappa| \equiv \frac{|\alpha|T^2}{mL^{d\sigma+2}} = 1, \quad \text{or, equivalently,} \quad T = \sqrt{\frac{mL^{d\sigma+2}}{|\alpha|}}, \tag{1.16}$$

hence relating the macroscopic length- and time-scales in a specific way. We remark that in the linear case a scaling analogous to (1.13) has been introduced in [37]. A brief discussion on several aspects of this scaling procedure is now in order. First note that if we choose $h = \varepsilon$, hence, in view of (1.15), $\varepsilon = \delta$; i.e., if we choose the *same* ratio for both the length- and the time-scales, then (1.13) simplifies to a one-parameter model given by

$$i\varepsilon\partial_t\psi = -\frac{\varepsilon^2}{2}\Delta\psi + V_\Gamma\left(\frac{x}{\varepsilon}\right)\psi + U(t,x)\psi + \kappa\,|\psi|^{2\sigma}\psi. \tag{1.17}$$

This is nothing but the standard *semiclassical scaling* for (nonlinear) Schrödinger-type equations including an additional *highly oscillatory periodic potential* V_Γ. Recently the

rigorous study of the corresponding asymptotic regime $\varepsilon \to 0$, known as the *combined semiclassical and adiabatic approximation*, attracted lots of interest. Particularly in the linear setting, i.e., $\kappa = 0$, different mathematical approaches are currently at hand, e.g., *WKB-type expansions* [6, 21], *Wigner transformation techniques* [30, 18], and the so-called *space-adiabatic perturbation theory* [33, 44], which gives the most sophisticated mathematical results so far. Including nonlinear effects, the literature is not so abundant. To the author's knowledge the only result in this direction is a recent paper by Carles, Markowich, and the current author [11]. The results given there, however, are valid only for weak nonlinearities, in the sense that we need to assume $\kappa \sim \mathcal{O}(\varepsilon)$. Therefore a different rescaling of the original NLS (1.5) has been introduced in [11].

Remark 1.1. Additionally there exists a related work on the semiclassical limit of the so-called Schrödinger–Poisson system [5] in a crystal. There, however, additional assumptions have to be imposed which are out of the realm of the present setting (like truly mixed-state initial data).

In the following our focus is not on the semiclassical regime. Rather, we shall study the asymptotic behavior of the scaled NLS (1.13) for $\varepsilon \ll 1$ but with a fixed h of order $\mathcal{O}(1)$. Note that, by (1.15), this implies $\delta \sim \mathcal{O}(\varepsilon^2)$, and hence we are considering our system on a much larger macroscopic time-scale T than we do by fixing a macroscopic length-scale L via (1.10). In particular, we are dealing with much larger times T, as in the semiclassical studies described above. Roughly speaking, the semiclassical regime can be seen to be an asymptotic description for *ballistic scales*, whereas we shall be dealing in the following with *dispersive scales* (sometimes also called *diffusive scales*). As we shall see, this indeed turns out to be the asymptotic regime where one can rigorously justify the effective-mass formalism discussed at the beginning. We note, however that, in contrast to what is noted in [3], the considered regime is *not* equivalent to the one obtained after rescaling time in the semiclassical equation (1.17) by $t \to \varepsilon t$. The reason for this is the different orders of magnitude in the external potential U and in the nonlinearity. To have a more concrete feeling of the involved time- and length-scales we come back to our original equation (1.1). Thus we consider (1.13) in $d = 3$, with $\sigma = 1$ and therefore $\kappa = 4\pi\hbar^2 a T^2/(mL^5)$ by (1.14). A particular example for the periodic potentials used in physical experiments is then given by [14, 36]

$$V(x) = \sum_{l=1}^{3} \frac{\hbar^2 \xi_l^2}{m} \sin^2(\xi_l x_l), \quad \xi_l \in \mathbb{R}, \tag{1.18}$$

where $\xi = (\xi_1, \xi_2, \xi_3)$ denotes the wave vector of the laser field which generates the optical lattice. Hence we readily identify $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ as $\lambda_l = 1/\xi_l$. Moreover, the slowly varying external potential U_0 is usually modeled to be static and of harmonic oscillator type (isotropic or anisotropic), i.e.,

$$U_0(x) = \frac{m\omega_0^2}{2} |x|^2, \quad \omega_0 \in \mathbb{R},\ x \in \mathbb{R}^3. \tag{1.19}$$

In this case, a natural choice for the macroscopic length-scale is therefore given by $L = a_0$, where a_0 denotes the length of the harmonic oscillator ground state corresponding to $U_0(x)$, i.e.,

$$a_0 := \sqrt{\frac{\hbar}{\omega_0 m}}. \tag{1.20}$$

The assumption $\varepsilon \ll 1$ is then of course equivalent to $\lambda \ll a_0$. In an actual physical experiment this requirement can be easily satisfied, as a typical ground state length would be $a_0 \approx 10^{-6}[m]$, whereas the wave vectors ξ_l of the laser fields are usually tuned from $10^6[1/m]$ to $10^9[1/m]$, the latter case therefore being suitable in our situation. The corresponding relation for the time-scales, however, is more subtle. From the condition (1.16) we see that T has to be chosen such

$$T^2 = \frac{a_0^5 m^2}{4\pi N|a|\hbar^2} \gg \tau^2, \qquad \text{since } \delta \ll 1.$$

With τ given by (1.9), this finally leads to the requirement

$$\frac{a_0^5}{4\pi N|a|} \gg \lambda^4.$$

In particular, in the so-called *moderate interaction regime*, characterized by the fact that $4\pi N|a| \approx a_0$ [4], this is again equivalent to $a_0 \gg \lambda$, and in this case we compute

$$T^2 = \frac{a_0}{4\pi N|a|\omega_0^2} \approx \frac{1}{\omega_0^2}.$$

Note that this is precisely what one would get in the corresponding linear situation. From a mathematical point of view the limit $\varepsilon \to 0$ with δ fixed corresponds to the so-called *homogenization limit* of (1.13). In view of the classical homogenization results as described in, e.g., [6, 23], the main new difficulty, apart from the appearing nonlinearity, stems from the *large factor* $1/\varepsilon^2$ in front of the periodic potential, which furnishes a highly *singularly perturbed* term. It is therefore not a surprise that, even in the linear case, this type of homogenization problem has been rigorously studied only very recently [2]. In particular, (linear) time-dependent Schrödinger-type equations have been considered in [3] and in [37]. The latter result relies on the use of *Wigner measures*, a technique which cannot be applied in the given nonlinear situation. The former work is more closely related to ours, as it combines classical homogenization techniques, most notably *two-scale convergence methods* [1], with Bloch wave decomposition [13]. However, we want to stress the fact that the nonlinear case we shall be dealing with is by no means a straightforward generalization of the linear results. More precisely, one should note that the scaling of (1.13) in general prohibits the derivation of suitable, i.e., *uniformly* in ε, a priori estimates, except for the basic L^2 estimate corresponding to the conservation of mass. In the majority of cases, the derivation of such uniform estimates is crucial to gaining sufficient control on the limiting behavior of the appearing nonlinearities, a problem which cannot be handled by using weak-convergence methods (such as Wigner measures or two-scale convergence).

Remark 1.2. Also, for the same reasons, our results do not fit into the framework of *H-measures* [43] or *G-convergence* [34].

We note that in [2, 3] the authors propose the use of a *factorization principle* in order to extend their results to the nonlinear case. However, this approach remains unproven there and, moreover, is known to be applicable only in situations where the initial data ψ_I is *concentrated at the minimum of the first Bloch band* (see [2, 3] for more details on this). In comparison to that, the results given below are indeed *independent* of the number of the Bloch band, and also they do *not* require that ψ_I be concentrated at a local minimum of the considered band. On the other hand, we do need the initial data ψ_I to be *well prepared* in a sense to be made more precise

below; cf. Assumption 4.3. Additionally we need to assume sufficient regularity on the potentials U, V as well as on the initial data ψ_I. The reasons for these assumptions are first the fact that we shall use a more traditional *multiple-scales expansion method*, similar to those introduced in [6]. This approach will allow us to obtain, in a rather transparent way, an asymptotic description of $\psi(t)$ for small $\varepsilon \ll 1$, and also to determine the corresponding *effective homogenized NLS*. Second, in order to prove that the given asymptotic solution is indeed *stable* under the nonlinear time-evolution governed by (1.13), we shall adapt an approach originally introduced to prove the accuracy of *nonlinear geometrical optics expansions* (cf. [15, 20, 39]) for the results most closely related to the present work. The given proof will then again rely heavily on the fact that we have sufficient regularity properties and well-prepared initial data. A similar strategy recently proved to be successful when applied to the weakly nonlinear semiclassical situation studied in [11]. The main goal of this paper, though, is not the introduction of new methods but rather a complete and correct treatment of the problem at hand. Moreover, one should keep in mind that for completely arbitrary initial data $\psi_I \in L^2(\mathbb{R}^d)$ one cannot expect an effective mass-type dynamics to be valid. In other words, to obtain an equation of the form (1.22) the initial data ψ_I needs to be (asymptotically) of the same type as stated below, at least in leading order, and the additional well-preparedness assumptions we shall need concern only the higher order terms within the asymptotic expansion. We note that only in linear cases does the superposition principle allow for more general states ψ_I.

At the expense of not completely well-defined assumptions (to be made precise later on), let us now state the typical *effective mass theorem* we shall prove in what follows.

THEOREM 1.3. *Let V_Γ and U be smooth real-valued potentials such that V_Γ is Γ-periodic and U is subquadratic. Assume that for some $k_0 \in \mathbb{R}$ the initial data ψ_I is of the following form:*

$$\psi_I(x) = f_I(x)\chi_n\left(\frac{x}{\varepsilon}, k_0\right) e^{ik_0\cdot x/\varepsilon} + \varepsilon\eta^\varepsilon(x),$$

where $f_I \in \mathcal{S}(\mathbb{R}^d;\mathbb{C})$; $\chi_n(y,k)$ is an eigenfunction of Bloch's spectral cell problem, corresponding to a simple eigenvalue $E_n(k)$; and the corrector $\eta^\varepsilon \sim \mathcal{O}(1)$ is such that ψ_I is well prepared, up to sufficiently large $K \in \mathbb{N}$, in the sense of Assumption 4.3 below. Then there exists an $\varepsilon_0 > 0$ such that for $\varepsilon \le \varepsilon_0$ the following asymptotic estimate holds:

$$\sup_{t\in[-\tau_0,\tau_0]} \|\psi(t) - v_0^\varepsilon(t)\|_{L^2(\mathbb{R}^d)} = \mathcal{O}(\varepsilon), \quad \tau_0 < \tau, \tag{1.21}$$

where $\tau > 0$ is the maximum existence-time for a smooth solution $f(t,x)$ for the homogenized NLS

$$ih\partial_t f = -\frac{h^2}{2}\,\mathrm{div}(M^*\nabla)f + U(x)\psi + \kappa^*\,|f|^{2\sigma} f, \tag{1.22}$$

with effective mass tensor $M^* = D^2E_n(k_0)$ and an effective coupling $\kappa^* \in \mathbb{R}$, given by (3.21) below. Moreover, the leading order approximate solution v_0^ε is found to be

$$v_0^\varepsilon(t,x) = f\left(t, x - \frac{ht}{\varepsilon}\nabla_k E_n(k_0)\right)\chi_n\left(\frac{x}{\varepsilon}, k_0\right) e^{ik_0\cdot x/\varepsilon} e^{-ihE_n(k_0)t/\varepsilon^2}. \tag{1.23}$$

We note that the estimate (1.21) implies a strong two-scale convergence statement as defined in [1, 2, 3]. Also note that, apart from the nonlinearity, our approach represents a refinement of the classical multiple-scales expansions given in [6], in the sense that we can include possibly appearing *large drifts* (clearly visible in (1.23) in the second argument of f) in order to resolve the underlying dispersive behavior. The possibility of large drifts present in the asymptotic solution can be seen as the aftermath of the ballistic regime known from the semiclassical situation, a fact which has already been noticed in linear situations [3, 16].

The paper is now organized as follows. In section 2 we collect some preliminary results and important notation used throughout this work. We then proceed along the lines of [11] and next present in section 3 the multiple-scales expansion method, whereas its nonlinear stability shall be proved in section 4.

2. Preliminaries. For simplicity we restrict ourselves in this work to *static external potentials* $U = U(x)$, although all results could be generalized to the case of time-dependent potentials $U(t,x)$, which are smooth w.r.t. $t \in \mathbb{R}$. (Indeed, we could as well include smoothly time-dependent coupling factors $\kappa(t) \in \mathbb{R}$.) Thus we study in the following the asymptotic behavior as $\varepsilon \to 0$ of

$$\begin{cases} ih\partial_t\psi = -\dfrac{h^2}{2}\,\Delta\psi + \dfrac{h^2}{\varepsilon^2}V_\Gamma\left(\dfrac{x}{\varepsilon}\right)\psi + U(x)\psi + \kappa\,|\psi|^{2\sigma}\psi, \\ \psi\big|_{t=0} = \psi_I(x). \end{cases} \tag{2.1}$$

All results given below are then valid for potentials which satisfy the following basic assumption.

Assumption 2.1. The potentials U and V_Γ are such that $V_\Gamma, U \in C^\infty(\mathbb{R}^d;\mathbb{R})$, and moreover they satisfy the following:

(i) V_Γ is Γ-periodic: $V_\Gamma(x+\gamma) = V_\Gamma(x)\ \forall x \in \mathbb{R}^d$, $\gamma \in \Gamma \simeq \mathbb{Z}^d$.

(ii) U is subquadratic: $\partial^\alpha U \in L^\infty(\mathbb{R}^d)\ \forall \alpha \in \mathbb{N}^d$ such that $|\alpha| \geq 2$.

Remark 2.2. These are the same assumptions that were used in [11]. In particular, they include the cases of isotropic harmonic potentials $U(x) = |x|^2$, as well as those corresponding to anisotropic ones, like $U(x) = \sum \omega_j^2 x_j^2$. Moreover, we can take U to be identically zero, or include a linear component such as $E \cdot x$, $E \in \mathbb{R}$, modeling constant electric fields, for example.

We proceed by recalling Bloch's famous eigenvalue problem [7].

2.1. Bloch's eigenvalue problem. From now on we denote by $\mathfrak{C}$ the *elementary lattice cell*, i.e., the centered *fundamental domain* of the lattice Γ, i.e.,

$$\mathfrak{C} := \left\{ \gamma \in \mathbb{R}^d :\ \gamma = \sum_{l=1}^{d} \gamma_l \zeta_l,\ \gamma_l \in \left[-\frac{1}{2}, \frac{1}{2}\right] \right\}, \tag{2.2}$$

whereas the corresponding basic cell of the dual lattice will be denoted by $\mathfrak{C}^*$. In solid-state physics $\mathfrak{C}^*$ is usually called the (first) *Brillouin zone*; hence we shall also write $\mathfrak{B} \equiv \mathfrak{C}^*$. Also let us introduce the so-called *Bloch Hamiltonian* (or shifted Hamiltonian) given by

$$H_\Gamma(k) := \frac{1}{2}\,(-i\nabla_y + k)^2 + V_\Gamma(y), \quad k \in \mathbb{R}^d. \tag{2.3}$$

Then Bloch's eigenvalue problem is given by the following spectral cell equation:

$$\begin{cases} H_\Gamma(k)\chi_n(y,k) = E_n(k)\chi_n(y,k) & n \in \mathbb{N},\ y \in \mathfrak{C}, \\ \chi_n(y+\gamma,k) = \chi_n(y,k) & \text{for } \gamma \in \Gamma, \end{cases} \tag{2.4}$$

and $E_n(k) \in \mathbb{R}$, $k \in \mathfrak{B}$, is then called the nth *Bloch eigenvalue* corresponding to the potential V_Γ. We shall now briefly collect some well-known facts for this eigenvalue problem (cf. [31, 40, 45]). Since V_Γ is smooth and periodic, $H_\Gamma(k)$, for every fixed $k \in \mathfrak{B}$, is self-adjoint on $H^2(\mathbb{T}^d)$, $\mathbb{T}^d = \mathbb{R}^d/\Gamma$, with compact resolvent. Hence its spectrum is given by

$$\sigma(H_\Gamma(k)) = \{E_n(k) \in \mathbb{R} \ ; \ k \in \mathfrak{B}, \, n \in \mathbb{N}^*\}.$$

The eigenvalues $E_n(k)$ can then be ordered according to their magnitude and multiplicity, i.e.,

$$E_1(k) \leq \cdots \leq E_n(k) \leq E_{n+1}(k) \leq \cdots .$$

Moreover, every $E_n(k)$ is periodic w.r.t. Γ^* and $E_n(k) = E_n(-k)$ holds. The set

$$\{E_n(k) \in \mathbb{R} : E_n(k) \leq E_{n+1}(k), \, k \in \mathfrak{B}\}$$

is then usually named the nth *energy band* (or Bloch band). The associated eigenfunction, the so-called *Bloch waves*, $\chi_n(y,k)$ form (for every fixed $k \in \mathfrak{B}$) a complete orthonormal basis in $L^2(\mathfrak{C})$ and are smooth w.r.t. $y \in \mathfrak{C}$. For the following we choose the usual normalization

$$\langle \chi_n(\cdot,k), \chi_m(\cdot,k) \rangle_{L^2(\mathfrak{C})} \equiv \int_{\mathfrak{C}} \overline{\chi_n(y,k)} \chi_m(y,k) dy = \delta_{n,m}, \quad n, \, m \in \mathbb{N}. \tag{2.5}$$

Regularity of the χ_n w.r.t. their dependence on $k \in \mathfrak{B}$ is more subtle. It has been shown [31] that for any $n \in \mathbb{N}$ there exists a closed subset $\mathfrak{A} \subset \mathfrak{B}$ such that $E_n(k)$ is analytic. Similarly the eigenfunctions $\chi_n(\cdot,k)$ are found to be analytic and quasi-periodic in k for all $k \in \mathfrak{O} := \mathfrak{B} \backslash \mathfrak{A}$. Moreover, it holds that

$$E_{n-1}(k) < E_n(k) < E_{n+1}(k) \quad \forall \, k \in \mathfrak{O}. \tag{2.6}$$

If this condition is satisfied for all $k \in \mathfrak{B}$, then $E_n(k)$ is said to be an *isolated Bloch band* [44]. Finally we remark that the set where one encounters the so-called *band crossings* is indeed of measure zero, i.e.,

$$\operatorname{meas} \mathfrak{A} = \operatorname{meas} \{k \in \mathfrak{B} \mid E_n(k) = E_m(k), \ n \neq m\} = 0.$$

Remark 2.3. Note that in the case $d = 1$ all band crossings can be removed through a proper analytic continuation of the bands; cf. [40].

From the eigenvalue equation (2.4) we obtain the following useful identities. Differentiating (2.4) w.r.t. k (assuming for the moment that everything is sufficiently smooth) yields

$$(\nabla_k H_\Gamma(k) - \nabla_k E_n(k))\chi_n + (H_\Gamma(k) - E_n(k))\nabla_k \chi_n = 0. \tag{2.7}$$

Hence, taking the scalar product with χ_n, we obtain a formula for $\nabla_k E_n(k)$ by

$$\begin{aligned} \langle \chi_n, \, \nabla_k H_\Gamma(k) \chi_n \rangle_{L^2(\mathfrak{C})} &\equiv \langle \chi_n, \, (-i\nabla_y + k)\chi_n \rangle_{L^2(\mathfrak{C})} \\ &= \nabla_k E_n(k), \end{aligned} \tag{2.8}$$

since H_Γ is self-adjoint. Similarly we obtain the following expression for the entries of the Hessian matrix $D^2 E_n(k)$:

$$\begin{aligned} \partial^2_{k_j k_l} E_n(k) = \delta_{j,l} &+ \langle \chi_n, \, (-i\partial_{y_j} + k_j)\chi_n \rangle + \langle \chi_n, \, (-i\partial_{y_l} + k_l)\partial_{k_j}\chi_n \rangle \\ &- \langle \chi_n, \, (\partial_{k_l} E_n(k))\partial_{k_j}\chi_n + (\partial_{k_j} E_n(k))\partial_{k_l}\chi_n \rangle, \end{aligned} \tag{2.9}$$

where $\delta_{j,l}$ denotes the Kronecker symbol for $j,l=1,\dots,d$. (Below we shall assume $E_n(k)$ to be a simple eigenvalue, which implies sufficient regularity to justify all differentiations above.)

2.2. Existence of smooth solutions for NLS. As a final preparatory step we state a basic existence and uniqueness result for NLS of the form (2.1). (See also [9, 42] for a general introduction.)

LEMMA 2.4. *Let Assumption* 2.1 *be satisfied, and let* $\psi_I \in \mathcal{S}(\mathbb{R}^d)$*, the Schwartz space. Let* $s > d/2$*. Then there exists* $t = t(\varepsilon,h) > 0$ *and a unique solution* $\psi \in C(]-t,t[\,;H^s(\mathbb{R}^d))$ *satisfying* (2.1)*. Moreover,* $x^\alpha\psi \in C(]-t,t[\,;H^s(\mathbb{R}^d))$ *for any* $\alpha\in\mathbb{N}^d$*,* $s\in\mathbb{N}$*, and the following conservation law holds:*

$$\frac{d}{dt}\,\|\psi(t)\|_{L^2}=0\,. \tag{2.10}$$

Proof. See the proof of Lemma 4.3 in [11]. ∎

Remark 2.5. In general, one cannot expect global-in-time existence for solutions to the NLS. For example, if κ is negative and if $\sigma > 2/d$, finite time blow-up may occur; see, e.g., [10, 42] (see also [9] for the case $\sigma = 2/d$).

3. Multiple-scales expansion. We now establish the asymptotic behavior of $\psi(t)$, the solution to (2.1), for $0<\varepsilon\ll 1$, by means of a multiple-scales expansion. In the following $h>0$ is kept fixed, though we shall not simply set it equal to 1, since we want to keep track of its appearance in order to compare our results with the semiclassical situation of [11]; cf. Remark 3.9 below. As in [3], we shall first consider the easier situation where no large drifts appear, and then include them in a second step, using a more general asymptotic expansion method.

3.1. Homogenization without drift. In this subsection we seek an asymptotic expansion for solutions to (1.13) in the following form:

$$\begin{cases} \psi(t,x) = u^\varepsilon\left(t,x,\dfrac{x}{\varepsilon}\right)\exp\left(i\left(\dfrac{k_0\cdot x}{\varepsilon}+\dfrac{\beta t}{\varepsilon^2}\right)\right), \\ u^\varepsilon(t,x,y)\sim \displaystyle\sum_{j=0}^{\infty}\varepsilon^j u_j(t,x,y), \end{cases} \tag{3.1}$$

where $k_0\in\mathbb{R}^d$ is induced by the initial condition and $\beta\in\mathbb{R}$ is some arbitrary constant to be determined below. The precise meaning of the symbol "$\sim$" in terms of an asymptotic series will be discussed in section 4 below. Moreover, we impose that $u^\varepsilon(t,x,y)\in\mathbb{C}$ satisfies

$$u^\varepsilon(\cdot,\cdot,y+\gamma)=u^\varepsilon(\cdot,\cdot,y)\quad\forall\, y\in\mathbb{R}^d,\,\gamma\in\Gamma,$$

in order to capture precisely those oscillations which are introduced via V_Γ.

Remark 3.1. This particular form of a multiple-scale ansatz is suggested by the linear results given in [2, 3]. Indeed, one could have started with a more general ansatz, imposing appropriate periodicity or quasi-periodicity assumptions. It then turns out that one ends up again with the same form as given in (3.1). Also, the ansatz (3.1) might not be so surprising when compared to the *two-scale WKB approach* used in [6, 11, 21] (see also Remark 3.9 below).

As usual in asymptotic expansion methods we have to henceforth assume that the initial condition ψ_I is compatible with (3.1).

Assumption 3.2 (well-prepared initial data I). The initial data ψ_I is assumed to be of the following form:

$$\psi_I(x) = u_I^\varepsilon\left(x, \frac{x}{\varepsilon}\right) e^{ik_0\cdot x/\varepsilon} \quad \text{for some given } k_0 \in \mathbb{R}, \tag{3.2}$$

where $u_I^\varepsilon(x,y)$ is Γ-periodic w.r.t. y and $u_I^\varepsilon \in \mathcal{S}(\mathbb{R}^{2d})$.[1]

Assuming for the moment that u^ε is sufficiently smooth, we (formally) plug the ansatz (3.1) into (2.1) and compare equal powers in ε. This yields

$$\frac{1}{\varepsilon^2}L_0u^\varepsilon + \frac{1}{\varepsilon}L_1u^\varepsilon + L_2u^\varepsilon + \kappa|u^\varepsilon|^{2\sigma}u^\varepsilon = 0, \tag{3.3}$$

where the linear differential operators L_0 and L_1 are defined by

$$\begin{aligned} L_0u^\varepsilon &:= h\beta + h^2H_\Gamma(k_0),\\ L_1u^\varepsilon &:= -h^2\nabla_x\cdot\nabla_y u^\varepsilon + ih^2k_0\cdot\nabla_x u^\varepsilon, \end{aligned} \tag{3.4}$$

with $H_\Gamma(k)$ being the Bloch Hamiltonian as given in (2.3). We also define

$$L_2u^\varepsilon := -ih\,\partial_t u^\varepsilon - \frac{h^2}{2}\,\Delta_x u^\varepsilon + U(x)u^\varepsilon. \tag{3.5}$$

Since $u^\varepsilon \sim \sum \varepsilon^j u_j$ we shall in the following expand (3.3) in powers of ε and derive conditions on u_j such that all resulting terms are zero up to sufficient high orders in ε. Setting the leading order term, i.e., the term of order $\mathcal{O}(\varepsilon^{-2})$, equal to zero gives

$$H_\Gamma(k_0)u_0 + \frac{\beta}{h}\,u_0 = 0, \tag{3.6}$$

from which we readily see that we need to choose

$$\beta = -hE_n(k_0). \tag{3.7}$$

Now, if we assume that $E_n(k_0)$ is indeed a *simple* Bloch-eigenvalue, then (3.6) implies that u_0 can be decomposed as

$$u_0(t,x,y) = f_0(t,x)\chi_n(y,k_0) \quad \forall\, t\in\mathbb{R},\ x\in\mathbb{R}^d, \tag{3.8}$$

with some yet undetermined function $f(t,x)\in\mathbb{C}$. This now leads directly to the following important assumption.

Assumption 3.3 (well-prepared initial data II). Initially, the leading order "amplitude" u_0 is assumed to be concentrated in a single Bloch band $E_n(k_0)$ corresponding to a simple eigenvalue of $H_\Gamma(k_0)$, i.e.,

$$u_0(0,x,y) \equiv f_0(0,x)\chi_n(y,k_0), \tag{3.9}$$

where $f_0(0,\cdot)\equiv f_I(\cdot)\in\mathcal{S}(\mathbb{R}^d;\mathbb{C})$ is some given initial data.

An important consequence of $E_n(k_0)$ being simple is that in this case it is known to be infinitely differentiable in a vicinity of k_0. We have seen that in leading order $\psi(t)$ can be written as

$$\psi(t,x) \sim f_0(t,x)\chi_n\left(\frac{x}{\varepsilon},k_0\right)e^{ik_0\cdot x/\varepsilon}e^{-ihE_n(k_0)t/\varepsilon^2} + \mathcal{O}(\varepsilon), \tag{3.10}$$

[1] That is, u_I^ε is smooth and rapidly decaying w.r.t. x and smooth w.r.t. y

where the f_0 is yet to be determined. To this end we proceed with our asymptotic expansion by setting terms of order $\mathcal{O}(\varepsilon^{-1})$ equal to zero. This yields

$$(H_\Gamma(k_0) - E_n(k_0))u_1 = \nabla_x \cdot \nabla_y u_0 - ik_0 \cdot \nabla_x u_0, \tag{3.11}$$

which, by inserting (3.8), can be rewritten as

$$\begin{aligned}(H_\Gamma(k_0) - E_n(k_0))u_1 =& -i\nabla_x f_0 \cdot (-i\nabla_y \chi_n + k_0 \chi_n) \\ =& -i\nabla_x f_0 \cdot \nabla_k H_\Gamma(k_0)\chi_n,\end{aligned} \tag{3.12}$$

where the second equality follows from the definition of $H_\Gamma(k)$ in (2.3). It remains to ask whether this equation is solvable for u_1. By Fredholm's alternative the necessary and sufficient condition for solving it is that the right-hand side be orthogonal (in $L^2(\mathfrak{C})$) to $\chi_n(y, k_0)$. Hence we have to impose that

$$\begin{aligned}0 =& -i\nabla_x f_0 \cdot \langle \chi_n, \nabla_k H_\Gamma(k_0)\chi_n \rangle_{L^2(\mathfrak{C})} \\ =& -i\nabla_x f_0 \cdot \nabla_k E_n(k_0),\end{aligned} \tag{3.13}$$

where for the second equality we used the identity (2.8). Thus we are led to the restriction that k_0 has to be a critical point of $E_n(k)$, i.e.,

$$\nabla_k E_n(k_0) = 0. \tag{3.14}$$

This situation is analogous to the one discussed in the first part of [3], though the arguments given there are different. Assuming that (3.14) indeed holds true, we get from (3.12), together with (2.7), that the order $\mathcal{O}(\varepsilon)$ *corrector* u_1 in general can be written in the following form:

$$u_1(t, x, y) = -i\nabla_x f_0(t, x) \cdot \nabla_k \chi_n(y, k_0) + f_1(t, x)\chi_n(y, k_0) \tag{3.15}$$

for any given function f_1. Note that we cannot choose $u_1(0, x, y)$ completely arbitrary, once $u_0(0, x, y)$ is fixed; i.e., the initial data u_I^ε given in Assumption 3.2 (formally) has to be of the form

$$u_I^\varepsilon(x, y) \sim (f_0(0, x) + \varepsilon f_1(0, x))\chi_n(y, k_0) - i\varepsilon \nabla_x f_0(0, x) \cdot \nabla_k \chi_n(y, k_0) + \mathcal{O}(\varepsilon^2) \tag{3.16}$$

in order to be consistent with our asymptotic description. In the following we shall choose $f_1(0, x) \equiv 0$ for simplicity. Proceeding with our ε-expansion of (3.3), we next consider terms of order $\mathcal{O}(1)$ to obtain the following equation:

$$L_0 u_2 + L_1 u_1 + L_2 u_0 + \kappa |u_0|^{2\sigma} u_0 = 0. \tag{3.17}$$

Again, by Fredholm's alternative, it can be solved for u_2 iff

$$\int_{\mathfrak{C}} \overline{\chi_n(y, k_0)} \left(L_1 u_1 + L_2 u_0 + \kappa |u_0|^{2\sigma} u_0 \right) dy = 0. \tag{3.18}$$

Plugging into this identity the precise forms of u_0 and u_1, respectively defined in (3.8) and (3.15), and using formula (2.9), given the fact that $\nabla_k E_n(k_0) = 0$, we obtain after some lengthy but straightforward calculations the following solvability condition:

$$\begin{cases} ih\partial_t f_0 = -\dfrac{h^2}{2} \operatorname{div}_x(M^* \nabla_x) f_0 + U(x) f_0 + \kappa^* |f_0|^{2\sigma} f_0, \\ f_0\big|_{t=0} = f_I(x). \end{cases} \tag{3.19}$$

This is nothing but the *homogenized NLS*, or the *effective mass NLS*, where the so-called *effective mass tensor* $M^* \in \mathbb{R}^{d\times d}$ is defined by

$$M^*_{j,l} := \partial^2_{k_j,k_l} E_n(k_0), \qquad j,l = 1,\dots,d. \tag{3.20}$$

Moreover, the *effective coupling constant* $\kappa^* \in \mathbb{R}$ within the nth Bloch band is defined by

$$\kappa^*(k_0) := \kappa \int_{\mathfrak{C}} |\chi_n(y,k_0)|^{2\sigma+2}\, dy. \tag{3.21}$$

The effective NLS (3.19) describes the dispersive dynamics, as $\varepsilon \to 0$, of (2.1) for long macroscopic time-scales. However, it should not be confused with the so-called effective Hamiltonian, as determined in [31, 44]. Note that in general M^* is *neither positive nor definite*, and thus (3.19) in general also includes the class of so-called *nonelliptic NLS* [17, 25, 42].

Remark 3.4. The formulas (3.19)–(3.21) can be shown to be exactly the same as in the physics literature [38, 41] and moreover simplify to the ones given in [3, 37] in the linear case. If M^* is scalar, its inverse $m^* = 1/M^*$ is called the *effective mass*.

In the next subsection we shall show how to get rid of the additional assumption (3.14) that k_0 is a critical point of $E_n(k)$.

3.2. General situation including drifts. In order to generalize the expansion to situations where $\nabla_k E_n(k) \neq 0$ we have to modify our multiple-scales ansatz. It turns out that instead of (3.1) we need to consider

$$\psi(t,x) \sim u^\varepsilon\left(t,\tilde{x},\frac{x}{\varepsilon}\right)\exp\left(i\left(\frac{k_0\cdot x}{\varepsilon} - \frac{hE_n(k_0)t}{\varepsilon^2}\right)\right), \tag{3.22}$$

where $u^\varepsilon \sim \sum \varepsilon^j u_j$ and the new spatial coordinate $\tilde{x}$ is given by

$$\tilde{x} := x - \frac{h}{\varepsilon}\,\omega(k_0)t, \quad \text{with } \omega(k_0) := \nabla_k E_n(k_0). \tag{3.23}$$

Thus $\tilde{x}$ comprises a *macroscopically large drift* with a drift-velocity proportional to $\nabla_k E_n(k_0)$. Note that the fast scale x/ε remains unchanged; hence in situations where $\nabla_k E_n(k) = 0$ we are clearly back to our old situation. As before, we plug (3.22) into (2.1), which yields

$$\frac{1}{\varepsilon^2}L_0 u^\varepsilon + \frac{1}{\varepsilon}\tilde{L}_1 u^\varepsilon + L_2 u^\varepsilon + \kappa|u^\varepsilon|^{2\sigma}u^\varepsilon = 0, \tag{3.24}$$

where the linear differential operators L_0, L_2 are defined as in (3.4), (3.5), respectively, but with x replaced by $\tilde{x}$. However, instead of L_1 we have

$$\tilde{L}_1 u^\varepsilon := -h^2\nabla_{\tilde{x}}\cdot\nabla_y u^\varepsilon + ih^2(k_0+\omega(k_0))\cdot\nabla_{\tilde{x}} u^\varepsilon. \tag{3.25}$$

Then, by exactly the same arguments as above, we obtain that the leading order amplitude is given by

$$u_0(t,\tilde{x},y) = f_0(t,\tilde{x})\chi_n(y,k_0) \quad \forall\, t\in\mathbb{R}, x\in\mathbb{R}^d. \tag{3.26}$$

However, setting next the $\mathcal{O}(\varepsilon^{-1})$-term equal to zero yields, instead of (3.12),

$$(H_\Gamma(k_0) - E_n(k_0))u_1 = -i\nabla_x f_0\cdot(\nabla_k H_\Gamma(k_0)\chi_n - \omega(k_0)\chi_n). \tag{3.27}$$

In this case, the solvability condition requires

$$\begin{aligned} 0 &= \langle \chi_n, (\nabla_k H_\Gamma(k_0) - \omega(k_0))\chi_n \rangle_{L^2(\mathfrak{C})} \\ &= \nabla_k E_n(k_0) - \omega(k_0), \end{aligned} \tag{3.28}$$

where we have used the normalization $\langle \chi_n, \chi_n \rangle_{L^2(\mathfrak{C})} = 1$. But (3.28), of course, is identically fulfilled by definition of $\omega(k_0) := \nabla_k E_n(k_0)$. Thus, by identity (2.7), we formally obtain the same $\mathcal{O}(\varepsilon)$-corrector u_1,

$$u_1(t, \tilde{x}, y) = -i\nabla_x f_0(t, \tilde{x}) \cdot \nabla_k \chi_n(y, k_0) + f_1(t, \tilde{x})\chi_n(y, k_0), \tag{3.29}$$

for any given function f_1. (As above, we set f_1 to be identically zero at $t = 0$, for simplicity.) Thus by transforming the x-coordinate into $\tilde{x}$ by (3.23), we can now proceed with our asymptotic expansion, having gained the additional freedom to include the case $\nabla_k E_n(k_0) \neq 0$. The equation of order $\mathcal{O}(1)$ now gives

$$L_0 u_2 + \tilde{L}_1 u_1 + L_2 u_0 + \kappa |u_0|^{2\sigma} u_0 = 0. \tag{3.30}$$

It is clear now that, as before, the corresponding solvability condition, i.e.,

$$\int_{\mathfrak{C}} \overline{\chi_n(y, k_0)} \left(\tilde{L}_1 u_1 + L_2 u_0 + \kappa |u_0|^{2\sigma} u_0 \right) dy = 0, \tag{3.31}$$

yields a homogenized NLS equation of the same form as in (3.19), namely,

$$ih\partial_t f_0 = -\frac{h^2}{2} \operatorname{div}(M^* \nabla) f_0 + U(x) f_0 + \kappa^* |f_0|^{2\sigma} f_0. \tag{3.32}$$

One can easily check that even though $\nabla_k E_n(k_0) \neq 0$ in this case, all additional terms appearing in (3.31) cancel out identically; hence (3.32) remains.

Remark 3.5. In the physics literature [38, 41] the variable-transformation $x \to \tilde{x} := x - h\omega(k_0)t/\varepsilon$ is sometimes reverted, leading to an additional convective term on the left-hand side of (3.32). This, however, can be considered as only a formal statement, since consequently the large factor ε^{-1} would appear then in the homogenized NLS, a somewhat inconsistent formalism.

To prove the existence of smooth solutions for the homogenized NLS we shall impose the following ellipticity assumption.

Assumption 3.6 (ellipticity). We assume that at $k_0 \in \mathfrak{B}$ the following holds:

$$\xi^T M^* \xi \equiv \sum_{k,l=1}^{d} \partial^2_{k_j k_l} E_n(k_0) \xi_j \xi_l \geq C|\xi|^2 \quad \text{for } \xi \in \mathbb{R}^d,\ C > 0. \tag{3.33}$$

Clearly, condition (3.33) is valid if $k_0 \in \mathfrak{B}$ is indeed a local minimum of $E_n(k)$. It may very well be possible to relax the above assumption; cf. Remark 3.8 below. Here we mainly introduced this condition for definiteness, since it is then straightforward to prove that the effective NLS (3.19) (or equivalently (3.32)) has a smooth solution, at least locally-in-time.

LEMMA 3.7. *Let Assumptions* 2.1 *and* 3.6 *be satisfied, and let* $f_I \in \mathcal{S}(\mathbb{R}^d)$. *Then there exists* $\tau = \tau(h) > 0$ *and a unique solution* $f_0 \in C(]-\tau, \tau[; H^s(\mathbb{R}^d))$, $s > d/2$, *satisfying* (3.19) *or equivalently* (3.32). *Moreover,* $x^\alpha f_0 \in C(]-\tau, \tau[; H^s(\mathbb{R}^d))$ *for any* $\alpha \in \mathbb{N}^d$, $s \in \mathbb{N}$.

Proof. By Assumption 3.6 we have that $-\operatorname{div}(M^*\nabla)$ is uniformly elliptic, and since U is subquadratic, the operator $-\operatorname{div}(M^*\nabla)+U$ is therefore well known to be essentially self-adjoint on $C_0^\infty(\mathbb{R}^d)$; cf. [40]. The existence of a smooth solution $f_0(t,\cdot)\in H^s(\mathbb{R}^d)$, $s>d/2$, therefore follows by the same arguments as it does for the standard NLS [9]. The higher order regularity is then also proved similarly to, e.g., [9, 22] (see also the proof of Lemma 4.3 in [11] for the main strategy). ∎

Remark 3.8. Indeed, one can expect similar existence results to be true under much weaker conditions, as given by (3.33); cf. [25]. It is beyond the scope of this work, though, to study the weakest possible assumptions needed (also including, for example, degenerate cases), but rather we refer the reader to [42] and the references given therein. Here we only remark that in the case where M^* is *diagonal*, the existence of smooth local-in-time solutions is known even in the nonelliptic case [17].

Thus, at least for $t\in\,]-\tau,\tau[$, we have that u_0, and hence also u_1, is smooth w.r.t. x,y and, moreover, in $H^s(\mathbb{R}^d)$ w.r.t. x for any $s\in\mathbb{N}$. It is clear that in general we cannot expect smooth global-in-time solutions of the effective NLS (3.19). However, a situation where one indeed has globally smooth solutions, i.e., $\tau=\infty$, is furnished by condition (3.33) together with Assumption 2.1 and imposing that, in addition, $\kappa>0$.

Remark 3.9. Let us briefly compare the obtained leading order asymptotic description of $\psi(t)$ with the one derived in [11] for the weakly nonlinear semiclassical scaling. If we formally set $h=\varepsilon$ in (3.10), (3.26), we obtain

$$\psi(t,x)\sim f_0(t,x-\omega(k_0)t)\chi_l\left(\frac{x}{\varepsilon},k_0\right)e^{i(k_0\cdot x-E_n(k_0)t)/\varepsilon}+\mathcal{O}(\varepsilon), \tag{3.34}$$

which is exactly of the same form as the two-scale WKB ansatz used in [11] (see also [6, 21] for the linear case). In this case the highly oscillatory WKB-phase is simply given by $\phi(t,x)=k_0\cdot x-E_n(k_0)t$. Note that this ϕ solves the nth band *semiclassical Hamilton–Jacobi equation* with vanishing external field and plane wave initial data, i.e.,

$$\begin{cases}\partial_t\phi+E_n(\nabla_x\phi)=0,\\ \phi\big|_{t=0}=k_0\cdot x.\end{cases} \tag{3.35}$$

Moreover, with this choice of ϕ (and since U vanishes), one easily checks that the transport equation for the leading order WKB-amplitude, as derived in [11], simplifies to

$$\begin{cases}\partial_t f_0+\omega(k_0)\cdot\nabla_x f_0=0,\\ f_0\big|_{t=0}=f_I(x).\end{cases} \tag{3.36}$$

Clearly, the solution of (3.36) is then simply given by

$$f_0(t,x)=f_I(t,x-\omega(k_0)t), \tag{3.37}$$

and is hence consistent with our approach.

3.3. Higher order expansions. We henceforth proceed with our ε-expansion of (3.3). Denote the projector onto the nth Bloch band corresponding to a simple eigenvalue $E_n(k)$ by

$$\mathbb{P}_n(k):=|\chi_n(y,k)\rangle\langle\chi_n(y,k)|$$

(using the convenient Dirac notation), and moreover define

$$\mathbb{Q}_n(k) := \mathrm{id} - \mathbb{P}_n(k).$$

This operator is smooth in a vicinity of k_0, and hence, by elliptic inversion, a partial inverse for $L_0 \equiv L_0(k_0)$ can be defined on its range; i.e., $L_0^{-1}\mathbb{Q}(k_0)$ is well defined and smooth. Coming back to (3.30), we can decompose u_2 as

$$u_2(t,x,y) = f_2(t,x)\chi_n\left((y,k_0)\right) + u_2^{\perp}(t,x,y), \tag{3.38}$$

where the function f_2 is yet unknown and $u_2^{\perp}$ is such that

$$\mathbb{P}_n(k_0)u_2^{\perp}(t,x,\cdot) = \langle \chi_n(\cdot,k_0), u_2^{\perp}(t,x,\cdot)\rangle_{L^2(\mathfrak{C})} = 0 \quad \forall\,(t,x) \in \left]-\tau,\tau\right[\times \mathbb{R}^d.$$

Now, $u_2^{\perp}$ is determined by (3.30) via

$$u_2^{\perp} = -L_0^{-1}\mathbb{Q}_n(k_0)\left(\tilde{L}_1 u_1 + L_2 u_0 + \kappa|u_0|^{2\sigma}u_0)\right), \tag{3.39}$$

which implies $u_2^{\perp} \in C(]-\tau,\tau[; H^s(\mathbb{R}^d))$, since this already holds for u_0 and u_1, by Lemma 3.7. As before, (3.39) henceforth induces a particular form of the $\mathcal{O}(\varepsilon^2)$-corrector in the initial amplitude $u_I^{\varepsilon} \sim \sum \varepsilon^j u_j$. The next higher order in ε leads us to the following *linear* problem (after a Taylor expansion of the nonlinearity around u_0):

$$L_0 u_3 + \tilde{L}_1 u_2 + L_2 u_1 + \kappa\left((2\sigma+1)|u_0|^{2\sigma}u_1 + 2\sigma|u_0|^{2\sigma-2}u_0^2\bar{u}_1\right) = 0. \tag{3.40}$$

The corresponding solvability condition then determines $f_1(t,x) \in \mathbb{C}$, i.e., the amplitude corresponding to the *polarized part* of the first order amplitude $u_1(t,x,y)$ given in (3.15). This then leads to a homogenized *linear* Schrödinger-type equation for $f_1(t,x)$, where we have the freedom to choose $f_1(0,x) = 0$. By this procedure, all higher order terms $u_j(t,x,y)$, $j \geq 1$, of the asymptotic solution (3.22) can be obtained, and it is now clear that we can always choose $g_j(0,x)$, i.e., the polarized part of $u_j(0,x,y)$, to be identically zero. In the *globally periodic case*, i.e., $U(x) = 0$ and $\kappa = 0$, the nonvanishing higher order terms $u_j^{\perp}(t,x,y)$, $j \geq 1$, are found to be combinations of higher order derivatives w.r.t. k and x, respectively, of χ_n and f_0; cf. [6, 13]. Although this is no longer true in our case, we still have that $u_j \in C(]-\tau,\tau[; H^s(\mathbb{R}^d))$ for all $j \geq 1$. At each step, though, an additional condition is imposed (recursively) for the initial data ψ_I. This can be seen to be analogous to the situation encountered in [11] and can be understood in the framework of so-called *super-adiabatic subspaces* as constructed in [33].

Remark 3.10. Note that the above given construction can be extended to the case where $E_n(k)$ is an *m-fold degenerate* family of eigenvalues, i.e.,

$$E_n(k_0) = E_*(k_0) \quad \forall\, n \in I \subset \mathbb{N},\ |I| = m,$$

under the additional assumption that there exists a smooth orthonormal basis $\chi_l(y,k_0)$, $l = 1,\dots,m$, of $\operatorname{ran}\mathbb{P}_I(k)$, where $\mathbb{P}_I(k_0)$ denotes the spectral projector corresponding to $E_*(k)$. In this case the leading order asymptotic description would be

$$\psi(t,x) \sim \sum_{l=1}^{m} f_{0,l}(t,\tilde{x})\chi_l\left(\frac{x}{\varepsilon},k_0\right)e^{ik_0\cdot x/\varepsilon}e^{-ihE_n(k_0)t/\varepsilon^2} + \mathcal{O}(\varepsilon). \tag{3.41}$$

In this case, however, we would be forced to consider a *coupled system of homogenized equations*. The corresponding analysis is then analogous to the given one but requires rather tedious computations, a situation which we wanted to avoid for simplicity. For an extensive study of such situations in the linear case we refer to the last section of [3].

4. Nonlinear stability of the asymptotic solution. To prove that the above given multiple-scale expansion indeed yields a good approximation of the exact solution $\psi(t)$ for $\varepsilon \ll 1$, a nonlinear stability result is needed. Note that due to the scaling of (2.1) we cannot hope for any uniform (w.r.t. ε) bound in, say, $H^s(\mathbb{R}^d)$ for $\psi(t)$. On the other hand, the uniform L^2 estimate (2.10) is clearly not sufficient to pass to the limit in the nonlinearity. This motivates the introduction of the following ε-scaled spaces.

DEFINITION 4.1. *For $s \in \mathbb{N}$ let*

$$Y_\varepsilon^s := \left\{ f^\varepsilon \in L^2(\mathbb{R}^d) \; ; \; \sup_{0<\varepsilon\leq 1} \|f^\varepsilon\|_{Y_\varepsilon^s} < +\infty \right\},$$

where

$$\|f^\varepsilon\|_{Y_\varepsilon^s} := \sum_{|\alpha|+|\beta|\leq s} \left\| (\varepsilon x)^\alpha (\varepsilon\partial)^\beta f^\varepsilon \right\|_{L^2(\mathbb{R}^d)}.$$

Remark 4.2. Similar spaces, but without the extra weight ε^α, have been used in the semiclassical study given in [11]. Both variants can be seen as an extension of the H_ε^s-spaces defined by

$$\|f^\varepsilon\|_{H_\varepsilon^s} := \sum_{|\beta|\leq s} \left\| (\varepsilon\partial)^\beta f^\varepsilon \right\|_{L^2(\mathbb{R}^d)}, \tag{4.1}$$

which in the context of geometrical optics expansion were first introduced in [20] (see also [39] and the references given therein). In our case the additional factor $(\varepsilon x)^\alpha$ is needed because we want to include subquadratic potentials $U(x)$ (and due to our scaling we cannot work in the X_ε^s-spaces introduced in [11]). If we would allow $U(x)$ to grow only sublinearly, we could work in H_ε^s as well.

Notation. Let $(\alpha^\varepsilon)_{0<\varepsilon\leq 1}$ and $(\beta^\varepsilon)_{0<\varepsilon\leq 1}$ be two families of positive numbers. From now on we shall write

$$\alpha^\varepsilon \lesssim \beta^\varepsilon$$

if there exists a $C > 0$, independent of $\varepsilon \in]0,1]$ (but possibly dependent on other parameters), such that

$$\alpha^\varepsilon \leq C\beta^\varepsilon \quad \forall \varepsilon \in]0,1].$$

Since for the following results the value of h, appearing in (2.1), is indeed irrelevant, we shall set $h \equiv 1$ throughout this section. Moreover, we shall no longer distinguish between the usual spatial coordinate x and its shifted value $\tilde{x}$, since our results apply in both situations. Next, let us precisely specify the class of well-prepared initial data which we need to consider.

Assumption 4.3 (well-prepared initial data III). The initial data ψ_I^ε satisfies Assumptions 3.2 and 3.3 such that for some $K \in \mathbb{N}$ the following holds:

$$u_I^\varepsilon(x) = \sum_{j=0}^{K} \varepsilon^j u_j(x,y)\Big|_{y=x/\varepsilon} + \mathcal{O}\left(\varepsilon^{K+1}\right) \quad \text{in } Y_\varepsilon^s \text{ for any } s \in \mathbb{N}. \tag{4.2}$$

Moreover, with u_0 and u_1 given by (3.9) and (3.16), respectively, and with $F(z) := |z|^{2\sigma}z$, the functions u_j, $j \geq 2$, are recursively given by

$$u_j = -L_0^{-1}\mathbb{Q}(k_0)\left(\tilde{L}_1 u_{j-1} + L_2 u_{j-2} + \kappa \frac{d^{j-2}}{ds^{j-2}}\, F\left(u_0 + \sum_{l=1}^{j-2} s^l u_l\right)\bigg|_{s=0}\right).$$

After what we have encountered in the construction of higher order asymptotic solutions, this assumption should not come as a surprise. In the linear case the Assumption 4.3 is needed if one aims for refined asymptotic estimates. As we shall see, the inclusion of higher order asymptotics in our case is needed to control the nonlinear term in the proof of the stability result. To this end we need the following existence result for well-prepared initial data.

LEMMA 4.4. *There exists $\psi_I \in \mathcal{S}(\mathbb{R}^d)$ such that Assumption* 4.3 *holds true for any $K \in \mathbb{N}$.*

Proof. The proof follows from Borel's theorem; cf. Theorem 4.2 in [39]. ∎

For the following, define the Nth order asymptotic solution by

$$v_N^\varepsilon(t,x) := \left(\sum_{j=0}^{N} \varepsilon^j u_j\left(t, x, \frac{x}{\varepsilon}\right)\right) e^{ik_0\cdot x/\varepsilon} e^{-ihE_n(k_0)t/\varepsilon^2}, \tag{4.3}$$

and moreover let

$$H^\varepsilon := -\frac{1}{2}\Delta + \frac{1}{\varepsilon^2}V_\Gamma\left(\frac{x}{\varepsilon}\right) + U(x) \tag{4.4}$$

denote the linear part of the Hamiltonian operator. (Note that the scaling of (4.4) is different from the standard semiclassical one as used in [11, 33].) In the foregoing section we obtained the following preliminary result.

PROPOSITION 4.5. *Let ψ_I satisfy Assumption* 4.3 *for any $K \in \mathbb{N}$ and let $\tau > 0$ be the existence-time of smooth solutions to* (3.32). *Then for any $N \in \mathbb{N}$, $v_N^\varepsilon(t)$ solves*

$$\begin{cases} i\partial_t v_N^\varepsilon - H^\varepsilon v_N^\varepsilon = \kappa |v_N^\varepsilon|^{2\sigma} v_N^\varepsilon + \varepsilon^N r_N^\varepsilon, \\ \qquad v_N^\varepsilon\big|_{t=0} = \psi_I + \varepsilon^{N+1}\eta_{N+1}^\varepsilon, \end{cases} \tag{4.5}$$

where $r_N^\varepsilon \in C(]-\tau,\tau[;\, H^s(\mathbb{R}^d))$, $\eta_{N+1}^\varepsilon \in \mathcal{S}(\mathbb{R}^d)$ are such that $r_N^\varepsilon \in L^\infty_{\mathrm{loc}}(]-\tau,\tau[;\, Y_\varepsilon^s)$ and $\|\,\eta_{N+1}^\varepsilon\,\|_{Y_\varepsilon^s} = \mathcal{O}(1)$ for any $s \in \mathbb{N}$.

The main result we shall prove is then given by the following theorem.

THEOREM 4.6. *Let ψ_I satisfy Assumption* 4.3 *for any $K \in \mathbb{N}$, $\tau > 0$ be the existence-time of smooth solutions to* (3.32), *and v_N^ε be given by* (4.3). *Then for any $\tau_0 < \tau$ there exists $\varepsilon_0 > 0$ such that for $0 < \varepsilon \leq \varepsilon_0$, the solution $\psi(t)$ to* (2.1) *is defined on the time-interval $[-\tau_0, \tau_0]$, and moreover*

$$\sup_{t\in[-\tau_0,\tau_0]} \|\psi(t) - v_N^\varepsilon(t)\|_{Y_\varepsilon^s} = \mathcal{O}\left(\varepsilon^{N+1}\right) \tag{4.6}$$

holds for any $N \in \mathbb{N}$ and $s \in \mathbb{N}$.

The above given result shows that if $f(t)$ does not blow up in finite time, then neither does $\psi(t)$, at least for ε sufficiently small. Further, notice that if $\tau = \infty$, then the above given estimate (4.6) holds for any bounded time-interval $[\tau_0, \tau_0] \in \mathbb{R}_t$, in contrast to the (nonlinear) semiclassical situation [11], where the appearance of

caustics usually causes the WKB approach to break down in finite time. Note that in this result, Assumption 4.3 on the initial data ψ_I has to be valid for any $K \in \mathbb{N}$. We shall show in Proposition 4.8 below how to relax this restriction.

Proof. The proof is similar to that given in [11]. Due to the different scaling of our equation, we shall present it here in more detail, which should benefit the reader. Define the difference between the exact and the asymptotic solutions as

$$w_N^\varepsilon(t,x) := \psi(t,x) - v_N^\varepsilon(t,x).$$

Then, from (2.1) and (4.5), w_N solves

$$\begin{cases} i\partial_t w_N^\varepsilon = H^\varepsilon w_N^\varepsilon + \kappa\left(|\psi|^{2\sigma}\psi - |v_N|^{2\sigma}v_N\right) - \varepsilon^N r_N^\varepsilon, \\ w_N^\varepsilon\big|_{t=0} = \varepsilon^{N+1}\eta_{N+1}^\varepsilon. \end{cases} \tag{4.7}$$

By Lemma 3.7 and the well-known Gagliardo–Nirenberg inequality, we have that v_N^ε is uniformly bounded in $L^\infty([-\tau_0,\tau_0]\times\mathbb{R}^d)$. We shall now prove that w_N^ε is also bounded in $L^\infty([-\tau_0,\tau_0]\times\mathbb{R}^d)$, by using a continuity argument which shows that $w_N\varepsilon$ is actually small in that space, for N sufficiently large. To this end we first note that the following important relation holds:

$$\|f^\varepsilon\|_{H^s} = \varepsilon^{-d/2}\|f^\varepsilon\|_{H_\varepsilon^s} \lesssim \varepsilon^{-d/2}\|f^\varepsilon\|_{Y_\varepsilon^s}, \tag{4.8}$$

where the scaling factor $\varepsilon^{-d/2}$ can be easily seen by Fourier transformation. This then directly leads us to an ε-scaled Gagliardo–Nirenberg-type inequality, i.e.,

$$\|w\|_{L^\infty(\mathbb{R}^d)} \lesssim \|w\|_{H^s(\mathbb{R}^d)} \lesssim \varepsilon^{-d/2}\|w\|_{Y_\varepsilon^s} \quad \text{for } s > \frac{d}{2}, \tag{4.9}$$

which we shall use heavily in the following. Multiplying (4.7) by $\overline{w_N^\varepsilon}$, integrating over $\mathbb{R}^d$, and taking the imaginary part yields

$$\partial_t\|w_N^\varepsilon(t)\|_{L^2} \lesssim \left\| |\psi|^{2\sigma}\psi - |v_N^\varepsilon|^{2\sigma}v_N^\varepsilon\right\|_{L^2} + \varepsilon^N\|r_N^\varepsilon(t)\|_{L^2}, \tag{4.10}$$

since H^ε is self-adjoint and $|\kappa| = 1$ by (1.16). To proceed further we recall the following Moser-type lemma, the proof of which is a straightforward generalization of those given in [11, 39].

LEMMA 4.7. *Let $R > 0$, $s \in \mathbb{N}$, and $F(z) = |z|^{2\sigma}z$ for $\sigma \in \mathbb{N}$. Then there exists $C = C(R,s,\sigma,d)$ such that if v satisfies*

$$\left\|(\varepsilon x)^\alpha(\varepsilon\partial)^\beta v\right\|_{L^\infty(\mathbb{R}^d)} \le R \qquad \forall |\alpha| + |\beta| \le s$$

and w satisfies $\|w\|_{L^\infty(\mathbb{R}^d)} \le R$, then it holds that

$$\sum_{|\alpha|+|\beta|\le s} \left\|(\varepsilon x)^\alpha(\varepsilon\partial)^\beta\left(F(v+w) - F(v)\right)\right\|_{L^2(\mathbb{R}^d)} \le C \sum_{|\alpha|+|\beta|\le s} \left\|(\varepsilon x)^\alpha(\varepsilon\partial)^\beta w\right\|_{L^2(\mathbb{R}^d)}.$$

We shall now use this lemma to factor out w_N^ε in the right-hand side of (4.10) and then take advantage of the smallness of the remainder. By construction, $w_N^\varepsilon(0,x) = \mathcal{O}(\varepsilon^{N+1})$ in any Y_ε^s. By Lemma 3.7 we can find for fixed $\tau_0 < \tau$ an $R > 0$ such that if $N + 1 > d/2$, then

$$\|w_N^\varepsilon(t)\|_{L^\infty} \le R \tag{4.11}$$

for ε sufficiently small. Hence, as long as (4.11) holds, (4.10) and the above given Lemma 4.7, with $s = 0$, imply

$$\partial_t \|w_N^\varepsilon(t)\|_{L^2} \leq C\|w_N^\varepsilon(t)\|_{L^2} + C\varepsilon^N \|r_N^\varepsilon(t)\|_{L^2}.$$

Thus by a Gronwall-type estimate, we get, as long as (4.11) holds, that

$$\|w_N^\varepsilon(t)\|_{L^2} \lesssim \varepsilon^N \tag{4.12}$$

for $t \leq \tau$. Next we shall show how to obtain similar estimates for the momenta and derivatives of w_N^ε. Applying the operator $\varepsilon\nabla_x$ to (4.7) yields (where, as before, $F(z) := |z|^{2\sigma} z$)

$$\begin{aligned} i\partial_t(\varepsilon\nabla_x w_N^\varepsilon) = H^\varepsilon(\varepsilon\nabla_x w_N^\varepsilon) + \kappa\,\varepsilon\nabla_x\left(F(\psi) - F(v_N^\varepsilon)\right) \\ + [\varepsilon\nabla_x, H^\varepsilon]\, w_N^\varepsilon - \varepsilon^{N+1}\nabla_x r_N^\varepsilon, \end{aligned}$$

and hence we obtain the following energy estimate:

$$\begin{aligned} \partial_t\|\varepsilon\nabla_x w_N^\varepsilon(t)\|_{L^2} \lesssim \|\varepsilon\nabla_x\left(F(\psi) - F(\mathfrak{v}_N^\varepsilon)\right)\|_{L^2} + \|[\varepsilon\nabla_x, H^\varepsilon]\, w_N^\varepsilon\|_{L^2} \\ + \varepsilon^N\|\varepsilon\nabla_x r_N^\varepsilon\|_{L^2}. \end{aligned} \tag{4.13}$$

On the other hand, we compute from (4.4) that

$$[\varepsilon\nabla_x, H^\varepsilon] = \frac{1}{\varepsilon^2}\nabla_x V_\Gamma\left(\frac{x}{\varepsilon}\right) + \varepsilon\nabla_x U(x).$$

Since ∇V_Γ is bounded and ∇U is sublinear, (4.13) consequently yields

$$\begin{aligned} \partial_t\|\varepsilon\nabla_x w_N^\varepsilon(t)\|_{L^2} \lesssim \|\varepsilon\nabla_x\left(F(\psi) - F(v_N^\varepsilon)\right)\|_{L^2} + \frac{1}{\varepsilon^2}\|w_N^\varepsilon\|_{L^2} + \|\varepsilon x\, w_N^\varepsilon\|_{L^2} \\ + \varepsilon^N\|\varepsilon\nabla_x r_N^\varepsilon\|_{L^2}. \end{aligned}$$

Thus, again using Lemma 4.7 (with $s = 1$) together with Proposition 4.5 and the estimate (4.12), we get

$$\partial_t\|\varepsilon\nabla_x w_N^\varepsilon(t)\|_{L^2} \lesssim \|\varepsilon\nabla_x w_N^\varepsilon\|_{L^2} + \|\varepsilon x\, w_N^\varepsilon\|_{L^2} + \varepsilon^{N-2}. \tag{4.14}$$

(Note the difference in the last term as compared to the semiclassical estimate obtained in [11].) To obtain an estimate for $\|\varepsilon x\, w_N^\varepsilon\|_{L^2}$, we proceed analogously to obtain the following moment estimate:

$$\begin{aligned} \partial_t\|\varepsilon x\, w_N^\varepsilon(t)\|_{L^2} \lesssim \|\varepsilon x\left(F(\psi) - F(v_N^\varepsilon)\right)\|_{L^2} + \|[\varepsilon x, H^\varepsilon]\, w_N^\varepsilon\|_{L^2} \\ + \varepsilon^N\|\varepsilon x\, r_N^\varepsilon\|_{L^2}. \end{aligned} \tag{4.15}$$

But, since $[\varepsilon x, H^\varepsilon] = -\varepsilon\nabla_x$, we get, as long as (4.11) holds,

$$\begin{aligned} \partial_t\|\varepsilon x\, w_N^\varepsilon(t)\|_{L^2} &\lesssim \|\varepsilon x\left(F(\psi) - F(v_N^\varepsilon)\right)\|_{L^2} + \|\varepsilon\nabla_x w_N^\varepsilon\|_{L^2} + \varepsilon^N\,\|\varepsilon x\, r_N^\varepsilon\|_{L^2} \\ &\lesssim \|\varepsilon x\, w_N^\varepsilon(t)\|_{L^2} + \|\varepsilon\nabla_x w_N^\varepsilon\|_{L^2} + \varepsilon^N. \end{aligned} \tag{4.16}$$

Putting (4.14) and (4.16) together, we have

$$\partial_t\left(\|\varepsilon\nabla_x w_N^\varepsilon\|_{L^2} + \|\varepsilon x\, w_N^\varepsilon(t)\|_{L^2}\right) \lesssim \|\varepsilon\nabla_x w_N^\varepsilon\|_{L^2} + \|\varepsilon x\, w_N^\varepsilon(t)\|_{L^2} + \varepsilon^{N-2}.$$

Hence a Gronwall lemma yields

$$\|w_N^\varepsilon(t)\|_{Y_\varepsilon^1} \lesssim \varepsilon^{N-2} \tag{4.17}$$

as long as (4.11) holds, and by induction one arrives at

$$\|w_N^\varepsilon(t)\|_{Y_\varepsilon^s} \lesssim \varepsilon^{N-2s}\,. \tag{4.18}$$

For $s > d/2$, and as long as (4.11) holds, the Gagliardo–Nirenberg-type inequality (4.9) therefore implies

$$\|w_N^\varepsilon(t)\|_{L^\infty(\mathbb{R}^d)} \lesssim \varepsilon^{-d/2}\|w_N^\varepsilon(t)\|_{Y_\varepsilon^s} \lesssim \varepsilon^{N-2s-d/2}.$$

Hence, if indeed $N-2s-d/2>0$ holds true, a continuity argument shows that (4.11) is valid up to times $|t|=\tau$, provided that ε is sufficiently small. In particular, w_N^ε, and hence ψ, is well defined up to times $|t|=\tau_0<\tau$, for $0<\varepsilon\le\varepsilon(\tau)$. It remains to prove the estimate (4.6). Fix $s,N\in\mathbb{N}$ and let $s_1\ge s$ be such that $s_1>d/2$, as well as $N_1\ge 2s_1+N+1$. From (4.18) we conclude that

$$\sup_{t\in[-\tau_0,\tau_0]} \left\|w_{N_1}^\varepsilon(t)\right\|_{Y_\varepsilon^{s_1}} \lesssim \varepsilon^{N_1-2s_1} \lesssim \varepsilon^{N+1}.$$

Since $N_1>N$, it is therefore straightforward that

$$\sup_{t\in[-\tau_0,\tau_0]} \left\|v_N^\varepsilon(t)-v_{N_1}^\varepsilon(t)\right\|_{Y_\varepsilon^{s_1}} \lesssim \varepsilon^{N+1},$$

and hence we deduce that (4.6) holds for any $s,N\in\mathbb{N}$. ∎

In the proof given above, the initial data ψ_I is assumed to be well prepared up to any order $K\in\mathbb{N}$. This rather strong assumption can be relaxed, as the next result will show. To this end we introduce the following notation.

Notation. For every $\alpha\in\mathbb{R}$ we denote by $[\alpha]\in\mathbb{N}$ the *ceiling of* α, i.e., the smallest integer which is larger than or equal to α.

PROPOSITION 4.8. *Let $\widetilde{\psi}(t)$ be the solution of* (2.1) *corresponding to an initial data $\widetilde{\psi}_I$, which satisfies Assumption* 4.3 *for any $K\in\mathbb{N}$. On the other hand, let $\psi(t)$ be the solution corresponding to an initial data ψ_I, where ψ_I is such that Assumption* 4.3 *is satisfied for $K\ge[3d/2]$. Then for any $\tau_0\in\,]-\tau,\tau[$ there exists $\varepsilon_0>0$ such that for $0<\varepsilon\le\varepsilon_0$, $\psi^\varepsilon(t)$ is defined up to times $|t|\le\tau_0$, and moreover the following holds:*

$$\sup_{t\in[-\tau_0,\tau_0]} \left\|\psi(t)-\widetilde{\psi}(t)\right\|_{Y_\varepsilon^s} = \mathcal{O}\left(\varepsilon^{K+1-2s}\right) \quad \textit{for } s\ge 0\,.$$

Proof. Since the proof follows the lines of that for Theorem 4.6, we shall be rather brief. As before, we introduce

$$\widetilde{w}(t,x) := \psi(t,x)-\widetilde{\psi}(t,x).$$

Then $\widetilde{w}(t)$ solves

$$\begin{cases} i\partial_t\widetilde{w} = H^\varepsilon\widetilde{w} + \kappa\left(|\psi|^{2\sigma}\psi - |\widetilde{\psi}|^{2\sigma}\widetilde{\psi}\right), \\ \widetilde{w}\big|_{t=0} = \mathcal{O}\left(\varepsilon^{K+1}\right) \quad \text{in } Y_\varepsilon^s \text{ for any } s\in\mathbb{N}. \end{cases}$$

(Note that there is no remainder r_N^ε in this case.) We can then argue as in the above given proof. We have that initially the following holds:

$$\|\widetilde{w}(0,\cdot)\|_{L^\infty} \lesssim \varepsilon^{-d/2}\|\widetilde{w}(0,\cdot)\|_{Y_\varepsilon^s} \lesssim \varepsilon^{K+1-d/2}, \quad \text{provided } s > \frac{d}{2}.$$

With $K+1 \geq [3d/2]+1 > d/2$, the same arguments as in the proof of Theorem 4.6 yield

$$\|\widetilde{w}(t)\|_{Y_\varepsilon^s} \lesssim \varepsilon^{K+1-2s},$$

as long as (4.11) holds. Since $K+1 > d$, we can choose $s > d/2$ such that $K+1-2s > d/2$; i.e., we can choose s such that $K+1 > [d/2+2s] = [3d/2]$. Therefore the above given estimate and (4.9) show that (4.11) holds up to times $|t| = \tau_0$, for $\varepsilon \ll 1$. ∎

Theorem 4.6 and Proposition 4.8 then finally lead to the following statement, proving also Theorem 1.3.

COROLLARY 4.9. *If ψ_I satisfies Assumption 4.3 with $K \geq [3d/2]$, then there exists $\varepsilon_0 > 0$ such that for $0 < \varepsilon \leq \varepsilon_0$ the solution $\psi(t)$ to (2.1) is defined on the time interval $[-\tau_0, \tau_0]$ for any $\tau_0 < \tau$ and the following estimate holds:*

$$\sup_{t\in[-\tau_0,\tau_0]} \|\psi(t) - v_0^\varepsilon(t)\|_{L^2(\mathbb{R}^d)} = \mathcal{O}(\varepsilon). \tag{4.19}$$

Additionally, if $K > [3d/2]$, then we have

$$\sup_{t\in[-\tau_0,\tau_0]} \|\psi(t) - v_0^\varepsilon(t)\|_{L^\infty(\mathbb{R}^d)} = \mathcal{O}(\varepsilon). \tag{4.20}$$

Proof. From Proposition 4.8 we deduce that Theorem 4.6 holds with $K \geq [3d/2]$. The L^2 estimate (4.19) then is nothing but (4.6) with $N = s = 0$. The L^∞ estimate (4.20) follows similarly from Theorem 4.6 and (4.9). ∎

In other words we deduce that the solution of (2.1) can be, up to an error of $\mathcal{O}(\varepsilon)$, approximated by the leading order asymptotic solution v_0, obtained from the multiple-scales expansion, if the initial data is well prepared, i.e., including correctors up to $K = [3d/2]$, which is slightly stronger than what was required for the semiclassical result given in [11]. There the analogous condition for the correctors was $K \geq d$. On the other hand, one might guess that the leading order estimates (4.19), (4.20) are true even if the initial data is "correct" only up to leading order. However, the techniques used in the above proofs do not allow the conclusion that this is indeed the case. Note, however, that the higher order correctors in the initial data tend to zero as $\varepsilon \to 0$ in practically every reasonable sense.

Remark 4.10. Finally, let us remark that one could also study the semiclassical asymptotic behavior of the homogenized NLS, i.e., the limit $h \to 0$ of (3.19), although, in view of the given scaling arguments and Remark 3.9, the word "semiclassical" should rather be understood here in a purely mathematical sense. To this end, the well-known WKB-type method derived in [19] can be adapted under suitable conditions on M^*. In this case, however, one can only hope for local-in-time results, i.e., results up to caustics. The combined limit $\varepsilon/h \to 0$, $h \to 0$ seems to be more subtle, in particular due to the somewhat hidden dependence of τ on h in the above results.

Acknowledgments. The author is grateful to R. Carles for helpful discussions on this work.

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[2] G. Allaire, Y. Capdeboscq, A. Piatnitski, V. Siess, and M. Vanninathan, *Homogenization of periodic systems with large potentials*, Arch. Ration. Mech. Anal., 174 (2004), pp. 179–220.

[3] G. Allaire and A. Piatnitski, *Homogenization of the Schrödinger equation and effective mass theorems*, Comm. Math. Phys., 258 (2005), pp. 1–22.

[4] W. Bao, D. Jaksch, and P. Markowich, *Numerical solution of the Gross–Pitaevskii equation for Bose–Einstein condensation*, J. Comput. Phys., 187 (2003), pp. 318–342.

[5] P. Bechouche, N. Mauser, and F. Poupaud, *Semiclassical limit for the Schrödinger–Poisson equation in a crystal*, Comm. Pure Appl. Math., 54 (2001), pp. 851–890.

[6] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North–Holland, Amsterdam, 1978.

[7] F. Bloch, *Über die Quantenmechanik der Elektronen in Kristallgittern*, Z. Phys., 52 (1928), pp. 555–600 (in German).

[8] J. Callaway, *Quantum Theory of the Solid State*, Academic Press, New York, 1991.

[9] T. Cazenave, *Semilinear Schrödinger Equations*, Courant Lect. Notes Math. 10, New York University Courant Math. Institute, New York, 2003.

[10] R. Carles, *Remarks on nonlinear Schrödinger equations with harmonic potential*, Ann. Henri Poincaré, 3 (2002), pp. 757–772.

[11] R. Carles, P. Markowich, and C. Sparber, *Semi-classical asymptotics for weakly nonlinear Bloch waves*, J. Statist. Phys., 117 (2004), pp. 369–400.

[12] D. Choi and Q. Niu, *Bose–Einstein condensates in an optical lattice*, Phys. Rev. Lett., 82 (1999), pp. 2022–2025.

[13] C. Conca and M. Vanninathan, *Homogenization of periodic structures via Bloch decomposition*, SIAM J. Appl. Math., 57 (1997), pp. 1639–1659.

[14] B. Deconinck, B. Frigyik, and J. N. Kutz, *Dynamics and stability of Bose–Einstein condensates: The nonlinear Schrödinger equation with periodic potential*, J. Nonlinear Sci., 12 (2002), pp. 169–205.

[15] P. Donnat and J. Rauch, *Dispersive nonlinear geometrical optics*, J. Math. Phys., 38 (1997), pp. 1484–1523.

[16] J. Garnier, *Homogenization in a periodic and time-dependent potential*, SIAM J. Appl. Math., 57 (1997), pp. 95–111.

[17] J. M. Ghidaglia and J. C. Saut, *Nonelliptic nonlinear Schrödinger equations*, J. Nonlinear Sci., 3 (1993), pp. 169–195.

[18] P. Gérard, P. Markowich, N. Mauser, and F. Poupaud, *Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math, 50 (1997), pp. 323–378.

[19] E. Grenier, *Semiclassical limit of the nonlinear Schrödinger equation in small time*, Proc. Amer. Math. Soc., 126 (1998), pp. 523–530.

[20] O. Gués, *Développement asymptotique de solutions exactes de systémes hyperboliques quasilinéaires*, Asymptot. Anal., 6 (1993), pp. 241–269 (in French).

[21] J. C. Guillot, J. Ralston, and E. Trubowitz, *Semiclassical asymptotics in solid-state physics*, Comm. Math. Phys., 116 (1998), pp. 401–415.

[22] N. Hayashi, N. Nakamitsu, and M. Tsutsumi, *Nonlinear Schrödinger equations in weighted Sobolev spaces*, Funkcial. Ekvac., 31 (1988), pp. 363–381.

[23] V. Jikov, S. Kozlov, and O. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer, New York, 1994.

[24] J. L. Joly, G. Métivier, and J. Rauch, *Diffractive nonlinear geometric optics with rectification*, Indiana Univ. Math. J., 47 (1998), pp. 1167–1241.

[25] C. E. Kenig, G. Ponce, and L. Vega, *Smoothing effects and local existence theory for the generalized nonlinear Schrödinger equations*, Invent. Math., 134 (1998), pp. 489–545.

[26] E. B. Kolomeisky, T. J. Newman, J. P. Straley, and X. Qi, *Low dimensional Bose liquids: Beyond the Gross–Pitaevskii approximation*, Phys. Rev. Lett., 85 (2000), pp. 1146–1149.

[27] V. V. Konotop and M. Salerno, *Modulational instability in Bose–Einstein condensates in optical lattices*, Phys. Rev. A, 65 (2002), pp. 21602–21606.

[28] M. Kraemer, C. Menotti, L. Pitaevskii, and S. Stringari, *Bose–Einstein condensates in 1D optical lattices: Compressibility, Bloch bands, and elementary excitations*, Eur. Phys. J. D. At. Mol. Opt. Phys., 27 (2003), pp. 247–263.

[29] E. Lieb, R. Seiringer, and J. Yngvason, *One-dimensional behavior of dilute, trapped Bose gases*, Comm. Math. Phys., 244 (2004), pp. 347–393.

[30] P. A. Markowich, N. Mauser, and F. Poupaud, *A Wigner-function approach to (semi)classical limits: Electrons in a periodic potential*, J. Math. Phys., 35 (1994), pp. 1066–1094.

[31] G. Nenciu, *Dynamics of band electrons in electric and magnetic fields: Rigorous justification of the effective Hamiltonian*, Rev. Modern Phys., 63 (1991), pp. 547–578.

[32] O. Morsch and E. Arimondo, *Ultracold atoms and Bose–Einstein condensates in optical lattices*, in Dynamics and Thermodynamics of Systems with Long-Range Interactions, Lecture Notes in Phys. 602, T. Dauxois, S. Ruffo, E. Arimondo, and M. Wilkens, eds., Springer, New York, 2002, pp. 312–331.

[33] G. Panati, H. Sohn, and S. Teufel, *Effective dynamics for Bloch electrons: Peierls substitution and beyond*, Comm. Math. Phys., 242 (2003), pp. 547–578.

[34] A. Pankov, *G-Convergence and Homogenization of Nonlinear Partial Differential Operators*, Math. Appl. 422, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

[35] F. Pedersen, *Simple derivation of the effective-mass equation using a multiple-scale technique*, European J. Phys., 18 (1997), pp. 43–45.

[36] L. Pitaevskii and S. Stringari, *Bose–Einstein Condensation*, Internat. Ser. Monogr. Phys. 116, Clarendon Press, Oxford, UK, 2003.

[37] F. Poupaud and C. Ringhofer, *Semiclassical limits in a crystal with external potentials and effective mass theorems*, Comm. Partial Differential Equations, 21 (1996), pp. 1897–1918.

[38] H. Pu, L. O. Baksmaty, W. Zhang, N. P. Bigelow, and P. Meystre, *Effective-mass analysis of Bose–Einstein condensates in optical lattices: Stabilization and levitation*, Phys. Rev. A, 67 (2003), pp. 43605–43612.

[39] J. Rauch, *Lectures on Nonlinear Geometrical Optics*, IAS/Park City Math. Ser. 5, AMS, Providence, RI, 1999.

[40] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* IV. *Analysis of Operators*, Academic Press, New York, 1978.

[41] M. J. Steel and W. Zhang, *Bloch Function Description of a Bose–Einstein Condensate in a Finite Optical Lattice*, preprint, arXiv: cond-mat/9810284.

[42] C. Sulem and P. L. Sulem, *The Nonlinear Schrödinger Equation*, Appl. Math. Sci. 139, Springer, New York, 1999.

[43] L. Tartar, *H-measures, A new approach for studying homogenisation, oscillations, and concentration effects in partial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 115 (1990), pp. 193–230.

[44] S. Teufel, *Adiabatic Perturbation Theory in Quantum Dynamics*, Lecture Notes in Math. 1821, Springer, New York, 2003.

[45] C. H. Wilcox, *Theory of Bloch waves*, J. Anal. Math., 33 (1978), pp. 146–167.

SIAM J. APPL. MATH.
Vol. 66, No. 3, pp. 843–872

THE ROLE OF COINFECTION IN MULTIDISEASE DYNAMICS*

MAIA MARTCHEVA† AND SERGEI S. PILYUGIN‡

Abstract. We investigate an epidemic model of two diseases. The primary disease is assumed to be a slowly progressing disease, and the density of individuals infected with it is structured by age since infection. Hosts that are already infected with the primary disease can become coinfected with a secondary disease. We show that in addition to the disease-free equilibrium, there exists a unique dominance equilibrium corresponding to each disease. Without coinfection there are no coexistence equilibria; however, with coinfection the number of coexistence equilibria may vary. For some parameter values, there exist two coexistence equilibria. We also observe competitor-mediated oscillatory coexistence. Furthermore, weakly subthreshold (which occur when exactly one of the reproduction numbers is below one) and strongly subthreshold (which occur when both reproduction numbers are below one) coexistence equilibria may exist. Some of those are a result of a two-parameter backward bifurcation. Bistability occurs in several regions of the parameter space. Despite the presence of coinfection, coexistence of the two diseases appears possible only for relatively small values of the reproduction numbers—for large values of the reproduction numbers the typical outcome of competition is the dominance of one of the diseases, including bistable dominance where the competition outcome is initial condition dependent.

Key words. coinfection, infection-age structure, subthreshold coexistence, backward bifurcation, Hopf bifurcation, oscillatory dominance, oscillatory coexistence, restricted pathogenic diversity

AMS subject classifications. 35B32, 35B35, 35B38, 35F25, 92D30

DOI. 10.1137/040619272

1. Introduction.

Coinfection is a simultaneous infection of one host with multiple pathogens that may be the causative agents of different diseases or variants of the same parasite. Coinfections are common for individuals infected with the human immunodeficiency virus (HIV). Since HIV compromises the immune system, the carrier becomes vulnerable to other infections commonly called opportunistic infections [10]. For instance, the case of HIV-HSV (herpes simplex virus) coinfection has been well documented. Such coinfection typically leads to reactivation of HSV, which accelerates the progression of HIV disease towards AIDS. HIV-HSV infected individuals are also more likely to unwittingly transmit HSV via an increased shedding common in HIV-infected patients. The treatment of HIV-HSV coinfected patients may present additional challenges since the HSV is likely to be more resistant to antiviral therapy [22]. Coinfections may also occur when a patient is already infected with a slowly progressing disease which lasts for decades. In tuberculosis (TB), for example, coinfection even with a minor illness can trigger a reactivation of TB.

Many mathematical studies exist on single diseases, both general theoretic and those treating a specific disease. At the same time, few studies exist that address the interaction of two or more diseases. On an epidemiological level, Courchamp et al. [7] studied a model of two feline retroviruses. Two recent articles—one by Allen,

*Received by the editors November 19, 2004; accepted for publication (in revised form) August 23, 2005; published electronically February 21, 2006.

http://www.siam.org/journals/siap/66-3/61927.html

†Corresponding author. Department of Mathematics, University of Florida, Gainesville, FL 32611-8105 (maia@math.ufl.edu). The research of this author was partially supported by NSF grants DMS-0137687 and DMS-0406119.

‡Department of Mathematics, University of Florida, Gainesville, FL 32611-8105 (pilyugin@math.ufl.edu). This author was visiting the Mathematical Biosciences Institute (MBI) at The Ohio State University for part of the time when this article was being written.

Langlais, and Phillips [1] and another by Gumel et al. [9]—also consider two infections in a single host. On an immunological level, the interactions between *Mycobacterium tuberculosis* and HIV-1 are investigated in [11]. Statistical aspects of mapping two diseases are the focus of [8]. Coinfection has been studied from a general perspective in [14]. In [16], a connection between superinfection and coinfection with different strains, and the impact of both on the coexistence and the evolution of virulence, is discussed.

In this paper, we study an epidemic model with two diseases that can coinfect a single host. We include infection-age structure in the primary disease to account for slowly progressing and/or persistent diseases that affect the immune status of individuals over time. Infection-age structure has been previously shown to cause qualitative changes, namely oscillatory behavior, in case of a single disease dynamics [2, 19, 15, 12, 20]. Although the model discussed here is relatively simple, we find complex dynamic behavior: oscillatory dominance and coexistence, two-parameter backward bifurcation, multiple and subthreshold coexistence equilibria, and bistability. These phenomena have important epidemiological consequences for disease management. Most of them have been illustrated in the single-disease case. In particular, backward bifurcation which leads to multiple and subthreshold equilibria has been attracting significant attention in the literature (see [13] and the references therein). However, to the best of our knowledge, backward bifurcations have not been studied in the context of multiple infectious agents.

This paper is organized as follows. In the next section, we introduce the two-disease coinfection model. In section 3, we introduce the reproduction numbers of the primary and secondary diseases $\mathcal{R}_1$, $\mathcal{R}_2$ and discuss the equilibria of the model. The values of the disease-free equilibrium and the two boundary equilibria are given explicitly. We also present sufficient conditions for the existence of a coexistence equilibrium. In section 4, we consider scenarios for extinction of either disease or both. Section 5 focuses on the local stability of equilibria. We show that both the primary disease equilibrium and coexistence equilibria can lose stability, leading to sustained oscillations. Section 6 is devoted to the derivation of necessary and sufficient conditions for the backward bifurcation in $\mathcal{R}_1$ and $\mathcal{R}_2$. In section 7, we present several numerical simulations to illustrate the various complex dynamic phenomena. In section 8, we discuss the epidemiological implications of our model. Section 9 contains a summary of our results and concludes the paper.

2. A model of coinfection of two diseases. Two diseases are spreading in a population of total size $N(t)$. They both compete for the same pool of susceptible individuals, whose number at time t is denoted by $S(t)$. We assume that the first disease is a slowly progressing one, and we structure the class of infected individuals with respect to the time since infection, a. The age-density is denoted by $i(a,t)$. The total number of individuals infected with the first disease is denoted by $I_1(t)$. Population members who eventually contract both diseases are assumed to be infected by the slowly progressing disease first. Consequently we call it the *primary disease*. A susceptible becomes infected with the primary disease at a rate $\beta_1(a)$. The number of individuals infected with the second disease is denoted by $I_2(t)$. The secondary disease is transmitted by the class I_2 to susceptibles at a rate β_2. An individual already infected with the primary disease can be coinfected with the secondary disease at a rate $\delta(a)$ and thus become jointly infected with both diseases. We denote the number of jointly infected (coinfected) individuals by $J(t)$. The individuals infected with both diseases can infect susceptibles with the primary disease at a rate γ_1 and

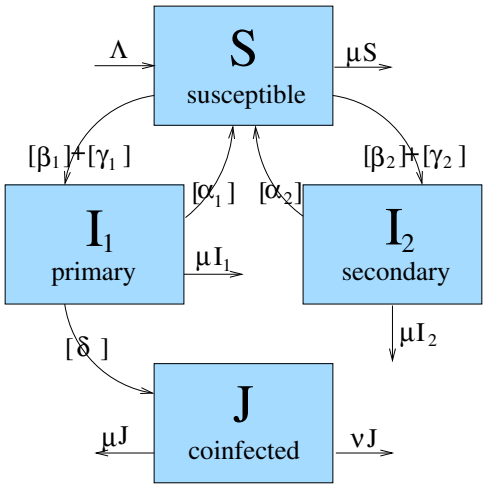


FIG. 2.1. *The flow diagram of the model* (2.1). *The primary infection rate is shown as* $[\beta_1]+[\gamma_1]$, *where* $[\beta_1] = \frac{S}{N}\int_0^\infty \beta_1(a)i(a)\,da$ *and* $[\gamma_1] = \gamma_1\frac{SJ}{N}$. *The secondary infection rate is shown as* $[\beta_2]+[\gamma_2]$, *where* $[\beta_2] = \beta_2\frac{SI_2}{N}$ *and* $[\gamma_2] = \gamma_2\frac{SJ}{N}$. *The coinfection rate is shown as* $[\delta] = \frac{I_2}{N}\int_0^\infty \delta(a)i(a)\,da$. *The primary and secondary recovery rates are shown as* $[\alpha_1] = \int_0^\infty \alpha_1(a)i(a)\,da$ *and* $[\alpha_2] = \alpha_2 I_2$. *The parameters* Λ, μ, *and* ν *represent the birth/recruitment rate, the background mortality rate, and the disease-induced mortality associated with coinfection, respectively.*

with the secondary disease at a rate γ_2. Figure 2.1 presents a schematic flow diagram of the mathematical model that takes the form

$$
\begin{aligned}
S' &= \Lambda - \frac{S}{N}\int_0^\infty \beta_1(a)i(a,t)\,da - \beta_2\frac{SI_2}{N} - (\gamma_1+\gamma_2)\frac{SJ}{N} - \mu S \\
&\qquad + \int_0^\infty \alpha_1(a)i(a,t)\,da + \alpha_2 I_2, \\
(\partial_t + \partial_a)i(a,t) &= -\alpha_1(a)i(a,t) - \delta(a)\frac{I_2}{N}i(a,t) - \mu i(a,t), \\
i(0,t) &= \frac{S}{N}\int_0^\infty \beta_1(a)i(a,t)\,da + \gamma_1\frac{SJ}{N}, \qquad (2.1) \\
I_2' &= \beta_2\frac{SI_2}{N} + \gamma_2\frac{SJ}{N} - (\mu+\alpha_2)I_2, \\
J' &= \frac{I_2}{N}\int_0^\infty \delta(a)i(a,t)\,da - (\mu+\nu)J,
\end{aligned}
$$

where μ is natural death rate. We assume that either disease by itself is not lethal but that the two in combination can be. The biological motivation of this assumption is the case of HIV/AIDS, where coinfections in late stages of the HIV (considered the primary disease) can be terminal. Specifically, we assume that jointly infected individuals do not recover and that coinfections cause disease-induced mortality at a rate ν. Individuals infected with either primary or secondary disease alone may be potentially treated and recover at rates $\alpha_1(a)$, α_2, respectively. The functions $\alpha_1(a)$, $\beta_1(a)$, and $\delta(a)$ are nonnegative and bounded. The parameters β_2, γ_1, γ_2, ν, α_2 are nonnegative, whereas $\Lambda > 0$ and $\mu > 0$. A standard argument can be used to show that the model (2.1) is well posed.

The total population size $N(t)$ is the sum of all individuals in all classes:

$$N(t) = S(t) + \int_0^\infty i(a,t)da + I_2(t) + J(t). \tag{2.2}$$

The total population size satisfies the equation $N'(t) = \Lambda - \mu N - \nu J$. We introduce the notation

$$\pi_1(a) = e^{-\int_0^a \alpha_1(s)ds}.$$

To understand the biological meaning of the quantity $\pi_1(a)$ we note that $\pi_1(a)e^{-\mu a}$ is the probability that an individual will remain infected with the primary disease a time units after infection. In addition, we define the quantity

$$\Delta = \int_0^\infty \alpha_1(a)\pi_1(a)e^{-\mu a}da, \tag{2.3}$$

which gives the probability of leaving the primary disease infectious period via recovery. Since individuals can leave the primary infected class only via recovery or death, the sum of the probabilities of recovery and death equals one; that is,

$$\int_0^\infty \alpha_1(a)\pi_1(a)e^{-\mu a}\,da + \mu\int_0^\infty \pi_1(a)e^{-\mu a}\,da = \int_0^\infty (\mu+\alpha_1(a))e^{-\int_0^a(\mu+\alpha_1(s))ds}\,da = 1.$$

It immediately follows that $\Delta < 1$.

3. Equilibria of the model with coinfection. We introduce the reproduction numbers of the two diseases. The reproduction number of the primary disease is

$$\mathcal{R}_1 = \int_0^\infty \beta_1(a)\pi_1(a)e^{-\mu a}\,da, \tag{3.1}$$

and the reproduction number of the secondary disease is

$$\mathcal{R}_2 = \frac{\beta_2}{\mu+\alpha_2}. \tag{3.2}$$

We note that the coinfection rate $\delta(a)$ does not affect the reproduction numbers since coinfection does not lead to additional infections. We will adopt the notation $s = S/N^*$, $i_2 = I_2/N^*$, $j = J/N^*$ and will use $i(a)$ to denote the normalized version of the equilibrium value of $i(a,t)$, $i^*(a)$. The quantity N^* is given by the sum (2.2) at an equilibrium. Let us define

$$\Gamma(a;i_2) = e^{-i_2\int_0^a \delta(\sigma)d\sigma}. \tag{3.3}$$

Notice that $\Gamma(a;0) = 1$. Setting the derivatives with respect to time to zero, we obtain a system of algebraic equations and one ODE for the equilibria of (2.1). The ODE in the system can be solved to yield

$$i(a) = i(0)\Gamma(a;i_2)\pi_1(a)e^{-\mu a}. \tag{3.4}$$

Substituting for i in the integrals, one obtains

$$\int_0^\infty \beta_1(a)i(a)\,da = i(0)\int_0^\infty \beta_1(a)\Gamma(a;i_2)\pi_1(a)e^{-\mu a}\,da = i(0)B(i_2)$$

and

$$\int_0^\infty \alpha_1(a)i(a)\,da = i(0)\int_0^\infty \alpha_1(a)\Gamma(a;i_2)\pi_1(a)e^{-\mu a}\,da = i(0)A(i_2).$$

Finally,

$$\int_0^\infty \delta(a)i(a)\,da = i(0)\int_0^\infty \delta(a)\Gamma(a;i_2)\pi_1(a)e^{-\mu a}\,da = i(0)D(i_2).$$

We notice that $A(i_2) < 1$ and $i_2D(i_2) + A(i_2) < 1$ because

$$\begin{aligned} i_2D(i_2) + A(i_2) &= \int_0^\infty (\alpha_1(a) + i_2\delta(a))e^{-\int_0^a(\alpha_1(\sigma)+i_2\delta(\sigma))\,d\sigma}e^{-\mu a}\,da \\ &< \int_0^\infty (\alpha_1(a) + i_2\delta(a))e^{-\int_0^a(\alpha_1(\sigma)+i_2\delta(\sigma))\,d\sigma}\,da = 1. \end{aligned}$$

With this notation the system for the equilibria becomes

$$\begin{aligned} &0 = \mu - si(0)B(i_2) - \beta_2 si_2 - (\gamma_1+\gamma_2)sj - \mu s + i(0)A(i_2) + \alpha_2 i_2 + \nu j,\\ &i(0) = i(0)sB(i_2) + \gamma_1 sj,\\ &0 = \beta_2 si_2 + \gamma_2 sj - (\mu+\alpha_2)i_2,\\ &0 = i(0)i_2D(i_2) - (\mu+\nu)j. \end{aligned} \tag{3.5}$$

This system has three boundary equilibria, as follows:

1. The disease-free equilibrium

$$\mathcal{E}_0 = (1,0,0,0)\,.$$

   The disease-free equilibrium always exists.

2. The primary disease equilibrium exists if and only if $\mathcal{R}_1 > 1$. The steady distribution of infectives in the primary disease equilibrium is given by

$$i(a) = i(0)\pi_1(a)e^{-\mu a}, \qquad \text{where} \quad i(0) = \frac{\mu\left(1-\frac{1}{\mathcal{R}_1}\right)}{1-\Delta}.$$

   Thus, the equilibrium is

$$\mathcal{E}_1 = \left(\frac{1}{\mathcal{R}_1}, i(a), 0, 0\right).$$

3. The secondary disease equilibrium exists if and only if $\mathcal{R}_2 > 1$ and is given by

$$\mathcal{E}_2 = \left(\frac{1}{\mathcal{R}_2}, 0, \left(1-\frac{1}{\mathcal{R}_2}\right), 0\right).$$

Notice that the values of the two dominance equilibria do not depend on the coinfection. These exact same equilibria are present even if $\delta(a) = 0$.

We introduce the invasion reproduction numbers for each of the diseases. The invasion reproduction number of the first disease measures the ability of the primary disease to invade an equilibrium of the secondary disease. We define the invasion reproduction number of the primary disease as

$$\hat{\mathcal{R}}_1 = \frac{1}{\mathcal{R}_2}B(\hat{i}_2) + \frac{\gamma_1}{\mu+\nu}\frac{1}{\mathcal{R}_2}\left(1-\frac{1}{\mathcal{R}_2}\right)D(\hat{i}_2), \quad \text{where } \hat{i}_2 = \left(1-\frac{1}{\mathcal{R}_2}\right). \tag{3.6}$$

The invasion reproduction number of the secondary disease measures its ability to invade an equilibrium of the primary disease, and it is defined as

$$\hat{\mathcal{R}}_2 = \frac{(\mathcal{R}_1 - 1)\mu\gamma_2 D(0)}{\mathcal{R}_1(\mu+\alpha_2)(\mu+\nu)(\mathcal{R}_1 - \mathcal{R}_2)(1-\Delta)} \qquad \text{if } \mathcal{R}_1 > \mathcal{R}_2. \tag{3.7}$$

It is important to point out that, due to the asymmetry of the model, $\hat{\mathcal{R}}_1$ is defined if $\mathcal{R}_2 > 1$, and $\hat{\mathcal{R}}_2$ is defined if $\mathcal{R}_1 > \max(1, \mathcal{R}_2)$. In addition, it is possible that $\hat{\mathcal{R}}_1 > 1$ even if $\mathcal{R}_1 < 1$; that is, the dominance equilibrium $\mathcal{E}_1$ of the primary disease does not exist, and yet the primary disease can invade the dominance equilibrium of the secondary disease. It is also possible that $\hat{\mathcal{R}}_2 > 1$ even if $\mathcal{R}_2 < 1$; that is, the dominance equilibrium $\mathcal{E}_2$ of the secondary disease does not exist, but the secondary disease can invade the dominance equilibrium of the primary disease.

LEMMA 3.1. *The curves* $\mathcal{C}_1 = \{(\mathcal{R}_1, \mathcal{R}_2) | \hat{\mathcal{R}}_1 = 1\}$ *and* $\mathcal{C}_2 = \{(\mathcal{R}_1, \mathcal{R}_2) | \hat{\mathcal{R}}_2 = 1\}$ *enclose a nontrivial region in the positive* $(\mathcal{R}_1, \mathcal{R}_2)$ *quadrant. The interior of this region always contains an unbounded component given by inequalities* $\hat{\mathcal{R}}_1, \hat{\mathcal{R}}_2 < 1$.

Proof. Using the fact that $B(0) = \mathcal{R}_1$, the curve $\mathcal{C}_1 = \{(\mathcal{R}_1, \mathcal{R}_2) | \hat{\mathcal{R}}_1 = 1\}$ is given by the graph

$$\mathcal{R}_1 = \frac{B(0)}{B(\hat{i}_2)} \left(\mathcal{R}_2 - \frac{\gamma_1 \hat{i}_2 D(\hat{i}_2)}{\mu + \nu} \right) =: \mathcal{F}_1(\mathcal{R}_2),$$

where $\mathcal{R}_2 \geq 1$ and $\hat{i}_2 = 1 - 1/\mathcal{R}_2 \geq 0$. The definition of $\mathcal{F}_1$ implies that $\mathcal{F}_1(1) = 1$, and for large values of $\mathcal{R}_2$ the function

$$\mathcal{F}_1(\mathcal{R}_2) \approx \frac{B(0)}{B(1)} \mathcal{R}_2 + \frac{B(0)B'(1)}{B^2(1)} - \frac{\gamma_1 B(0) D(1)}{(\mu+\nu)B(1)} \qquad \text{for } \mathcal{R}_2 \gg 1$$

is approximately linear in $\mathcal{R}_2$ with a slope $B(0)/B(1) > 1$. On the other hand, the curve $\mathcal{C}_2 = \{(\mathcal{R}_1, \mathcal{R}_2) | \hat{\mathcal{R}}_2 = 1\}$ is given by the graph

$$\mathcal{R}_2 = \mathcal{R}_1 - \left(1 - \frac{1}{\mathcal{R}_1}\right) \frac{\mu\gamma_2 D(0)}{(\mu+\alpha_2)(\mu+\nu)(1-\Delta)} =: \mathcal{F}_2(\mathcal{R}_1),$$

where $\mathcal{R}_1 \geq 1$. It is easy to see that $\mathcal{F}_2(1) = 1$, and for large values of $\mathcal{R}_1$ the function

$$\mathcal{F}_2(\mathcal{R}_1) \approx \mathcal{R}_1 - \frac{\mu\gamma_2 D(0)}{(\mu+\alpha_2)(\mu+\nu)(1-\Delta)} \qquad \text{for } \mathcal{R}_1 \gg 1$$

is approximately linear in $\mathcal{R}_1$ with a unit slope. Consequently, when both $\mathcal{R}_1$ and $\mathcal{R}_2$ are large, the curve $\mathcal{C}_1$ lies below and to the right of the curve $\mathcal{C}_2$. The unbounded region enclosed by these curves is therefore given by the inequalities $\mathcal{R}_2 < \mathcal{F}_2(\mathcal{R}_1)$ and $\mathcal{R}_1 < \mathcal{F}_1(\mathcal{R}_2)$, which are equivalent to the inequalities $\hat{\mathcal{R}}_1, \hat{\mathcal{R}}_2 < 1$. ∎

In Figure 3.1, we present a simple diagram depicting the curves $\mathcal{C}_1$ and $\mathcal{C}_2$ and identify various parts of the region enclosed by these curves. The following theorem establishes the existence of at least one coexistence equilibrium for any point in the $(\mathcal{R}_1, \mathcal{R}_2)$ plane that lies between the curves $\mathcal{C}_1$ and $\mathcal{C}_2$. In what follows, we will refer to the region between the curves $\mathcal{C}_1$ and $\mathcal{C}_2$ as the *coexistence region*.

THEOREM 3.2. *Let*

$$\begin{aligned}
\mathcal{D}_- &= \{\mathcal{R}_1, \mathcal{R}_2 > 0 | 1 < \mathcal{R}_1 \leq \mathcal{R}_2, \hat{\mathcal{R}}_1 > 1\},\\
\mathcal{D}_+ &= \{\mathcal{R}_1, \mathcal{R}_2 > 0 | 1 < \mathcal{R}_2 < \mathcal{R}_1, \hat{\mathcal{R}}_1 > 1, \hat{\mathcal{R}}_2 > 1\},\\
\mathcal{D}_1 &= \{\mathcal{R}_1, \mathcal{R}_2 > 0 | \mathcal{R}_2 < 1 < \mathcal{R}_1, \hat{\mathcal{R}}_2 > 1\},\\
\mathcal{D}_2 &= \{\mathcal{R}_1, \mathcal{R}_2 > 0 | \mathcal{R}_1 < 1 < \mathcal{R}_2, \hat{\mathcal{R}}_1 > 1\},\\
\mathcal{D}_3 &= \{\mathcal{R}_1, \mathcal{R}_2 > 0 | 1 < \mathcal{R}_2 < \mathcal{R}_1, \hat{\mathcal{R}}_1 < 1, \hat{\mathcal{R}}_2 < 1\};
\end{aligned}$$

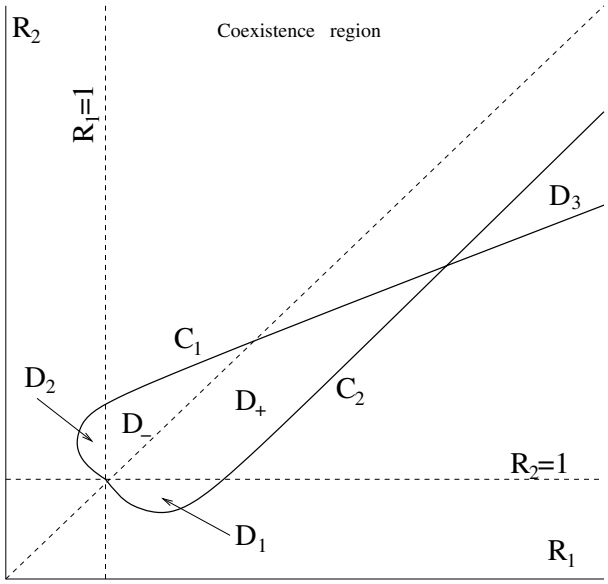


FIG. 3.1. *A schematic diagram representing the coexistence region in the* $(\mathcal{R}_1, \mathcal{R}_2)$ *plane. The boundary of the coexistence region is formed by the curves* $\mathcal{C}_1 : \hat{\mathcal{R}}_1 = 1$ *and* $\mathcal{C}_2 : \hat{\mathcal{R}}_2 = 1$. *The coexistence region (as shown) consists of the following five components:* D_-, *where* $1 < \mathcal{R}_1 < \mathcal{R}_2$ *and* $\hat{\mathcal{R}}_1 > 1$; D_+, *where* $1 < \mathcal{R}_2 < \mathcal{R}_1$ *and* $\hat{\mathcal{R}}_1$; $\hat{\mathcal{R}}_2 > 1$; D_1, *where* $\mathcal{R}_2 < 1 < \mathcal{R}_1$ *and* $\hat{\mathcal{R}}_2 > 1$; D_2, *where* $\mathcal{R}_1 < 1 < \mathcal{R}_2$ *and* $\hat{\mathcal{R}}_1 > 1$; *and finally,* D_3, *where* $1 < \mathcal{R}_2 < \mathcal{R}_1$ *and* $\hat{\mathcal{R}}_1, \hat{\mathcal{R}}_2 < 1$. D_3 *is the only component of the coexistence region that is always nonempty; the other components may or may not exist, depending on the parameter values.*

then for any $(\mathcal{R}_1, \mathcal{R}_2)$ *in the coexistence region* $\mathcal{D}_c = \mathcal{D}_- \cup \mathcal{D}_+ \cup \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3$ *there exists at least one coexistence equilibrium for the two diseases.*

Proof. The fourth equation in (3.5) implies that at a coexistence equilibrium,

$$j = \frac{i(0) i_2 D(i_2)}{\mu + \nu}.$$

Substituting this expression into the second equation in (3.5), we find that

$$s = \left(B(i_2) + \frac{\gamma_1 i_2 D(i_2)}{\mu + \nu} \right)^{-1} =: \mathcal{S}(i_2). \tag{3.8}$$

Now we substitute (3.8) into the third equation in (3.5) and solve for j to obtain the expressions

$$j = \frac{(\mu + \alpha_2)(1 - \mathcal{R}_2 \mathcal{S}(i_2)) i_2}{\gamma_2 \mathcal{S}(i_2)}, \quad \mathcal{R}_2 = \frac{\beta_2}{\mu + \alpha_2},$$

and

$$i(0) = \frac{(\mu + \nu) j}{i_2 D(i_2)} = \frac{(\mu + \nu)(\mu + \alpha_2)(1 - \mathcal{R}_2 \mathcal{S}(i_2))}{\gamma_2 \mathcal{S}(i_2) D(i_2)}.$$

Finally, we express j and i_1 as follows:

$$j = \frac{i_2 (\mu + \alpha_2)(1 - \mathcal{R}_2 \mathcal{S}(i_2))}{\gamma_2 \mathcal{S}(i_2)} =: \mathcal{J}(i_2) \tag{3.9}$$

and

(3.10)
$$i_1 = \int_0^\infty i(a)\, da = i(0)G(i_2) = \frac{(\mu+\nu)(\mu+\alpha_2)G(i_2)(1-\mathcal{R}_2\mathcal{S}(i_2))}{\gamma_2\mathcal{S}(i_2)D(i_2)} =: \mathcal{I}(i_2),$$

where

$$G(i_2) = \int_0^\infty \Gamma(a;i_2)\pi_1(a)e^{-\mu a}da > 0, \quad G(0) = \frac{1-\Delta}{\mu}.$$

Since we are working with rescaled variables, the relation $s + i_1 + i_2 + j = 1$ implies that

$$\mathcal{M}(i_2) := i_2 + \mathcal{S}(i_2) + \mathcal{I}(i_2) + \mathcal{J}(i_2) = 1.$$

We note that the function $\mathcal{S}(i_2)$ is positive for all $i_2 \geq 0$, and both functions $\mathcal{I}(i_2)$, $\mathcal{J}(i_2)$ are positive if $i_2 > 0$ and $\mathcal{R}_2\mathcal{S}(i_2) < 1$. To prove the existence of a coexistence equilibrium it suffices to show the existence of a positive root of the equation $\mathcal{M}(i_2) = 1$ that satisfies $\mathcal{R}_2\mathcal{S}(i_2) < 1$. We also note that the function $\mathcal{M}(i_2)$ can be equivalently expressed as

$$\mathcal{M}(i_2) := i_2 + \mathcal{S}(i_2) + \frac{(\mu+\alpha_2)(1-\mathcal{R}_2\mathcal{S}(i_2))}{\gamma_2\mathcal{S}(i_2)}\left(i_2 + \frac{G(i_2)(\mu+\nu)}{D(i_2)}\right).$$

We observe that $\mathcal{S}(0) = 1/\mathcal{R}_1$, and therefore

$$\begin{aligned}\mathcal{M}(0) &= \frac{1}{\mathcal{R}_1} + \frac{(\mu+\nu)(\mu+\alpha_2)G(0)(1-\mathcal{R}_2\mathcal{S}(0))}{\gamma_2\mathcal{S}(0)D(0)} \\ &= \frac{1}{\mathcal{R}_1} + \frac{(1-\Delta)(\mu+\nu)(\mu+\alpha_2)(\mathcal{R}_1-\mathcal{R}_2)}{\mu\gamma_2 D(0)}.\end{aligned}$$

Suppose that $(\mathcal{R}_1, \mathcal{R}_2) \in \mathcal{D}_-$, that is, $1 < \mathcal{R}_1 \leq \mathcal{R}_2$ and $\hat{\mathcal{R}}_1 > 1$. It follows that $\mathcal{M}(0) \leq 1/\mathcal{R}_1 < 1$ because $\mathcal{R}_1 \leq \mathcal{R}_2$. Using the definition of $\hat{\mathcal{R}}_1$, we find that $\hat{\mathcal{R}}_1 > 1$ implies that $\mathcal{R}_2\mathcal{S}(\hat{i}_2) < 1$. Therefore, there exist the following three possibilities:

1. If $\mathcal{M}(\hat{i}_2) = 1$, then we are done because both $\mathcal{I}(\hat{i}_2)$ and $\mathcal{J}(\hat{i}_2)$ are positive.
2. If $\mathcal{M}(\hat{i}_2) > 1$, then one of the following holds:
   - There exists $i_2^* \in [0, \hat{i}_2)$ such that $\mathcal{R}_2\mathcal{S}(i_2^*) = 1$ and $\mathcal{R}_2\mathcal{S}(i_2) < 1$ for all $i_2 \in (i_2^*, \hat{i}_2]$. In this case, we have that
   $$\mathcal{M}(i_2^*) = i_2^* + \frac{1}{\mathcal{R}_2} = 1 + i_2^* - \hat{i}_2 < 1,$$
   and since $\mathcal{M}(\hat{i}_2) > 1$, there exists a number $i_2 \in (i_2^*, \hat{i}_2)$ such that $\mathcal{M}(i_2) = 1$, where both $\mathcal{I}(i_2)$ and $\mathcal{J}(i_2)$ are positive.
   - $\mathcal{R}_2\mathcal{S}(i_2) < 1$ for all $i_2 \in (0, \hat{i}_2]$. Then since $\mathcal{M}(0) < 1$ and $\mathcal{M}(\hat{i}_2) > 1$, there exists a number $i_2 \in (0, \hat{i}_2)$ such that $\mathcal{M}(i_2) = 1$, where both $\mathcal{I}(i_2)$ and $\mathcal{J}(i_2)$ are positive.
3. If $\mathcal{M}(\hat{i}_2) < 1$, then one of the following holds:
   - There exists $i_2^* \in (\hat{i}_2, 1]$ such that $\mathcal{R}_2\mathcal{S}(i_2^*) = 1$ and $\mathcal{R}_2\mathcal{S}(i_2) < 1$ for all $i_2 \in (\hat{i}_2, i_2^*)$. In this case, we have that
   $$\mathcal{M}(i_2^*) = i_2^* + \frac{1}{\mathcal{R}_2} = 1 + i_2^* - \hat{i}_2 > 1,$$
   and since $\mathcal{M}(\hat{i}_2) < 1$, there exists a number $i_2 \in (\hat{i}_2, i_2^*)$ such that $\mathcal{M}(i_2) = 1$, where both $\mathcal{I}(i_2)$ and $\mathcal{J}(i_2)$ are positive.

- $\mathcal{R}_2\mathcal{S}(i_2) < 1$ for all $i_2 \in [\hat{i}_2, 1)$. Then we have that $\mathcal{M}(1) \geq 1 + \mathcal{S}(1) > 1$, and since $\mathcal{M}(\hat{i}_2) < 1$, there exists a number $i_2 \in (\hat{i}_2, 1)$ such that $\mathcal{M}(i_2) = 1$, where both $\mathcal{I}(i_2)$ and $\mathcal{J}(i_2)$ are positive.

Therefore, there exists a coexistence equilibrium for all $(\mathcal{R}_1, \mathcal{R}_2) \in \mathcal{D}_-$.

Now suppose that $(\mathcal{R}_1, \mathcal{R}_2) \in \mathcal{D}_+$, that is, $1 < \mathcal{R}_2 < \mathcal{R}_1$ and $\hat{\mathcal{R}}_1, \hat{\mathcal{R}}_2 > 1$. The inequality $\hat{\mathcal{R}}_2 > 1$ implies that $\mathcal{M}(0) < 1$, and the inequality $\hat{\mathcal{R}}_1 > 1$ implies that $\mathcal{R}_2\mathcal{S}(\hat{i}_2) < 1$. From this point forward, the proof of this case is analogous to the proof of the case $(\mathcal{R}_1, \mathcal{R}_2) \in \mathcal{D}_-$.

Suppose that $(\mathcal{R}_1, \mathcal{R}_2) \in \mathcal{D}_2$, that is, $\mathcal{R}_1 < 1 < \mathcal{R}_2$ and $\hat{\mathcal{R}}_1 > 1$. As before, the inequality $\hat{\mathcal{R}}_1 > 1$ implies that $\mathcal{R}_2\mathcal{S}(\hat{i}_2) < 1$. On the other hand, we have that $\mathcal{R}_2\mathcal{S}(0) = \mathcal{R}_2/\mathcal{R}_1 > 1$. Now, if $\mathcal{M}(\hat{i}_2) \leq 1$, the proof is analogous to the proof of the case $(\mathcal{R}_1, \mathcal{R}_2) \in \mathcal{D}_-$. If $\mathcal{M}(\hat{i}_2) > 1$, then there exists $i_2^* \in (0, \hat{i}_2)$ such that $\mathcal{R}_2\mathcal{S}(i_2^*) = 1$ and $\mathcal{R}_2\mathcal{S}(i_2) < 1$ for all $i_2 \in (i_2^*, \hat{i}_2]$. In addition, we have that $\mathcal{M}(i_2^*) = i_2^* + 1/\mathcal{R}_2 = 1 + i_2^* - \hat{i}_2 < 1$. Consequently, there exists a number $i_2 \in (i_2^*, \hat{i}_2)$ such that $\mathcal{M}(i_2) = 1$. This concludes the proof of the case $(\mathcal{R}_1, \mathcal{R}_2) \in D_2$.

Suppose that $(\mathcal{R}_1, \mathcal{R}_2) \in \mathcal{D}_1$, that is, $\mathcal{R}_2 < 1 < \mathcal{R}_1$ and $\hat{\mathcal{R}}_2 > 1$. As before, the inequality $\hat{\mathcal{R}}_2 > 1$ implies that $\mathcal{M}(0) < 1$, but the value of $\hat{i}_2 = 1 - 1/\mathcal{R}_2 < 0$ since $\mathcal{R}_2 < 1$. Instead, we have that $\mathcal{R}_2\mathcal{S}(0) = \mathcal{R}_2/\mathcal{R}_1 < 1$. Suppose that there exists $i_2^* \in (0, 1)$ such that $\mathcal{R}_2\mathcal{S}(i_2^*) = 1$ and $\mathcal{R}_2\mathcal{S}(i_2) < 1$ for all $i_2 \in (0, i_2^*)$. Then we have that $\mathcal{M}(i_2^*) = i_2^* + 1/\mathcal{R}_2 > 1/\mathcal{R}_2 > 1$ and there exists a number $i_2 \in (0, i_2^*)$ such that $\mathcal{M}(i_2) = 1$. If no such i_2^* exists, we have that $\mathcal{R}_2\mathcal{S}(i_2) < 1$ for all $i_2 \in (0, 1)$. Since $\mathcal{M}(1) \geq 1 + \mathcal{S}(1) > 1$, there exists a number $i_2 \in (0, 1)$ such that $\mathcal{M}(i_2) = 1$. This concludes the proof of the case $(\mathcal{R}_1, \mathcal{R}_2) \in \mathcal{D}_1$.

Finally, suppose that $(\mathcal{R}_1, \mathcal{R}_2) \in \mathcal{D}_3$, that is, $1 < \mathcal{R}_2 < \mathcal{R}_1$ and $\hat{\mathcal{R}}_1, \hat{\mathcal{R}}_2 < 1$. The inequality $\hat{\mathcal{R}}_2 < 1$ implies that $\mathcal{M}(0) > 1$, and the inequality $\hat{\mathcal{R}}_1 < 1$ implies that $\mathcal{R}_2\mathcal{S}(\hat{i}_2) > 1$. In addition, since $\mathcal{R}_1 > \mathcal{R}_2$, we have that $\mathcal{R}_2\mathcal{S}(0) < \mathcal{R}_1\mathcal{S}(0) = 1$. Therefore, there exists $i_2^* \in (0, \hat{i}_2)$ such that $\mathcal{R}_2\mathcal{S}(i_2^*) = 1$ and $\mathcal{R}_2\mathcal{S}(i_2) < 1$ for all $i_2 \in [0, i_2^*)$. Since $\mathcal{M}(i_2^*) = 1 + i_2^* - \hat{i}_2 < 1$, there exists a number $i_2 \in (0, i_2^*)$ such that $\mathcal{M}(i_2) = 1$. This concludes the proof of the theorem. ∎

Remark. Each of the subregions comprising the coexistence region $\mathcal{D}_c$ has a clear epidemiological interpretation. These regions are presented in Figure 3.1.

If $(\mathcal{R}_1, \mathcal{R}_2) \in \mathcal{D}_- \cup \mathcal{D}_+$, then both dominance equilibria $\mathcal{E}_1$ and $\mathcal{E}_2$ exist, and each disease can invade the equilibrium of the other disease. The difference between $\mathcal{D}_-$ and $\mathcal{D}_+$ is that $\hat{\mathcal{R}}_2$ is defined only for $(\mathcal{R}_1, \mathcal{R}_2) \in \mathcal{D}_+$.

If $(\mathcal{R}_1, \mathcal{R}_2) \in \mathcal{D}_1$, then only the dominance equilibrium of the primary disease $\mathcal{E}_1$ exists, and the secondary disease can invade the equilibrium of the primary disease. Although the secondary disease cannot persist in the absence of the primary disease, the presence of the primary disease mediates the coexistence.

If $(\mathcal{R}_1, \mathcal{R}_2) \in \mathcal{D}_2$, then only the dominance equilibrium of the secondary disease $\mathcal{E}_2$ exists, and the primary disease can invade the equilibrium of the secondary disease. Although the primary disease cannot persist in the absence of the secondary disease, the presence of the secondary disease mediates the coexistence.

If $(\mathcal{R}_1, \mathcal{R}_2) \in \mathcal{D}_3$, then both dominance equilibria $\mathcal{E}_1$ and $\mathcal{E}_2$ exist, but neither disease can invade the equilibrium of the other disease.

4. Extinction of one or both diseases. In this section we provide the conditions that guarantee that one of the diseases or both of them will be eliminated from the population. These are global conditions in the sense that if they are satisfied, extinction occurs for all other values of the parameters and all initial conditions. As we show in section 6, there could be a backward bifurcation with respect to both $\mathcal{R}_1$

and $\mathcal{R}_2$, and therefore there could exist multiple coexistence equilibria even if $\mathcal{R}_1 < 1$ and $\mathcal{R}_2 < 1$. Thus, $\mathcal{R}_1 < 1$ and $\mathcal{R}_2 < 1$ by themselves do not necessarily imply extinction of one or both diseases. In what follows, we show that the diseases vanish if, in addition, $\gamma_1 = 0$ or $\gamma_2 = 0$.

We denote the initial conditions by $S(0) = S^0$, $i(a,0) = i_0(a)$, $I_2(0) = I_2^0$, and $J(0) = J^0$.

THEOREM 4.1. *Assume that $i_0(a)$ is integrable. If $\gamma_1 = 0$ or $\gamma_2 = 0$ and $\mathcal{R}_1 < 1$, $\mathcal{R}_2 < 1$, then both diseases become extinct in the sense that $\lim_{t\to\infty} i(a,t) = 0$ pointwise for every a, $I_2 \to 0$ as $t \to \infty$, and $J \to 0$ as $t \to \infty$.*

Proof. Assume $\gamma_1 = 0$. Let $\mathcal{B}(t) = i(0,t)$. Neglecting the term dependent on I_2, we obtain a differential inequality for the primary disease. Integrating this inequality along the characteristic lines, we have

$$i(a,t) \le \begin{cases} i_0(a-t)\dfrac{\pi(a)}{\pi(a-t)}e^{-\mu t}, & a \ge t, \\ \mathcal{B}(t-a)\pi(a)e^{-\mu a}, & a < t. \end{cases} \tag{4.1}$$

Since $\gamma_1 = 0$ we have

$$\mathcal{B}(t) \le \int_0^t \beta_1(a)\mathcal{B}(t-a)\pi_1(a)e^{-\mu a}da + e^{-\mu t}\int_t^\infty \beta_1(a)i_0(a-t)da.$$

Consequently, taking a limsup of both sides as $t \to \infty$, we obtain $\limsup_{t\to\infty} \mathcal{B} \le \mathcal{R}_1 \limsup_{t\to\infty} \mathcal{B}(t)$. Since $\mathcal{R}_1 < 1$ and $\limsup_{t\to\infty} \mathcal{B} < \infty$, this inequality can be satisfied only if $\limsup_{t\to\infty} \mathcal{B}(t) = 0$. This, in particular, implies that $i(a,t)$ approaches zero as $t \to \infty$ for every fixed a. From the equation for J we then have the following inequality:

$$J(t) \le e^{-(\mu+\nu)t}J_0 + \int_0^t e^{-(\mu+\nu)s}\int_0^\infty \delta(a)i(a,t-s)dads.$$

Since $\delta(a)$ is bounded and the integral of $i(a,t)$ goes to zero, $I_1(t) \to 0$ as $t \to \infty$, we get that $\limsup_{t\to\infty} J(t) = 0$. Consequently, the equality for I_2 in (2.1) leads to the following differential inequality: $I_2' \le \beta_2 I_2 + \gamma_2 J(t) - (\mu + \alpha_2)I_2$. Integrating this inequality, we obtain

$$I_2(t) \le e^{-(\mu+\alpha_2)t}I_2(0) + \beta_2\int_0^t e^{-(\mu+\alpha_2)\tau}I_2(t-\tau)d\tau + \gamma_2\int_0^t e^{-(\mu+\alpha_2)\tau}J(t-\tau)d\tau.$$

Taking a limsup as $t \to \infty$ on both sides of this inequality, we obtain $\limsup_{t\to\infty} I_2(t) \le \mathcal{R}_2 \limsup_{t\to\infty} I_2(t)$. Since $\mathcal{R}_2 < 1$, this inequality implies $\limsup_{t\to\infty} I_2(t) = 0$.

If $\gamma_2 = 0$, then the proof is symmetrical and somewhat analogous. Thus, it will be omitted. That concludes the proof of this theorem. □

As a special case of the theorem above, we have the following results on extinction of one of the diseases.

COROLLARY 4.2. *Assume that $i_0(a)$ is integrable. If $\gamma_1 = 0$ and $\mathcal{R}_1 < 1$, then the primary disease becomes extinct in the sense that $\lim_{t\to\infty} i(a,t) = 0$ pointwise for every a. As a consequence, $J(t) \to 0$ as $t \to \infty$.*

A similar result for the secondary disease is also valid.

COROLLARY 4.3. *If $\gamma_2 = 0$ and $\mathcal{R}_2 < 1$, then the secondary disease becomes extinct; that is, $I_2(t) \to 0$ as $t \to \infty$. As a consequence, $J(t) \to 0$ as $t \to \infty$.*

A special instance which deserves consideration is the one with $\delta(a) = 0$. In this case there is no coinfection, and the jointly infected class J vanishes. Only three equilibria are possible—the coexistence equilibrium does not exist. The main question is: does the competitive exclusion principle hold for the two diseases with no coinfection? This question can be answered positively in the case when all coefficients are constant and the model (2.1) consists of ODEs only. Then, with $\delta = 0$, it becomes a particular case of a more general model considered in [4]. The results there imply that the competitive exclusion holds and that only the disease with higher reproduction number persists in the population; the other one becomes extinct.

We have not been able to establish whether competitive exclusion for the model (2.1) with $\delta(a) = 0$ is the only possible outcome in the strictly age-structured case. Although there is no coexistence equilibrium, coexistence might still be possible in the form of, say, a stable oscillatory solution. Such a situation has been found to occur in model ecosystems such as the chemostat [3, 5, 18]. This option is even more plausible here, given that the dominance equilibrium of the primary disease can lose stability due to the age-structure, and oscillatory solutions are present (see section 5.2 for more detailed discussion). Despite the oscillatory solutions, simulations lead to extinction of the disease with lower reproduction number. Thus, we conjecture that competitive exclusion is still the norm. A rigorous justification, however, remains an open problem.

5. Local stability of equilibria. In this section we investigate the local stability of the equilibria. In particular, we derive conditions for the stability of the disease-free equilibrium and of the secondary disease dominance equilibrium. We also show that Hopf bifurcation occurs in the coexistence equilibrium. The stability of equilibria determines conditions under which the ultimate outcome will be elimination of both diseases, dominance of the primary disease, dominance of the secondary disease, or endemic presence of both of them.

To investigate the stability of the equilibria, we linearize the model (2.1). In particular, let $x(t)$, $y(a,t)$, $z(t)$, and $w(t)$ be the perturbations of, respectively, S^*, $i^*(a)$, I_2^*, and J^*. That is, $S = S^* + x$, $i = i^* + y$, $I_2 = I_2^* + z$, $J = J^* + w$. Thus the perturbations satisfy a linear system. Further, we consider the eigenvalue problem for the linearized system. We will denote the eigenvector again with x, $y(a)$, z, and w. These satisfy the following linear eigenvalue problem (here s, i, i_2, and j are the proportions in the corresponding equilibrium):

$$
\begin{aligned}
\lambda x &= -s\int_0^\infty \beta_1(a)y(a)da - xi(0)B(i_2) - \beta_2 sz - \beta_2 x i_2 + \int_0^\infty \alpha_1(a)y(a)da \\
&\qquad - (\gamma_1+\gamma_2)sw - (\gamma_1+\gamma_2)xj - \mu x + \alpha_2 z, \\
y'(a) &= -\lambda y - \alpha_1(a)y - \delta(a)i_2 y - \delta(a)i(a)z - \mu y, \\
y(0) &= s\int_0^\infty \beta_1(a)y(a)da + xi(0)B(i_2) + \gamma_1 sw + \gamma_1 xj, \\
\lambda z &= \beta_2 sz + \beta_2 i_2 x + \gamma_2 sw + \gamma_2 jx - (\mu+\alpha_2)z, \\
\lambda w &= i_2\int_0^\infty \delta(a)y(a)da + zi(0)D(i_2) - (\mu+\nu)w.
\end{aligned}
\tag{5.1}
$$

5.1. Stability of the disease-free equilibrium. For the disease-free equilibrium we have $i(0) = 0$, $i_2 = 0$, $j = 0$, and $s = 1$. Thus the system above simplifies to the following system:

(5.2)
$$\begin{aligned}
&\lambda x = -\int_0^\infty \beta_1(a)y(a)da - \beta_2 z - (\gamma_1+\gamma_2)w - \mu x + \int_0^\infty \alpha_1(a)y(a)da + \alpha_2 z,\\
&y'(a) = -\lambda y - \alpha_1(a)y - \mu y,\\
&y(0) = \int_0^\infty \beta_1(a)y(a)da + \gamma_1 w,\\
&\lambda z = \beta_2 z + \gamma_2 w - (\mu+\alpha_2)z,\\
&\lambda w = -(\mu+\nu)w.
\end{aligned}$$

From this system we will establish the following result regarding the local stability of the disease-free equilibrium $\mathcal{E}_0$.

PROPOSITION 5.1. *If $\mathcal{R}_1 < 1$ and $\mathcal{R}_2 < 1$, then the disease-free equilibrium $\mathcal{E}_0$ is locally asymptotically stable. If $\mathcal{R}_1 > 1$ or $\mathcal{R}_2 > 1$, then the disease-free equilibrium $\mathcal{E}_0$ is unstable.*

Proof. To see this, first notice that from the last equation we have either $\lambda = -(\mu+\nu)$, which is the first eigenvalue, or $w = 0$. From the second-to-last equation we have $\lambda z = \beta_2 z - (\mu+\alpha_2)z$, where either $\lambda = \beta_2 - (\mu+\alpha_2)$ or $z = 0$. This eigenvalue $\lambda = \beta_2 - (\mu+\alpha_2) < 0$ if and only if $\mathcal{R}_2 < 1$. Thus, if $\mathcal{R}_2 > 1$, the disease-free equilibrium $\mathcal{E}_0$ is unstable because this eigenvalue is positive. Further, from the second equation we have that the remaining eigenvalues satisfy the equation, also referred to as the characteristic equation,

$$\int_0^\infty \beta_1(a)e^{-(\lambda+\mu)a}\pi_1(a)da = 1. \tag{5.3}$$

Denoting the left-hand side of the equation above by $\mathcal{G}(\lambda)$, where λ is in general a complex number, assume $\Re\lambda \geq 0$. For such λ we have $|\mathcal{G}(\lambda)| \leq \mathcal{G}(\Re\lambda)$. Furthermore, $\mathcal{G}(\Re\lambda)$ is a decreasing function of $\Re\lambda$. Consequently,

$$|\mathcal{G}(\lambda)| \leq \mathcal{G}(\Re\lambda) \leq \mathcal{G}(0) = \mathcal{R}_1 < 1.$$

Thus, if both $\mathcal{R}_1 < 1$ and $\mathcal{R}_2 < 1$, all eigenvalues have negative real part, and the disease-free equilibrium $\mathcal{E}_0$ is locally asymptotically stable. If only $\mathcal{R}_1 > 1$, then if we consider $\mathcal{G}(\lambda)$ for λ real, we see that $\mathcal{G}(\lambda)$ is a decreasing function of λ approaching zero as λ approaches infinity. Since $\mathcal{G}(0) = \mathcal{R}_1 > 1$, that implies that there is a positive eigenvalue $\lambda^* > 0$, and the disease-free equilibrium $\mathcal{E}_0$ is unstable. This concludes the proof. ∎

5.2. Stability of the primary disease equilibrium. In this subsection we discuss the local stability of the equilibrium $\mathcal{E}_1$ and derive conditions for dominance of the primary disease. We show that the equilibrium $\mathcal{E}_1$ can lose stability, and dominance of the first disease is possible in the form of sustained oscillation. In this case $i_2 = 0$, $j = 0$, $s = \frac{1}{\mathcal{R}_1}$, and $i(a) = i(0)\pi_1(a)e^{-\mu a}$, where

$$i(0) = \frac{\mu\left(1 - \frac{1}{\mathcal{R}_1}\right)}{1-\Delta}. \tag{5.4}$$

The eigenvalue problem takes the form

$$
\begin{aligned}
\lambda x &= -s\int_0^\infty \beta_1(a)y(a)da - xi(0)B(0) - \beta_2 s\, z \\
&\quad -(\gamma_1+\gamma_2)sw - \mu x + \int_0^\infty \alpha_1(a)y(a)da + \alpha_2 z, \\
y'(a) &= -\lambda y - \alpha_1(a)y - \delta(a)i(a)z - \mu y, \\
y(0) &= s\int_0^\infty \beta_1(a)y(a)da + xi(0)B(0) + \gamma_1 s\, w, \\
\lambda z &= \beta_2 sz + \gamma_2 sw - (\mu+\alpha_2)z, \\
\lambda w &= zi(0)D(0) - (\mu+\nu)w.
\end{aligned} \tag{5.5}
$$

From the last equation we have

$$w = \frac{zi(0)D(0)}{\lambda+\mu+\nu}.$$

Substituting in the equation for z, assuming z is nonzero, and canceling z, we arrive at the following characteristic equation:

$$\frac{\gamma_2 si(0)D(0)}{(\lambda+\mu+\nu)(\lambda+\mu+\alpha_2-\beta_2 s)} = 1. \tag{5.6}$$

We are now ready to establish the first result.

PROPOSITION 5.2. *Let $\mathcal{R}_1 > 1$ and $\mathcal{R}_1 > \mathcal{R}_2$. Then the equilibrium $\mathcal{E}_1$ is unstable if the secondary disease can invade the equilibrium of the primary disease, that is, $\hat{\mathcal{R}}_2 > 1$. If $\hat{\mathcal{R}}_2 < 1$, then all solutions to the characteristic equation* (5.6) *have negative real part.*

Proof. To see these results, denote by $\mathcal{G}(\lambda)$ the left-hand side of the characteristic equation (5.6). First, we notice that, using the values of s and $i(0)$, we have

$$\mathcal{G}(0) = \frac{\gamma_2 si(0)D(0)}{(\mu+\nu)(\mu+\alpha_2-\beta_2 s)} = \frac{\mu\gamma_2\left(1-\frac{1}{\mathcal{R}_1}\right)D(0)}{(1-\Delta)(\mu+\alpha_2)(\mu+\nu)(\mathcal{R}_1-\mathcal{R}_2)} = \hat{\mathcal{R}}_2. \tag{5.7}$$

First, in the case $\hat{\mathcal{R}}_2 > 1$ we have that $\mathcal{G}(0) > 1$. In addition, if $\mathcal{G}(\lambda)$ is considered as a function of a real variable, we see that $\mathcal{G}(\lambda) \to 0$ as $\lambda \to \infty$. Since $\mathcal{R}_1 > \mathcal{R}_2$, $\mathcal{G}(\lambda)$ is also a continuous function of λ for $\lambda \geq 0$. Consequently, there exists $\lambda^* > 0$ such that $\mathcal{G}(\lambda^*) = 1$. Thus, $\mathcal{E}_1$ is unstable.

In the case when $\hat{\mathcal{R}}_2 < 1$ we have for λ's with $\Re\lambda \geq 0$

$$
\begin{aligned}
|\mathcal{G}(\lambda)| &= \frac{\gamma_2 si(0)D(0)}{|\lambda+\mu+\nu||\lambda+\mu+\alpha_2-\beta_2 s|} \\
&\leq \frac{\gamma_2 si(0)D(0)}{(\Re\lambda+\mu+\nu)(\Re\lambda+\mu+\alpha_2-\beta_2 s)} \leq \mathcal{G}(0) = \hat{\mathcal{R}}_2 < 1.
\end{aligned}
$$

Consequently, the equation $\mathcal{G}(\lambda) = 1$ has no solutions with nonnegative real parts. This concludes the proof of the proposition. ∎

We note that the fact that all solutions to the characteristic equation (5.6) have negative real part does not yet imply that $\mathcal{E}_1$ is stable, since there is a second characteristic equation associated with this case. For stability both characteristic equations must have only roots with negative real parts.

Next, we extend the result above to the case $\mathcal{R}_1 < \mathcal{R}_2$. In particular, we have the following result.

PROPOSITION 5.3. *Let $\mathcal{R}_1 > 1$. If $\mathcal{R}_1 < \mathcal{R}_2$, then the equilibrium $\mathcal{E}_1$ is unstable.*

Proof. To see that, we rewrite the characteristic equation (5.6) in the form

$$\frac{\gamma_2 s i(0) D(0)}{\lambda + \mu + \nu} = \lambda + \mu + \alpha_2 - \beta_2 s. \tag{5.8}$$

We notice that $\mu + \alpha_2 - \beta_2 s = (\mu + \alpha_2)(1 - \frac{\mathcal{R}_2}{\mathcal{R}_1})$, which is negative. Let $\lambda^* = -(\mu + \alpha_2 - \beta_2 s) > 0$. Thus for $\lambda \geq \lambda^*$ the expression $\lambda + \mu + \alpha_2 - \beta_2 s$, considered as a function of the real variable λ, is increasing from zero to infinity. On the other hand, for $\lambda \geq \lambda^*$ the expression

$$\frac{\gamma_2 s i(0) D(0)}{\lambda + \mu + \nu}$$

is decreasing from some positive value to zero. Thus, there is a unique positive (actually larger than λ^*) solution of (5.8). Consequently, $\mathcal{E}_1$ is unstable. This completes the proof. ∎

We continue with our consideration of the system (5.5). If we assume that $z = 0$, that implies $w = 0$. In this case the remaining two equations become

$$\begin{aligned} &\lambda x = -s \int_0^\infty \beta_1(a) y(a) da - x i(0) B(0) - \mu x + \int_0^\infty \alpha_1(a) y(a) da, \\ &y'(a) = -\lambda y - \alpha_1(a) y - \mu y, \\ &y(0) = s \int_0^\infty \beta_1(a) y(a) da + x i(0) B(0). \end{aligned} \tag{5.9}$$

Solving the differential equation, substituting in the equation for x and the initial condition, we obtain a system in x and $y(0)$ which has a nontrivial solution if and only if the following characteristic equation is satisfied:

$$(\lambda + \mu) s B_2(\lambda) = \lambda + \mu + i(0) B(0)(1 - A_2(\lambda)), \tag{5.10}$$

where the following notation has been used:

$$B_2(\lambda) = \int_0^\infty \beta_1(a) \pi_1(a) e^{-(\lambda+\mu)a} da, \quad A_2(\lambda) = \int_0^\infty \alpha_1(a) \pi_1(a) e^{-(\lambda+\mu)a} da.$$

If we define

$$E_2(\lambda) = \int_0^\infty \pi_1(a) e^{-(\lambda+\mu)a} da,$$

we can notice that integration by parts leads to the equality $1 - A_2(\lambda) = (\lambda + \mu) E_2(\lambda)$. Consequently the characteristic equation (5.10) has one eigenvalue equal to $-\mu$. The remaining eigenvalues satisfy the following reduced characteristic equation:

$$s B_2(\lambda) = 1 + i(0) B(0) E_2(\lambda). \tag{5.11}$$

This equation clearly does not have real nonnegative solutions since for λ real and nonnegative the left-hand side is smaller than one, while the right-hand side is larger than one. However, the dominant eigenvalue is not necessarily real—it may be complex

with nonnegative real part. Thus, the dominance equilibrium of the primary disease may lose stability, and oscillations are possible. We include an example and results of simulations later in this section. First, we show that the mechanism responsible for the instability of the primary disease equilibrium is the presence of infection-age structure and variable infectivity. Indeed, if $\beta_1(a) = \beta_1$ and $\alpha_1(a) = \alpha_1$ are constants, then $i(0)B(0) = \beta_1 i$, where i is the proportion infected with primary disease. In addition, $B_2(\lambda) = \beta_1 E_2(\lambda)$ and $E_2(\lambda) = (\lambda + \mu + \alpha_1)^{-1}$. Hence, in the constant coefficient case the characteristic equation (5.11) becomes $\lambda + \mu + \alpha_1 + \beta_1 i - \beta_1 s = 0$. Since $\beta_1 s = \mu + \alpha_1$, the only eigenvalue is $-\beta_1 i$ and is clearly negative. We formulate this result in the following proposition.

PROPOSITION 5.4. *Let $\beta_1(a) = \beta_1$ and $\alpha_1(a) = \alpha_1$ be constants. Let $\mathcal{R}_1 > 1$. Assume that $\mathcal{R}_1 > \mathcal{R}_2$ and that the secondary disease cannot invade the equilibrium of the primary disease; that is, $\hat{\mathcal{R}}_2 < 1$. Then the equilibrium $\mathcal{E}_1$ is locally asymptotically stable. If $\hat{\mathcal{R}}_2 > 1$, the equilibrium $\mathcal{E}_1$ is unstable.*

We conclude this section with an example that the presence of infection-age structure may lead to loss of stability of the dominance equilibrium of the primary disease and oscillations. For this specific example the characteristic equation (5.11) has a complex root with a positive real part. Simulations show the presence of a stable oscillatory solution with persistence of the primary disease only.

Consider the following values for the parameters: $\delta(a) = 0$, $\mu = 0.05$, $\gamma_1 = 0.1$, $\nu = 0$. The recovery rate for the primary disease is

$$\alpha_1(a) = \begin{cases} 0, & 0 \le a < 3, \\ 1.58259, & a \ge 3. \end{cases} \tag{5.12}$$

The transmission coefficient for the primary disease is

$$\beta_1(a) = \begin{cases} 2.33193e^{2a}, & 0 \le a < 1, \\ 0, & a \ge 1. \end{cases} \tag{5.13}$$

The parameters related to the secondary disease are not relevant as $I_2 \to 0$ and $J \to 0$, but they were chosen as follows: $\beta_2 = 0.2$, $\alpha_2 = 0.1$, $\gamma_2 = 8$. The recruitment rate $\Lambda = 1$. With these parameters the reproduction numbers are $\mathcal{R}_1 = 7.207464746$ and $\mathcal{R}_2 = 1.3333$. The characteristic equation (5.11) has a root $0.05 + i\frac{\pi}{2}$ (here i denotes the imaginary unit, $i = \sqrt{-1}$). The initial conditions for the primary disease are chosen close to the equilibrium: $S_0 = 2.77$, $I_2(0) = 0$, $J_0 = 0$,

$$i_0(a) = \begin{cases} 5.2e^{-0.05a}, & 0 \le a < 3, \\ 5.2e^{-1.58259(a-3)}e^{-0.05a}, & a \ge 3. \end{cases} \tag{5.14}$$

The results of the simulations are given in Figure 5.1. The step-size is 0.01, and the integration in age is for up to 100 units. Both figures give the dynamics of the proportion of all cases of the primary disease in the total population as a function of time, that is $\frac{I_1}{N}$, where I_1 is the integral in age of $i(a,t)$.

In the first figure all oscillations are presented. They are so dense that the space they occupy looks like a solid. The oscillations grow in magnitude up to time unit 1000, and then they stabilize in magnitude. The solution takes a long time to stabilize into sustained oscillations because the real part of the eigenvalue with positive real part is relatively small: 0.05. In the second figure a zoomed-in picture is presented for the oscillations between time units 1900 and 1950.

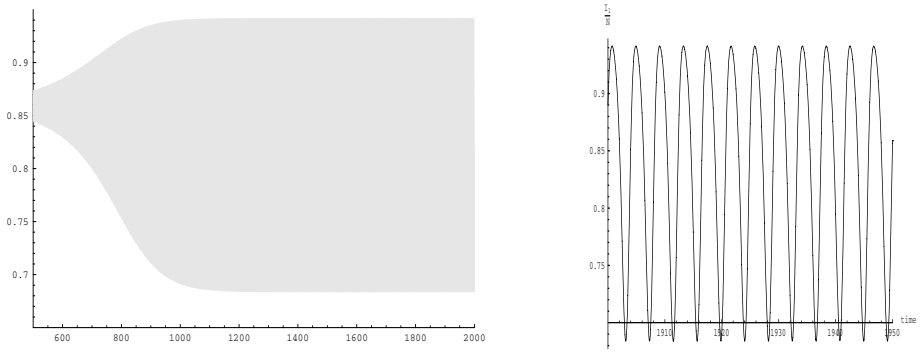


FIG. 5.1. *Left: the proportion of individuals infected with the primary disease* $\frac{I_1}{N}$*, where* I_1 *is the integral in age of* $i(a,t)$ *for up to* 2000 *time units. The horizontal axis shows the time. Right: a sample of the solution between the time units* 1900 *and* 1950*. The numerical solution exhibits sustained oscillations.*

5.3. Stability of the secondary disease equilibrium. In this subsection we establish the local stability properties of the equilibrium $\mathcal{E}_2$ whenever it exists. Thus, unlike $\mathcal{E}_1$, the presence of host age-structure does not lead to oscillations in the dominance equilibrium of the secondary disease. In this case $i(0) = 0$, $j = 0$, $s = \frac{1}{\mathcal{R}_2}$, and $i_2 = \hat{i}_2 = 1 - \frac{1}{\mathcal{R}_2}$. The linear eigenvalue problem becomes

$$
\begin{aligned}
\lambda x &= -s\int_0^\infty \beta_1(a)y(a)da - \beta_2 sz - \beta_2 x i_2 \\
&\quad -(\gamma_1+\gamma_2)sw - \mu x + \int_0^\infty \alpha_1(a)y(a)da + \alpha_2 z, \\
y'(a) &= -\lambda y - \alpha_1(a)y - \delta(a)i_2 y - \mu y, \\
y(0) &= s\int_0^\infty \beta_1(a)y(a)da + \gamma_1 sw, \\
\lambda z &= \beta_2 sz + \beta_2 i_2 x + \gamma_2 sw - (\mu+\alpha_2)z, \\
\lambda w &= i_2\int_0^\infty \delta(a)y(a)da - (\mu+\nu)w.
\end{aligned}
\tag{5.15}
$$

From the last equation we have

$$
w = \frac{i_2}{\lambda+\mu+\nu}\int_0^\infty \delta(a)y(a)da.
$$

From the equation for $y(a)$ we have

$$
y(a) = y(0)\Gamma(a;i_2)\pi_1(a)e^{-(\lambda+\mu)a}.
$$

Substituting in the equation for the initial condition $y(0)$ and assuming that $y(0) \neq 0$, we obtain the following characteristic equation:

$$
sB_1(\lambda) + \frac{\gamma_1 s i_2}{\lambda+\mu+\nu}D_1(\lambda) = 1, \tag{5.16}
$$

where we have used the notation

$$B_1(\lambda) = \int_0^\infty \beta_1(a)\Gamma(a; i_2)\pi_1(a)e^{-(\lambda+\mu)a}da,$$
$$D_1(\lambda) = \int_0^\infty \delta(a)\Gamma(a; i_2)\pi_1(a)e^{-(\lambda+\mu)a}da.$$

Clearly, $B_1(0) = B(i_2)$ and $D_1(0) = D(i_2)$. Now we are ready to prove the main result in this subsection.

THEOREM 5.5. *Let $\mathcal{R}_2 > 1$. Assume that the primary disease cannot invade the equilibrium of the secondary disease; that is, $\hat{\mathcal{R}}_1 < 1$. Then the equilibrium $\mathcal{E}_2$ is locally asymptotically stable, and the secondary disease dominates in the population. If $\hat{\mathcal{R}}_1 > 1$, the equilibrium $\mathcal{E}_2$ is unstable.*

Proof. Denote by $\mathcal{G}(\lambda)$ the left-hand side of the characteristic equation (5.16). We notice that

$$\mathcal{G}(0) = \hat{\mathcal{R}}_1.$$

First we assume that $\hat{\mathcal{R}}_1 > 1$. We consider $\mathcal{G}(\lambda)$ as a function of a real variable. We have $\mathcal{G}(0) = \hat{\mathcal{R}}_1 > 1$. In addition, $\mathcal{G}(\lambda) \to 0$ as $\lambda \to \infty$. Consequently, there exists $\lambda^* > 0$ such that $\mathcal{G}(\lambda^*) = 1$ and the equilibrium $\mathcal{E}_2$ is unstable.

Next, we assume $\hat{\mathcal{R}}_1 < 1$. For λ's with real part $\Re\lambda \geq 0$ we have

$$\begin{aligned}|\mathcal{G}(\lambda)| &\leq s|B_1(\lambda)| + \frac{\gamma_1 s i_2}{|\lambda+\mu+\nu|}|D_1(\lambda)| \\ &\leq sB_1(\Re\lambda) + \frac{\gamma_1 s i_2}{\Re\lambda+\mu+\nu}D_1(\Re\lambda) \\ &\leq sB(i_2) + \frac{\gamma_1 s i_2}{\mu+\nu}D(i_2) = \hat{\mathcal{R}}_1 < 1.\end{aligned}$$

Thus, the characteristic equation (5.16) has no solution with nonnegative real part. Furthermore, for $y(0) = 0$ we have that $y(a) = 0$ and $w = 0$. The remaining two equations become

$$\begin{aligned}\lambda x &= -\beta_2 s z - \beta_2 x i_2 - \mu x + \alpha_2 z, \\ \lambda z &= \beta_2 s z + \beta_2 i_2 x - (\mu+\alpha_2)z.\end{aligned} \tag{5.17}$$

We express x from the first equation,

$$x = \frac{(-\beta_2 s + \alpha_2)z}{\lambda+\mu+\beta_2 i_2},$$

and substitute in the equation for z. Assuming that z is nonzero, we cancel it to obtain the following characteristic equation:

$$(\lambda+\mu)(\lambda+\mu+\alpha_2+\beta_2 i_2-\beta_2 s) = 0.$$

Noticing that $\beta_2 s = \mu+\alpha_2$, we obtain the eigenvalues $-\mu$ and $-\beta_2 i_2$, which are both negative. Consequently, the equilibrium $\mathcal{E}_2$ is locally asymptotically stable. This concludes the proof. ∎

5.4. Loss of stability of a coexistence equilibrium—Oscillatory coexistence. The stability of the coexistence equilibria depends on the analysis of the perturbation equations (5.1). For the general case, however, it is difficult to derive the corresponding characteristic equation, let alone analyze the positions of its roots. Since the main thrust is that a characteristic equation of this complexity is likely to have roots with positive real part, we address the more interesting and tractable question of whether a Hopf bifurcation of a coexistence equilibrium can occur in the absence of age structure, that is, in the case when $\beta_1(a) = \beta_1$, $\alpha_1(a) = \alpha_1$, and $\delta(a) = \delta$. We established that in the constant coefficient case the two dominance equilibria are locally stable. Some additional but simple argument shows that in the absence of the second disease, the solutions converge to the dominance equilibrium, provided that the reproduction number is larger than one and that no oscillations are possible. Thus, if a Hopf bifurcation occurs, the loss of stability of the coexistence equilibrium is due to the presence of the competitor.

It turns out that a Hopf bifurcation of the coexistence equilibrium occurs for a limiting and much simpler form of the original system (2.1) taken with constant coefficients corresponding to $\alpha_1 = \alpha_2 = \gamma_1 = \nu = 0$. Since $\nu = 0$, the total population size is asymptotically constant, $N(t) \to \frac{\Lambda}{\mu} = N^*$. We will restrict our analysis to this invariant subspace [21]. We further rescale all state variables by $1/N^*$ and consider the system

$$\begin{aligned} i_1' &= \beta_1 s i_1 - \mu i_1 - \delta i_1 i_2, \\ i_2' &= \beta_2 s i_2 + \gamma_2 s j - \mu i_2, \\ j' &= \delta i_1 i_2 - \mu j, \end{aligned} \tag{5.18}$$

where $s \equiv 1 - i_1 - i_2 - j$. We establish the existence of Hopf bifurcation for values of the parameters satisfying the inequalities $\beta_2 < \mu < \beta_1 < \gamma_2$. In this case, we have that

$$\mathcal{R}_1 = \frac{\beta_1}{\mu} > 1 > \frac{\beta_2}{\mu} = \mathcal{R}_2,$$

which implies that $\mathcal{E}_2$ does not exist, and the secondary disease alone is always eliminated. Hence, the coexistence of both diseases must be mediated by the presence of the competitor, that is, the primary disease.

It is convenient to fix the parameters μ, β_1, β_2, and γ_2 and treat δ as a bifurcation parameter. Solving for positive coexistence equilibria, we find

$$s = \frac{\mu\gamma_2 + \gamma_2\delta - \beta_1\mu}{\gamma_2(\delta + \beta_1) - \beta_1\beta_2}, \qquad i_1 = \frac{\mu^2 - \beta_2\mu s}{\gamma_2\delta s}, \qquad i_2 = \frac{\beta_1 s - \mu}{\delta}, \qquad j = \frac{\delta i_1 i_2}{\mu}.$$

Under the above conditions, s and i_1 are automatically positive. The value of i_2 is positive if and only if $\delta > \delta^*$, where

$$\delta^* = \frac{\mu\beta_1(\beta_1 - \beta_2)}{\gamma_2(\beta_1 - \mu)}.$$

Since we consider δ as a bifurcation parameter, we view the values of the positive equilibrium as functions of delta: $s = s(\delta)$, $i_1 = i_1(\delta)$, $i_2 = i_2(\delta)$, $j = j(\delta)$. The variational matrix of the system (5.18) at the positive equilibrium (i_1, i_2, j) is given

by

$$A(\delta) = \begin{pmatrix} -\beta_1 i_1 & -(\beta_1+\delta)i_1 & -\beta_1 i_1 \\ -\beta_2 i_2 - \gamma_2 j & \beta_2 s - \mu - \beta_2 i_2 - \gamma_2 j & \gamma_2(s-j) - \beta_2 i_2 \\ \delta i_2 & \delta i_1 & -\mu \end{pmatrix}. \tag{5.19}$$

Calculating the determinant of $A(\delta)$, we find that $\det A(\delta) = \delta i_1 i_2[\beta_1\mu - \gamma_2(\mu+\delta)]$. Hence, $\det A(\delta) < 0$ whenever $\delta > \delta^*$ (since then $i_2 > 0$). Since $s(\delta^*) = \frac{\mu}{\beta_1}$, $i_1(\delta^*) = 1 - \frac{\mu}{\beta_1}$, and $i_2(\delta^*) = j(\delta^*) = 0$, we have that

$$A(\delta^*) = \begin{pmatrix} \mu - \beta_1 & (\beta_1+\delta)\frac{\mu-\beta_1}{\beta_1} & \mu - \beta_1 \\ 0 & \left(\frac{\beta_2}{\beta_1} - 1\right)\mu & \frac{\gamma_2\mu}{\beta_1} \\ 0 & \delta^*\frac{\beta_1-\mu}{\beta_1} & -\mu \end{pmatrix}.$$

The eigenvalues of $A(\delta^*)$ are given by $\lambda_1 = \mu - \beta_1 < 0$, $\lambda_2 = (\frac{\beta_2}{\beta_1} - 2)\mu < 0$, $\lambda_3 = 0$. Using the continuity of eigenvalues with respect to δ, we conclude that $A(\delta)$ has three negative eigenvalues when δ is slightly greater than δ^*. Hence, the positive equilibrium is stable for "small" $\delta > \delta^*$.

Next we argue that the positive equilibrium is unstable for sufficiently large values of δ. First we notice that $\lim_{\delta\to\infty} s(\delta) = 1$. Furthermore,

$$\lim_{\delta\to\infty} \delta i_1(\delta) = \frac{\mu^2 - \beta_2\mu}{\gamma_2}, \quad \lim_{\delta\to\infty} \delta i_2(\delta) = \beta_1 - \mu, \quad \lim_{\delta\to\infty} j(\delta) = 0.$$

Thus, we have that

$$\lim_{\delta\to\infty} A(\delta) = A_\infty = \begin{pmatrix} 0 & -\frac{\mu^2-\beta_2\mu}{\gamma_2} & 0 \\ 0 & \beta_2 - \mu & \gamma_2 \\ \beta_1 - \mu & \frac{\mu^2-\beta_2\mu}{\gamma_2} & -\mu \end{pmatrix}.$$

The characteristic polynomial of A_∞ has the form

$$p_\infty(\lambda) = \lambda^3 + (2\mu - \beta_2)\lambda^2 + (\beta_1 - \mu)(\mu^2 - \beta_2\mu).$$

Since $2\mu - \beta_2 > 0$ and $(\beta_1 - \mu)(\mu^2 - \beta_2\mu) > 0$, $p_\infty(\lambda)$ has one real negative and two complex roots with positive real parts. We conclude that the positive equilibrium changes stability as we increase δ. Since the determinant of the variational matrix remains negative for all $\delta > \delta^*$, the change of stability corresponds to a Hopf bifurcation. The rigorous analysis of this bifurcation is outside of the scope of this paper.

A continuation argument can establish that this bifurcation must also occur when the parameters α_1, α_2, γ_1, and ν are small and positive. In Figure 5.2 we demonstrate the presence of oscillatory coexistence when α_1, α_2, and γ_1 are small and positive. The figure shows a periodic orbit in the three-dimensional space of the variables $I_1(t)$, $I_2(t)$, and $J(t)$. In this example the parameter values are taken as $\beta_1 = 10$, $\beta_2 = 0.2$, $\alpha_1 = 1$, $\alpha_2 = 0.1$, $\mu = 1$, $\delta = 4$, $\nu = 0$, $\gamma_1 = 0.1$, $\gamma_2 = 80$, $\Lambda = 1$. Since $\frac{\Lambda}{\mu} = 1$ the values of I_1, I_2, and J are also the values of the proportions. The reproduction number of the primary disease is $\mathcal{R}_1 = 5$, while the reproduction number of the secondary disease is below one, $\mathcal{R}_2 = 0.18182$. Despite the fact that $\mathcal{R}_1 \gg \mathcal{R}_2$, the prevalence for the secondary disease I_2 is much higher than that of the primary disease I_1—a result of the very high rate at which the jointly infected individuals can infect with the secondary disease $\gamma_2 = 80$.

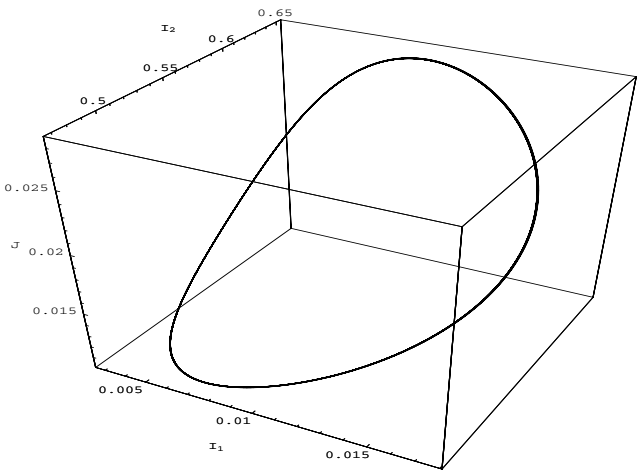


FIG. 5.2. *Existence of a periodic orbit for the age-independent case. Values of the parameters are as in the text.*

6. Backward bifurcation. In this section we analyze the existence of backward bifurcations in the system (2.1). In the single disease case, a backward bifurcation occurs when the equilibrium number (or proportion) of infectives bifurcates at the critical value of the reproduction number $\mathcal{R} = 1$ not forward but backward, and there are nontrivial equilibria when the reproduction number is below one.

In the case of two diseases, an analogous phenomenon occurs when the equilibrium number (or proportion) of infectives with each disease i_1 and i_2 bifurcates backward in both parameters $\mathcal{R}_1$ and $\mathcal{R}_2$, and nontrivial equilibria exist for values of both reproduction numbers below one. We will call this phenomenon *a two-parameter backward bifurcation*. In what follows we derive necessary and sufficient conditions for two-parameter backward bifurcation.

We treat $\beta_1(a)$ and β_2 (equivalently, $\mathcal{R}_1$ and $\mathcal{R}_2$) as bifurcation parameters and assume that all other parameters are fixed. Specifically, we define

$$\beta_1(a, \varepsilon_1) = \tilde{\beta}_1(a)(1 + \varepsilon_1 \tilde{v}_1(a)), \quad \beta_2(\varepsilon_2) = \tilde{\beta}_2(1 + \varepsilon_2),$$

so that $\tilde{\beta}_2 = \alpha_2 + \mu$ and the functions $\tilde{\beta}_1(a)$ and $\tilde{v}_1(a)$ are normalized as follows:

$$\int_0^\infty \tilde{\beta}_1(a)\pi_1(a)e^{-\mu a}\, da = \int_0^\infty \tilde{\beta}_1(a)\tilde{v}_1(a)\pi_1(a)e^{-\mu a}\, da = 1.$$

In this setting, the choice $\varepsilon_1 = \varepsilon_2 = 0$ corresponds to the basic reproduction numbers of both strains being equal to unity, that is, $\mathcal{R}_1 = \mathcal{R}_2 = 1$. We also introduce the auxiliary functions

$$B(i_2, \varepsilon_1) = \int_0^\infty \tilde{\beta}_1(a)(1 + \varepsilon_1 \tilde{v}_1(a))\Gamma(a; i_2)\pi_1(a)e^{-\mu a}\, da, \tag{6.1}$$

$$G(i_2) = \int_0^\infty \Gamma(a; i_2)\pi_1(a)e^{-\mu a}\, da, \tag{6.2}$$

where $\Gamma(a; i_2)$ is given by (3.3). Previously, we have shown that the fraction of sus-

ceptible individuals at the coexistence equilibrium must equal

$$s = \left(B(i_2, \varepsilon_1) + \frac{\gamma_1 i_2 D(i_2)}{\mu + \nu} \right)^{-1} = S(i_2, \varepsilon_1),$$

where i_2 is the fraction of individuals infected by secondary infection. Solving for $i(0)$, we find that

$$i(0) = (\mu + \nu)(\mu + \alpha_2) \frac{1 - (1 + \varepsilon_2) S(i_2, \varepsilon_1)}{\gamma_2 S(i_2, \varepsilon_1) D(i_2)} = T(i_2, \varepsilon_1, \varepsilon_2),$$

and thus the total fraction of individuals infected by primary infection is given by $i_1 = G(i_2) T(i_2, \varepsilon_1, \varepsilon_2)$. The fraction of individuals carrying both infections can be expressed as

$$j = \frac{i_2 D(i_2) T(i_2, \varepsilon_1, \varepsilon_2)}{\mu + \nu}.$$

The relation $s + i_1 + i_2 + j = 1$ now can be written as

$$M(i_2, \varepsilon_1, \varepsilon_2) = i_2 + S(i_2, \varepsilon_1) + G(i_2) T(i_2, \varepsilon_1, \varepsilon_2) + \frac{i_2 D(i_2) T(i_2, \varepsilon_1, \varepsilon_2)}{\mu + \nu} = 1.$$

Since

$$S(0,0) = \frac{1}{B(0,0)} = \left(\int_0^\infty \tilde{\beta}_1(a) \pi_1(a) e^{-\mu a} \, da \right)^{-1} = 1,$$

we find that $T(0,0,0) = 0$ and $M(0,0,0) = 1$. For a given pair $(\varepsilon_1, \varepsilon_2)$, we define the equilibrium values of $i_2(\varepsilon_1, \varepsilon_2)$ as an implicit solution of the equation $M(i_2, \varepsilon_1, \varepsilon_2) = 1$. The corresponding equilibrium values $i_1(\varepsilon_1, \varepsilon_2)$ are obtained from $i_1 = G(i_2) T(i_2, \varepsilon_1, \varepsilon_2)$. The backward bifurcation occurs whenever both functions $i_1(\varepsilon_1, \varepsilon_2)$ and $i_2(\varepsilon_1, \varepsilon_2)$ have positive values for (perhaps some) $\varepsilon_1, \varepsilon_2 < 0$. To pose the conditions for backward bifurcation we need all partial derivatives $\frac{\partial i_m}{\partial \varepsilon_n}(0,0)$, where $m, n = 1, 2$.

We compute the required partial derivatives. First, we have

$$\frac{\partial B}{\partial i_2}(0,0) = -\int_0^\infty \tilde{\beta}_1(a) \pi_1(a) e^{-\mu a} \left(\int_0^a \delta(s) \, ds \right) da = -\hat{\delta} < 0.$$

Next, if we define

$$\sigma = \hat{\delta} - \frac{\gamma_1 D(0)}{\mu + \nu}, \quad \tau = \frac{(\mu + \nu)(\mu + \alpha_2)}{\gamma_2 D(0)},$$

then the remaining partial derivatives are given by

$$\frac{\partial S}{\partial \varepsilon_1}(0,0) = -1, \quad \frac{\partial S}{\partial i_2}(0,0) = \sigma, \quad \frac{\partial T}{\partial \varepsilon_1}(0,0,0) = \tau,$$
$$\frac{\partial T}{\partial \varepsilon_2}(0,0,0) = -\tau, \quad \frac{\partial T}{\partial i_2}(0,0,0) = -\tau\sigma.$$

Finally, we have that

$$\begin{aligned} \frac{\partial M}{\partial i_2}(0,0,0) &= 1 + (1 - G(0)\tau)\sigma, \\ \frac{\partial M}{\partial \varepsilon_1}(0,0,0) = -1 + G(0)\tau, &\quad \frac{\partial M}{\partial \varepsilon_2}(0,0,0) = -G(0)\tau. \end{aligned} \tag{6.3}$$

Using the implicit function theorem, we find the derivatives of i_2 and, as a result, those of i_1:

$$\frac{\partial i_2}{\partial \varepsilon_1}(0,0) = \frac{1-G(0)\tau}{1+(1-G(0)\tau)\sigma}, \qquad \frac{\partial i_1}{\partial \varepsilon_1}(0,0) = \frac{G(0)\tau}{1+(1-G(0)\tau)\sigma}, \tag{6.4}$$

$$\frac{\partial i_2}{\partial \varepsilon_2}(0,0) = \frac{G(0)\tau}{1+(1-G(0)\tau)\sigma}, \qquad \frac{\partial i_1}{\partial \varepsilon_2}(0,0) = \frac{-G(0)\tau(1+\sigma)}{1+(1-G(0)\tau)\sigma}. \tag{6.5}$$

Since $G(0)\tau > 0$, all of these partial derivatives are negative if and only if

$$1+(1-G(0)\tau)\sigma < 0 \quad \text{and} \quad 1-G(0)\tau > 0. \tag{6.6}$$

Note that (6.6) enforces $\sigma < -1$.

Since we consider a two-parameter bifurcation, it may occur for all pairs $(\varepsilon_1, \varepsilon_2)$ or only for some pairs $(\varepsilon_1, \varepsilon_2)$. We will call a backward bifurcation *total* if the positive equilibrium exists for *all* pairs $(\varepsilon_1, \varepsilon_2)$ with sufficiently small $\varepsilon_k < 0$, $k = 1, 2$. We will call a backward bifurcation *partial* if the positive equilibrium exists for *some* pairs $(\varepsilon_1, \varepsilon_2)$ with sufficiently small $\varepsilon_k < 0$, $k = 1, 2$. In what follows, we argue that the model (2.1) admits only total backward bifurcations.

Indeed, a partial backward bifurcation occurs if and only if there exist pairs of positive numbers (ω_1, ω_2) such that

$$\omega_1 \frac{\partial i_1}{\partial \varepsilon_1}(0,0) + \omega_2 \frac{\partial i_1}{\partial \varepsilon_2}(0,0) < 0,$$
$$\omega_1 \frac{\partial i_2}{\partial \varepsilon_1}(0,0) + \omega_2 \frac{\partial i_2}{\partial \varepsilon_2}(0,0) < 0.$$

In contrast, total backward bifurcation occurs if the above inequalities are valid for all pairs of nonnegative numbers (ω_1, ω_2). These inequalities are equivalent to

$$\frac{(1-G(0)\tau)+\omega G(0)\tau}{1+(1-G(0)\tau)\sigma} < 0, \tag{6.7}$$

$$\frac{G(0)\tau(1-\omega(1+\sigma))}{1+(1-G(0)\tau)\sigma} < 0, \tag{6.8}$$

where $\omega = \omega_2/\omega_1 > 0$ (we assume $\omega_1 > 0$). We also note that $G(0)\tau > 0$.

Suppose that $1+(1-G(0)\tau)\sigma > 0$. Then (6.7)–(6.8) imply that $\sigma > -1$ and

$$\frac{1}{1+\sigma} < \omega < \frac{G(0)\tau - 1}{G(0)\tau},$$

and thus $G(0)\tau < (G(0)\tau - 1)(1+\sigma)$. The last inequality clearly contradicts $1+(1-G(0)\tau)\sigma > 0$. No backward bifurcations occur in this case.

Now suppose that $1+(1-G(0)\tau)\sigma < 0$. Then (6.7)–(6.8) imply that

$$\omega > \frac{G(0)\tau - 1}{G(0)\tau}, \quad 1-\omega(1+\sigma) > 0.$$

If $1+\sigma > 0$, then the second inequality implies that

$$\frac{G(0)\tau - 1}{G(0)\tau} < \omega < \frac{1}{1+\sigma},$$

and thus $G(0)\tau > (G(0)\tau - 1)(1+\sigma)$, which is a contradiction. If $1+\sigma < 0$, then $1-\omega(1+\sigma) > 0$ holds for all $\omega > 0$. On the other hand, we must have that $\sigma < 0$ and

$$G(0)\tau < \frac{1+\sigma}{\sigma} < 1.$$

Therefore, $(1-G(0)\tau) + \omega G(0)\tau > 0$ also holds for all $\omega \geq 0$. In this case, the backward bifurcation is total. We conclude that only total backward bifurcations may occur in this model, and the criterion is given by (6.6). We summarize this result in the following proposition.

PROPOSITION 6.1. *The model* (2.1) *exhibits the backward bifurcation if and only if*

$$1 + (1 - G(0)\tau)\sigma < 0 \quad \text{and} \quad 1 - G(0)\tau > 0. \tag{6.9}$$

If at least one of the inequalities in (6.9) *does not hold, then the model* (2.1) *does not admit any nontrivial equilibria with* $\mathcal{R}_1, \mathcal{R}_2 < 1$. *If both inequalities in* (6.9) *hold, then the backward bifurcation is total; that is, there exists a sufficiently small* $0 < \varepsilon_0 < 1$ *such that the model* (2.1) *admits nontrivial equilibria for all pairs of the reproduction numbers* $(\mathcal{R}_1, \mathcal{R}_2)$ *such that* $1 - \varepsilon_0 < \mathcal{R}_1, \mathcal{R}_2 < 1$.

7. Numerical results. In this section we consider several types of complex behavior which stem largely from the presence of coinfection. These regimes have important consequences for the development and eradication of one or both diseases.

We consider the following phenomena: subthreshold coexistence equilibria, multiple coexistence equilibria, and bistable dominance. Subthreshold coexistence equilibria may be generated by two-parameter backward bifurcation. These are multiple coexistence equilibria (two in our case), but multiple coexistence equilibria may also exist superthreshold. Finally, we consider the bistability of the dominance equilibria, which is defined as dominance of one of the diseases depending on the initial conditions. All these are illustrated in Figure 7.1. The figure is generated with the following values of the parameters: $\alpha_1 = 14$, $\alpha_2 = 25$, $\mu = 3.9$, $\nu = 0.1$, $\delta = 20$,

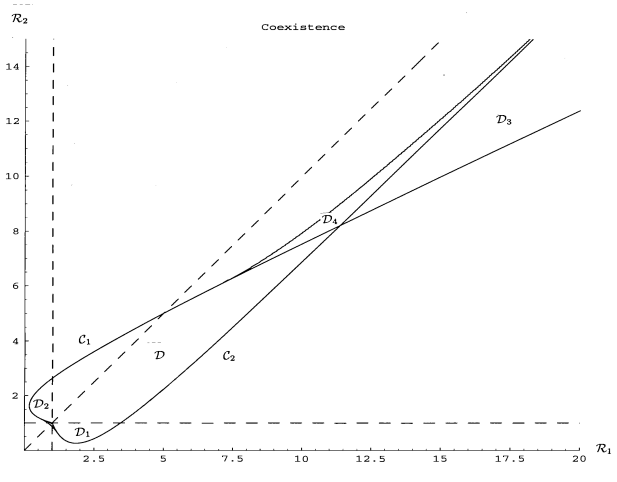


FIG. 7.1. *Boundaries of coexistence and stability of dominance equilibria. Parameters as in text.*

$\gamma_1 = 20$, $\gamma_2 = 20$. The values of $\mathcal{R}_1$ and $\mathcal{R}_2$ are treated as operating parameters that are directly related to the values of β_1 and β_2. In this section, we consider only the case where $\beta_1(a) \equiv \beta_1$ is age-independent. The values of $\mathcal{R}_1$ and $\mathcal{R}_2$ are plotted on the x and y axes, respectively. The upper of the two curves that originate at $(1, 1)$ is obtained from the equation $\hat{\mathcal{R}}_1 = 1$, while the lower is obtained from the equation $\hat{\mathcal{R}}_2 = 1$. We denoted these curves by $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively. The geometry of these curves was analyzed in Lemma 3.1.

7.1. Backward bifurcation and subthreshold equilibria. The presence of subthreshold equilibria has important implications for the control of a single disease. It means that the disease might not be eradicated by reducing its reproduction number slightly below one. Instead, it is necessary to reduce the reproduction number below the minimal transition value $\mathcal{R}^*$ such that there are no nontrivial equilibria for values of the reproduction number below $\mathcal{R}^*$.

When multiple diseases are present the situation is more complex. We call a coexistence equilibrium *subthreshold* if it occurs when at least one of the reproduction numbers is below one. Furthermore, there are two distinct cases with different consequences for the control of the diseases. In the first scenario, coexistence equilibria occur when exactly one of the reproduction number is below one. We will call those *weakly subthreshold* equilibria. In Figure 7.1 weakly subthreshold coexistence equilibria occur both in the case $\mathcal{R}_1 < 1$, $\mathcal{R}_2 > 1$ and in the case $\mathcal{R}_1 > 1$, $\mathcal{R}_2 < 1$. Those are to be found to the right of the curve $\mathcal{C}_1$ but to the left of the line $\mathcal{R}_1 = 1$ (Figure 7.1, region $\mathcal{D}_2$) and above the curve $\mathcal{C}_2$ but below the line $\mathcal{R}_2 = 1$ correspondingly (Figure 7.1, region $\mathcal{D}_1$). Given coinfection $\delta \neq 0$, a necessary condition for the first area to be nonempty is that $\gamma_1 \neq 0$; similarly, the second area can be nonempty only if $\gamma_2 \neq 0$. In both of these areas there is a unique coexistence equilibrium not obtained as a result of a backward bifurcation. In terms of disease control the presence of weakly subthreshold equilibria leads to the fact that reducing only one of the reproduction numbers below unity does not necessarily lead to the disappearance of the corresponding disease. Thus, eradicating only one of the two diseases may be difficult, particularly as the curves $\mathcal{C}_1$ and $\mathcal{C}_2$ pass very close to the corresponding axes. However, if both reproduction numbers are brought slightly below unity, both diseases will be eliminated. We note here that we may have weakly subthreshold coexistence equilibria only with $\mathcal{R}_1 < 1$, $\mathcal{R}_2 > 1$ without having such with $\mathcal{R}_1 > 1$, $\mathcal{R}_2 < 1$ or vice versa (not shown). In this case only the primary disease cannot be eliminated by reducing $\mathcal{R}_1$ below one, while the secondary will be eliminated if $\mathcal{R}_2$ is reduced below one. Weakly subthreshold equilibria also appear as a consequence of backward bifurcation (see Figure 7.2) and are discussed more in the next subsection.

In the second scenario, coexistence equilibria occur when both of the reproduction numbers are below one. We call those *strongly subthreshold* coexistence equilibria. Strongly subthreshold equilibria in our model are the result of a backward bifurcation in two parameters, namely $\mathcal{R}_1$ and $\mathcal{R}_2$. We established necessary and sufficient conditions for the two-parameter backward bifurcation in the previous section. If we consider J^* as the coexistence variable and we view it as a function of $\mathcal{R}_1$ and $\mathcal{R}_2$, then the surface $J^* = f(\mathcal{R}_1, \mathcal{R}_2)$ bifurcates backwards along the curves $\mathcal{C}_1$ and $\mathcal{C}_2$ near the critical point $(1, 1)$ and then turns around and heads in the direction of increasing values of $\mathcal{R}_1$ and $\mathcal{R}_2$. The projection of the turning curve on the plane $(\mathcal{R}_1, \mathcal{R}_2)$ is the curve that connects $\mathcal{C}_1$ and $\mathcal{C}_2$ (see Figure 7.2, region $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3$). In analogy with the single disease case, we will call this curve the *minimal transition curve*. In Figure 7.2 the area enclosed by the curves $\mathcal{C}_1$, $\mathcal{C}_2$ and the minimal transition curve,

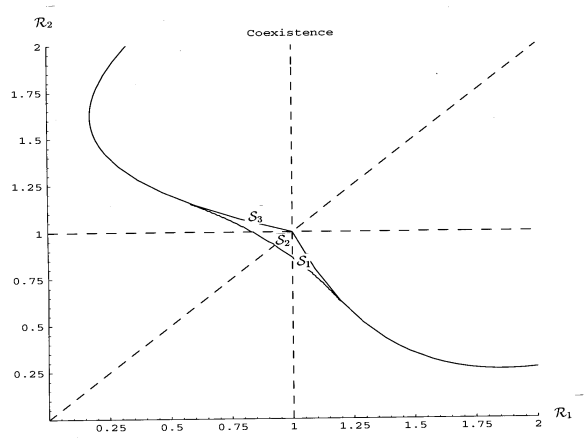


FIG. 7.2. *This is a zoom-in of the area from Figure* 7.1 *near the critical point where both reproduction numbers are near one.*

$\mathcal{S}$, is the projection of the overlapping branches of the surface $J^* = f(\mathcal{R}_1, \mathcal{R}_2)$. Thus, in this area there are two distinct coexistence equilibria. Figure 7.2 is a zoom-in of the part of Figure 7.1 near the critical point $(1, 1)$.

Next, we show that backward bifurcation occurs if and only if the angle between the tangent lines to the curves $\mathcal{C}_1$ and $\mathcal{C}_2$ (see Figure 7.1) at the point $(1, 1)$ is sufficiently small—smaller than 180°. Let l_1 be the tangent to $\mathcal{C}_1$ with slope m_1, and l_2 be the tangent to the curve $\mathcal{C}_2$ with slope m_2. Along the curve $\mathcal{C}_1$ we have $i_1(\varepsilon_1, \varepsilon_2) = 0$. Along the curve $\mathcal{C}_2$ we have $i_2(\varepsilon_1, \varepsilon_2) = 0$. The slopes of l_1 and l_2 are given by $\frac{d\varepsilon_2}{d\varepsilon_1}$, which is obtained for each curve by differentiating implicitly. Thus, by (6.4) and (6.5),

$$m_1 = \frac{1}{1+\sigma}, \qquad m_2 = -\frac{1 - G(0)\tau}{G(0)\tau}.$$

The angle between the tangents is obtuse if $m_2 < m_1 < 0$. The angle between the tangents is larger than 180° and backward bifurcation does not occur if $m_1 < m_2 < 0$. Consequently, the conditions for the angle to be obtuse are

$$\frac{G(0)\tau - 1}{G(0)\tau} < \frac{1}{1+\sigma} < 0.$$

It is easy to see that these inequalities are equivalent to the inequalities (6.9).

The fact that backward bifurcation occurs only if the angle between the tangent lines of the curves $\mathcal{C}_1$ and $\mathcal{C}_2$ at the point $(1, 1)$ is obtuse implies that strongly subthreshold coexistence equilibria are present only in conjunction with both types of weakly subthreshold coexistence equilibria. Thus, the existence of strongly subthreshold coexistence equilibria through two-parameter backward bifurcation is the analogue of the backward bifurcation in the single disease case. It has the same implication for the disease control—reducing both reproduction numbers slightly below one does not lead to the eradication of either disease. It is necessary to reduce both reproduction numbers in the square $[0, 1] \times [0, 1]$ below the minimal transition curve.

7.2. Multiple coexistence equilibria. Bistability. The presence of multiple equilibria, and particularly of multiple stable equilibria, can have significant impact on the outcome of the disease, as for a fixed set of parameters this outcome depends on the initial status of the population. For the present model and parameter values, as in Figure 7.1, results in previous sections and simulations suggested the presence of multiple coexistence equilibria in two areas.

The first such area is the subthreshold area $\mathcal{S}$ illustrated also in Figure 7.2. As we discussed above, the multiple equilibria there are obtained from backward bifurcation. In this case there are two coexistence equilibria. If they are ordered in increasing order of J^*, simulations suggest that the lower one is unstable, while the upper one is locally stable. In the subregion $\mathcal{S}_2$ there is also the disease-free equilibrium which is locally stable. Thus, in that region the two diseases might coexist, or they might both disappear depending on the initial conditions. Looking at Figure 7.2, we see that the area of backward bifurcation overlaps also with the regions $\mathcal{R}_1 > 1$, $\mathcal{R}_2 < 1$ forming region $\mathcal{S}_1$ and $\mathcal{R}_1 < 1$, $\mathcal{R}_2 > 1$ forming region $\mathcal{S}_3$. Consequently, we have multiple weakly subthreshold coexistence equilibria. In those regions the disease-free equilibrium is unstable. However, in addition to the locally stable coexistence equilibrium, in the region $\mathcal{S}_1$ the equilibrium $\mathcal{E}_1$ is also locally stable, while in the second region $\mathcal{S}_3$ the equilibrium $\mathcal{E}_2$ is also locally stable. Thus, the ultimate outcome is either dominance of one of the diseases or coexistence, depending on the initial conditions.

The second area where multiple coexistence equilibria exist is the superthreshold area in Figure 7.1, where the curves $\mathcal{C}_1$ and $\mathcal{C}_2$ cross and a third curve touches both of them forming a curvilinear triangle, denoted by $\mathcal{D}_4$. There are two coexistence equilibria in that area; the lower one there is stable, while the upper one is unstable. The disease-free equilibrium is again unstable. Both dominance equilibria $\mathcal{E}_1$ and $\mathcal{E}_2$ exist; however, $\mathcal{E}_1$ is unstable and $\mathcal{E}_2$ is locally stable. Consequently, if the combination of the reproduction numbers forms a point in that area, there are two possible outcomes for the long-term dynamics of the diseases: dominance of the secondary disease or coexistence. Which of the two will materialize depends on the initial status of the population.

A unique coexistence equilibrium exists in the area $\mathcal{D} = \mathcal{D}_- \cup D_+ \cup D_1 \cup D_2$ between the curves $\mathcal{C}_1$ and $\mathcal{C}_2$, which for most parameter values is locally stable. When it loses stability, oscillatory coexistence occurs.

7.3. Bistable dominance. One of distinctive features of this model is that the two curves that define the boundaries of stability of the dominance equilibria $\mathcal{E}_1$ and $\mathcal{E}_2$ always intersect (see Lemma 3.1 for details).

In the constant coefficient case, there is a unique intersection of the curves $\mathcal{C}_1$ and $\mathcal{C}_2$ that occurs at the point $(\mathcal{R}_1^*, \mathcal{R}_2^*)$, where $\mathcal{R}_1^* > 1$ and $\mathcal{R}_2^* > 1$. This intersection creates a region between the curves $\mathcal{C}_1$ and $\mathcal{C}_2$ with $\mathcal{R}_1 > \mathcal{R}_1^*$ and $\mathcal{R}_2 > \mathcal{R}_2^*$ (see Figure 7.1, area $\mathcal{D}_3$), where both dominance equilibria $\mathcal{E}_1$ and $\mathcal{E}_2$ are locally stable and the outcome of the competition between the diseases depends on the initial conditions. In other words, based only on the parameters values we cannot predict which disease will persist in the population. Figures 7.3 and 7.4 show the possible outcomes with two sets of initial conditions which differ only in the value of J_0. In the first figure $J_0 = 0.05$, while in the second $J_0 = 0.04$.

The curve $\mathcal{C}_1$ will not cross below the diagonal if $\gamma_1 = 0$, given that there is coinfection ($\delta \neq 0$). Thus the bistable dominance occurs as a result of the possibility that the jointly infected individuals can infect with the primary disease $\gamma_1 \neq 0$.

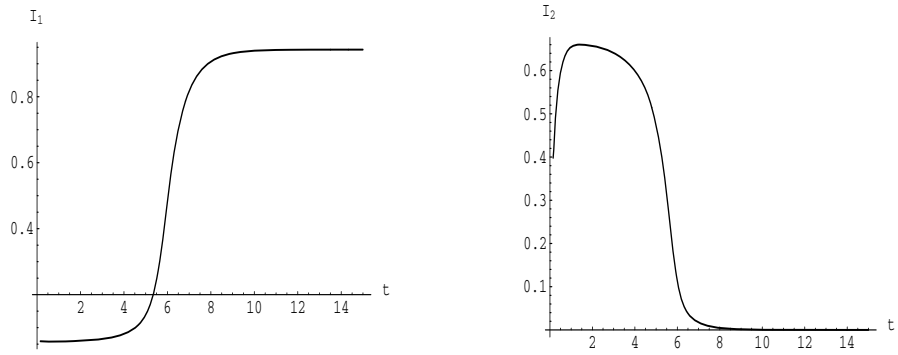


FIG. 7.3. *Primary disease persists, secondary disease dies out. Parameters are as in Figure* 7.1 *with* $\mathcal{R}_1 = 17.5$ *and* $\mathcal{R}_2 = 12$ *(from region* $\mathcal{D}_3$*). Initial conditions are* $S_0 = 0.2$, $I_1(0) = 0.01$, $I_2(0) = 0.05$, $J_0 = 0.05$.

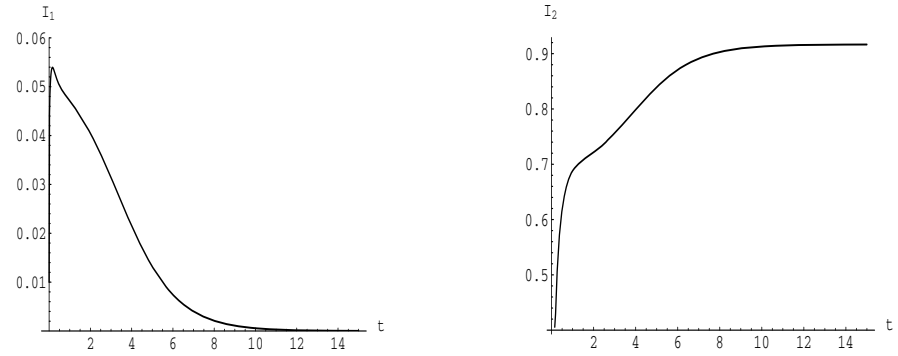


FIG. 7.4. *Primary disease dies out, secondary disease persists. Parameters as in Figure* 7.1 *with* $\mathcal{R}_1 = 17.5$ *and* $\mathcal{R}_2 = 12$ *(from region* $\mathcal{D}_3$*). Initial conditions are* $S_0 = 0.2$, $I_1(0) = 0.01$, $I_2(0) = 0.05$, $J_0 = 0.04$.

8. Discussion. At any given time thousands of diseases cocirculate in a population. Many of them participate in joint infections of a single host. New diseases like SARS appear; others fade only to re-emerge later with strains that are more difficult to treat. The complexity of interactions of the diseases through the host population can have a significant impact on the dynamics and management of each disease.

In this paper we introduce and investigate a simple epidemiological model with two diseases that can coinfect a single host. We compute the reproduction numbers and the invasion reproduction numbers of both diseases. We observe a variety of complex dynamic phenomena with significant consequences for disease control.

1. *Cooperative subthreshold coexistence.* First, we establish that the dominance equilibria $\mathcal{E}_1$ and $\mathcal{E}_2$ are present only if $\mathcal{R}_1 > 1$ and $\mathcal{R}_2 > 1$, correspondingly. That implies that neither disease can exist by itself when its reproduction number is below one. However, the "cooperation" of the two leads to subthreshold coexistence. Consequently, both diseases can persist concurrently for values of the reproduction numbers below one. We call this phenomenon *cooperative subthreshold coexistence.* We show two types of cooperative subthreshold coexistence: weakly subthreshold coexistence (occurs when exactly one of the reproduction numbers is below one) and strongly subthreshold coexistence (occurs when both reproduction numbers are below one). The strongly subthreshold coexistence is a result of backward bifurcation in both $\mathcal{R}_1$

and $\mathcal{R}_2$. Weakly subthreshold coexistence can result from backward bifurcation or from expansion of the coexistence region between the curves $\mathcal{C}_1$ and $\mathcal{C}_2$ to the below threshold areas. We derive necessary and sufficient conditions for existence of backward bifurcation. We show that the bifurcation is always *total*; that is, it occurs for all pairs of $(\mathcal{R}_1, \mathcal{R}_1)$ which are close to $(1, 1)$. A sufficient condition for backward bifurcation is that the angle between the tangents to those curves at the critical point $(1, 1)$ be obtuse, which occurs if both γ_1 and γ_2 are large. We establish that $\gamma_1 = 0$ leads to extinction of the primary disease if $\mathcal{R}_1 < 1$, and $\gamma_2 = 0$ with $\mathcal{R}_2 < 1$ leads to extinction of the secondary disease. No backward bifurcation occurs in these cases. One consequence of the observation is that public health mechanisms that lead to reduction of spread of either disease by the jointly infected individuals—like isolating those who are infected with both diseases—can have very dramatic effects on the eradication of one or both diseases. Furthermore, disease-induced mortality in the jointly infected class ν is a mechanism that impedes the backward bifurcation. This suggests that diseases which are more lethal in a combination are easier to manage from an epidemiological perspective.

2. *Restricted pathogenic diversity. Bistable dominance.* The dynamics of two diseases is reminiscent of the dynamics of two variants of the same pathogen. In many instances coexistence in stable form occurs in unbounded domains of the parameter space [17]. This is not the case here. The curves $\mathcal{C}_1$ and $\mathcal{C}_2$ intersect, thus making the region $\hat{\mathcal{R}}_1 > 1$, $\hat{\mathcal{R}}_2 > 1$ finite (Figure 7.1). We find coexistence in domains outside that one—namely, the area of backward bifurcation (Figure 7.2) and the area of two coexistence equilibria adjacent to the cross-point of $\mathcal{C}_1$ and $\mathcal{C}_2$. It appears from the simulations that these two areas are also finite. Consequently, stable coexistence is limited to finite regions in the $(\mathcal{R}_1, \mathcal{R}_2)$ plane and does not occur if the reproduction numbers are sufficiently large. We call this *restricted pathogenic diversity.* In other words, if evolution maximizes the reproduction numbers, then under this scenario it works against pathogenic diversity. It is interesting to know what mechanisms would lead to such an effect. In our case this is the ability of the jointly infected individuals to spread the primary diseases, $\gamma_1 \neq 0$. The intersection of the curves $\mathcal{C}_1$ and $\mathcal{C}_2$ also leads to emergence of a region between them where $\hat{\mathcal{R}}_1 < 1$, $\hat{\mathcal{R}}_2 < 1$. Simulations suggest that in this region there is still a unique coexistence equilibrium which is unstable. At the same time the two boundary equilibria are both locally stable. A situation like this has been described as occurring in a two-sex two-strain model of STD [6]. The result is bistable dominance—which disease persists and which dies out depend on the initial conditions, and the outcome can be very sensitive (Figures 7.3 and 7.4). In fact, bistability is somewhat common for the model (2.1). We find bistability in several regions of the $(\mathcal{R}_1, \mathcal{R}_2)$ plane, particularly where multiple coexistence equilibria exist. In all remaining cases, however, one of the possible outcomes is stable coexistence; the other is either dominance of one of the diseases or extinction.

9. Summary. In this paper, we have analyzed an epidemic model of two diseases with age-since-infection structure in the primary disease. We have obtained expressions for the basic reproduction numbers $\mathcal{R}_i$ for both diseases, and showed that the unique primary (resp., secondary) single disease equilibrium exists if and only if $\mathcal{R}_1 > 1$ (resp., $\mathcal{R}_2 > 1$). We have also shown that the disease-free equilibrium is locally stable if $\mathcal{R}_1, \mathcal{R}_2 < 1$ and unstable if $\mathcal{R}_i > 1$ for some $i = 1, 2$ (Proposition 5.1), and obtained sufficient conditions for the extinction of one or both diseases (section 4).

We have computed the invasion reproduction numbers $\hat{\mathcal{R}}_i$ for both single disease equilibria. We presented the necessary condition for the local stability of the pri-

mary disease equilibrium in Propositions 5.2 and 5.3. In the case of the secondary disease equilibrium, we presented the necessary and sufficient condition for the local stability in Proposition 5.5. In Theorem 3.2, we presented sufficient conditions for the presence of coexistence equilibria. In Proposition 6.1, we showed that multiple coexistence equilibria may exist via the backward bifurcation. In the absence of the age structure, we showed that a coexistence equilibrium can lose stability via a Hopf bifurcation (section 5.4). In general, the stability of coexistence equilibria remains an open problem. Finally, we presented results of numerical simulations that illustrate different dynamic outcomes of the interactions between the two diseases.

Acknowledgments. The authors are grateful to the anonymous referee and the handling editor for their valuable comments and suggestions.

REFERENCES

[1] L. J. S. Allen, M. Langlais, and C. J. Phillips, *The dynamics of two viral infection in a single host population with applications to hantavirus*, Math. Biosci., 186 (2003), pp. 191–217.
[2] V. Andreasen, *Multiple time scales in the dynamics of infectious diseases*, in Mathematical Approaches to Problems in Resource Management and Epidemiology, Lecture Notes in Biomath. 81, Springer-Verlag, Berlin, 1989, pp. 142–151.
[3] J. Arino, S. S. Pilyugin, and G. S. K. Wolkowicz, *Considerations on yield, nutrient uptake, cellular growth, and competition in chemostat models*, Canad. Appl. Math. Quart., 11 (2003), pp. 107–142.
[4] H. J. Bremermann and H. R. Thieme, *Competitive exclusion principle for pathogen virulence*, J. Math. Biol., 27 (1989), pp. 179–190.
[5] G. Butler and P. Waltman, *Bifurcation from a limit cycle in a two predator-one prey ecosystem modeled on a chemostat*, J. Math. Biol., 12 (1981), pp. 295–310.
[6] C. Castillo-Chavez, W. Huang, and J. Li, *Competitive exclusion and coexistence of multiple strains in an SIS STD model*, SIAM J. Appl. Math., 59 (1999), pp. 1790–1811.
[7] F. Courchamp, C. Suppo, E. Fromont, and C. Bouloux, *Dynamics of two feline retroviruses (FIV and FeLV) within one population of cats*, Proc. R. Soc. London B, 264 (1997), pp. 785–794.
[8] A. R. Dabney and J. C. Wakenfield, *Issues in the mapping of two diseases*, Stat. Meth. Med. Res., 14 (2005), pp. 83–112.
[9] A. B. Gumel, S. M. Moghadas, Y. Yuan, and P. Yu, *Bifurcation and stability analyses of a* 13*-D SEIC model using normal form reduction and numerical simulation*, Dyn. Contin. Discrete Impuls. Syst. Ser. B Appl. Algorithms, 10 (2003), pp. 317–330.
[10] C. C. Hung and S. C. Chang, *Impact of highly active antiretroviral therapy on incidence and management of human immunodeficiency virus-related opportunistic infections*, J. Antimicrob. Chemother., 54 (2005), pp. 849–853.
[11] D. Kirschner, *Dynamics of co-infection with M. tuberculosis and HIV*-1, Theoret. Pop. Biol., 55 (1999), pp. 94–109.
[12] M. Martcheva and C. Castillo-Chavez, *Diseases with chronic stage in a population with varying size*, Math. Biosci., 182 (2003), pp. 1–25.
[13] M. Martcheva and H. R. Thieme, *Progression age enhanced backward bifurcation in an epidemic model with super-infection*, J. Math. Biol., 46 (2003), pp. 385–424.
[14] R. May and M. Nowak, *Coinfection and the evolution of parasite virulence*, Proc. R. Soc. London B, 261 (1995), pp. 209–215.
[15] F. Milner and A. Pugliese, *Periodic solutions: A robust numerical method for an S-I-R model of epidemics*, J. Math. Biol., 39 (1999), pp. 471–492.
[16] J. Mosquera and F. Adler, *Evolution of virulence: A unified framework for coinfection and superinfection*, J. Theoret. Biol., 195 (1998), pp. 293–313.
[17] M. Nuño, Z. Feng, M. Martcheva, and C. Castillo-Chavez, *Dynamics of two-strain influenza with isolation and partial cross-immunity*, SIAM J. Appl. Math., 65 (2005), pp. 964–982.
[18] S. S. Pilyugin and P. Waltman, *Multiple limit cycles in the chemostat with variable yield*, Math. Biosci., 182 (2003), pp. 151–166.
[19] H. R. Thieme and C. Castillo-Chavez, *How may infection-age-dependent infectivity affect the dynamics of HIV/AIDS?*, SIAM J. Appl. Math., 53 (1993), pp. 1447–1479.

[20] H. R. Thieme, *Stability change of the endemic equilibrium in age-structured models for the spread of SIR type infectious diseases*, in Differential Equations Models in Biology, Epidemiology and Ecology, Lecture Notes in Biomath. 92, Springer-Verlag, New York, 1991, pp. 139–158.

[21] H. R. Thieme, *Convergence results and a Poincaré–Bendixson trichotomy for asymptotically autonomous differential equations*, J. Math. Biol., 30 (1992), pp. 755–763.

[22] WHO, *Herpes Simplex Virus Type 2: Programmatic and Research Priorities in Developing Countries*, report of a WHO/UNAIDS/LSHTM workshop, The World Health Organization, London, 2001; available online at http://www.who.int/docstore/hiv/herpes-meeting/.

SIAM J. APPL. MATH.
Vol. 66, No. 3, pp. 873–895

ON THE ROUTE TO EXTINCTION IN NONADIABATIC SOLID FLAMES*

J. H. PARK†, A. BAYLISS†, B. J. MATKOWSKY†, AND A. A. NEPOMNYASHCHY‡

Abstract. We consider nonadiabatic gasless solid fuel combustion employing a reaction sheet model. We derive an integrodifferential equation for the location of the interface separating the fresh fuel from the burned products. There are two parameters in our model, the Zeldovich number Z, related to the activation energy of the exothermic chemical reaction, and the heat loss parameter Γ. For any value of Z there is an extinction limit Γ_m, so that if $\Gamma > \Gamma_m$, the combustion wave cannot be sustained. For all values of Z and $\Gamma < \Gamma_m$ the model admits a uniformly propagating combustion wave. This solution is subject to a pulsating instability for Z sufficiently large. The effect of heat losses is destabilizing in the sense that pulsations occur for smaller values of Z when heat loss is considered.

We consider the dynamics of the combustion wave as Γ increases, thus, describing the dynamics of the model on the route to extinction. We consider values of Z below the adiabatic stability limit, so that for $\Gamma = 0$ the only stable steady state solution is the uniformly propagating combustion wave. We find that for Z near the adiabatic stability limit, the effect of heat loss is to promote a period doubling cascade leading to chaotic behavior prior to extinction. We also find an interval of laminar behavior within the chaotic window, corresponding to a secondary period doubling sequence. Specifically, we find solutions of period $12T, 24T, 48T$. We show that for smaller values of Z the full period doubled sequence does not necessarily occur. Rather, extinction follows after a finite, possibly small, number of periodic solutions.

Key words. nonadiabatic combustion wave dynamics, solid flame, chaos, period doubling sequence

AMS subject classifications. 35K55, 35B32, 80A25

DOI. 10.1137/050638564

1. Introduction.

The model of gasless solid fuel combustion describes the SHS (Self-propagating High Temperature Synthesis) process of materials synthesis. In this process reactants are ground into a powder, cold pressed, and ignited at one end. A high temperature combustion wave then propagates through the sample converting reactants into products. When gas plays no significant role in the process, the resulting gasless combustion wave is referred to as a "solid flame" and can be modeled, as in this paper, as gasless, solid fuel combustion. The process was pioneered in the former Soviet Union, and has subsequently been the focus of a great deal of research, e.g., [12, 13, 15]. The SHS process enjoys a number of advantages over conventional technology, in which the sample is placed into a furnace and "baked" until it is "well done." The advantages include (i) simpler equipment; (ii) significantly shorter synthesis times; (iii) greater economy, since the internal energy of the chemical reactions is employed rather than the costly external energy of the furnace; (iv) greater product purity; due to volatile impurities being burned off by the very high combustion temperatures of

*Received by the editors August 21, 2005; accepted (in revised form) October 6, 2005; published electronically February 21, 2006. This work was supported by NSF grants DMS02-02485, DMS00-72491, NSF-IGERT grant DGE99-87577, and the Japan Technion Society Research Fund.
http://www.siam.org/journals/siap/66-3/63856.html
†Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208 (jangpark@northwestern.edu, a-bayliss@northwestern.edu, b-matkowsky@northwestern.edu).
‡Department of Mathematics and Minerva Center for Nonlinear Physics of Complex Systems, The Technion-Israel Institute of Technology, Haifa, Israel (nepom@math.technion.ac.il).

the propagating combustion wave; and (v) no intrinsic limit on the size of the sample to be synthesized, as exists in conventional technology.

It is known that in many instances the combustion wave does not propagate in a uniform spatial and temporal manner, but rather nonuniformities can develop in the front speed and in the temperature along the front. Since the mode of propagation determines the microstructure of the product, i.e., the nature of the final product, a study of different modes of propagation is important for technological applications of the SHS process. For example, if the temperature is different at two different spatial points during the synthesis process, the product at these points will be different. It is known that for sufficiently large activation energies (more precisely Zeldovich numbers, defined below) the uniformly propagating combustion wave is unstable. In this case the only stable planar mode is the planar pulsating mode (autooscillatory combustion), in which there is no spatial structure along the front, i.e., the combustion wave remains planar. However, the front speed and temperature on the front oscillate in time in a periodic, quasiperiodic or chaotic fashion, (see, e.g., [1, 2, 3, 4, 5, 8, 9, 11, 14, 16, 17]). We note that for some parameters there exist various types of nonplanar modes, e.g., hot spots traveling on a helical path on the surface of a cylindrical sample, and other modes exhibiting yet more complex spatiotemporal dynamics. However, in this paper we only consider planar modes of propagation.

Most planar computations consider adiabatic combustion and study the effect of increasing the activation energy or Zeldovich number on the nonlinear dynamics. Computations have shown that in this case the nonlinear dynamics exhibits a period doubling sequence [3, 4, 5]. While the period doubling route to chaos for adiabatic solid fuel combustion as the Zeldovich number Z increases has been fairly well established, the role of heat loss has not been considered as extensively. In this paper we consider nonadiabatic combustion and focus on the nonlinear dynamics occurring prior to extinction. We consider the case that Z is below the pulsating stability boundary, so that pulsations arise solely due to heat loss. This problem was considered in [6] which demonstrated that increasing heat loss promotes the pulsating instability, and in [11] which computed period T and period $2T$ behavior. It should be noted that the route to chaos does not always proceed by period doubling. For example, if the effect of melting of the solid fuel prior to combustion is accounted for, the route to chaos is via intermittency [3, 16]. Moreover, there are a number of successive windows in which there is period T behavior, followed by period $2T$ behavior, followed by a return to period T behavior prior to the onset of intermittency.

We consider the route to extinction for one-dimensional nonadiabatic gasless solid fuel combustion. Specifically, we consider the reaction sheet model and derive an integrodifferential equation for the motion of the interface separating the fresh fuel from the burned products. The model, which has a basic solution describing a uniformly propagating planar wave, depends on various parameters, including the Zeldovich number Z (a nondimensional measure of the activation energy of the exothermic chemical reaction) and the nondimensional heat loss parameter Γ. Increasing Z above a critical value Z_c or increasing Γ above a critical value Γ_c destabilizes the basic solution to pulsations. We consider in particular the parameter regime below the pulsating boundary Z_c so that without heat loss the uniformly propagating combustion wave is stable and there are no pulsations. We show that as the heat loss Γ increases the system undergoes transitions to chaos via a period doubling sequence. We are able to compute nT periodic solutions where $n = 1, 2, 4, 8, 16$ denotes the the degree of doubling of the period, i.e., a T solution is a singly periodic solution while a $2T$ solution is a solution which has undergone one period doubling. While we have not

computed period doubled solutions beyond $16T$, we did compute apparently chaotic solutions near the extinction limit. We also show that within the chaotic window, there is a region containing a secondary period doubling sequence. Specifically, we computed solutions with periods $12T$, $24T$, and $48T$. We note that the adiabatic problem ($\Gamma = 0$) was reformulated as an integral equation in [7] which investigated T periodic relaxation oscillations and [5] which computed a period doubling route to chaos as Z increased, thus recovering the results in [3, 4].

In section 2 we describe the mathematical model and the basic solution which describes a uniformly propagating planar combustion wave. In section 3 we consider the linear stability of the basic solution and identify the pulsating stability boundary, while in section 4 we derive the integrodifferential equation which describes the motion of the interface. In section 5 we describe the numerical method which we employ and in section 6 we describe the results of our computations. Finally, in section 7 we summarize our results.

2. Mathematical model and basic solution. We consider a planar combustion front propagating in a solid fuel in the direction $-x_*$. We account for heat losses which are assumed to be proportional to the difference between the local temperature T_* and an ambient temperature T_0. The temperature field $T_*(x_*, t_*)$ and the concentration field $C_*(x_*, t_*)$ of the deficient reaction component are governed by

$$\frac{\partial T_*}{\partial t_*} = \kappa \frac{\partial^2 T_*}{\partial x_*^2} - \gamma(T_* - T_0) + QW(T_*, C_*), \qquad \frac{\partial C_*}{\partial t_*} = -W(T_*, C_*), \tag{1}$$

$$-\infty < x_* < \infty, \quad 0 < t_* < \infty,$$

where κ is the thermal diffusivity of the fuel, the constant γ is the heat loss coefficient, and the constant Q denotes the heat release of the exothermic chemical reaction, all scaled by ρc, where ρ denotes the fuel density and c its specific heat. We have assumed that the fuel sample is sufficiently long to permit a traveling wave to be established, so that the domain is taken to be $-\infty < x_* < \infty$. The reaction rate $W(T_*, C_*)$ is generally taken to be of Arrhenius form

$$W(T_*, C_*) = kC_* e^{\frac{-E}{RT_*}}, \tag{2}$$

where E is the activation energy of the reaction, R is the universal gas constant and k is the preexponential factor. The reaction zone in combustion reactions is typically thin since the activation energy is typically large. Therefore, we employ the reaction sheet assumption, i.e., we replace the Arrhenius reaction kinetics by delta-function kinetics [6, 8, 9],

$$W = w \exp\left[\frac{E}{2RT_a^2}(T_* - T_a)\right] \delta(x_* - \xi_*(t_*)), \tag{3}$$

located at the interface $x_* = \xi_*(t_*)$ separating the fresh fuel where $C_* = C_{*0}$ from the burned products where $C_* = 0$ corresponding to complete consumption of the fuel. Here,

$$T_a = T_0 + QC_{*0} \tag{4}$$

is the adiabatic interface temperature, i.e., the burned temperature in the absence of the heat losses, and w is a constant proportional to k. The reaction sheet assumption in solid fuel combustion is motivated by analogy with the case of gaseous

combustion, where it was systematically derived [10] for large activation energies. For solid fuel combustion, as considered here, it was "derived" by formal truncation of an asymptotic series for large Z [8].

The condition of complete consumption of fuel determines the dependence of the interface velocity on the reaction rate

$$\left(-\frac{d\xi_*}{dt_*}\right) C_{*0} = w \exp\left[\frac{E}{2RT_a^2}(T_*(\xi_*(t_*), t_*) - T_a)\right]. \tag{5}$$

Solving the system (1), (3), (5) with the boundary conditions

$$T_*(-\infty, t) = T_*(\infty, t) = T_0 \tag{6}$$

under the assumption of steady propagation,

$$\xi_*(t_*) = \xi_*(0) - v_f t_*, \tag{7}$$

we obtain the temperature profiles,

$$T_*(x_*, t_*) = T_0 + (T_{*f} - T_0) e^{k_-(x_* - \xi_*(t_*))}, \quad x_* - \xi_*(t_*) \le 0, \tag{8}$$

$$T_*(x_*, t_*) = T_0 + (T_{*f} - T_0) e^{-k_+(x_* - \xi_*(t_*))}, \quad x_* - \xi_*(t_*) \ge 0,$$

where

$$k_\pm = \frac{v_f}{2\kappa}\left[\mp 1 + \sqrt{1 + \frac{4\kappa\gamma}{v_f^2}}\,\right], \qquad T_{*f} = T_0 + \frac{QC_{0*}}{\sqrt{1 + 4\kappa\gamma/v_f^2}}. \tag{9}$$

Substituting (7) and (8) into (5) and employing (4), we find the following equation for the interface velocity v_f

$$v_f = \frac{w}{C_{0*}} \exp\left[\frac{EQC_{0*}}{2RT_a^2}\left(\frac{1}{\sqrt{1 + 4\kappa\gamma/v_f^2}} - 1\right)\right]. \tag{10}$$

In contrast to the adiabatic case $\gamma = 0$, where $v_f = w/C_{0*}$ is the sole solution of (10), for $\gamma > 0$, (10) may have several solutions.

We define the nondimensional quantities

$$Z = \frac{QEC_{*0}}{2RT_a^2} \tag{11}$$

(Zeldovich number),

$$H = \frac{4\kappa\gamma C_{*0}^2}{w^2}, \qquad V = \frac{v_f C_{*0}}{w} \tag{12}$$

and rewrite (10) in nondimensional form as

$$V = \exp\left[Z\left(\frac{1}{\sqrt{1 + H/V^2}} - 1\right)\right]. \tag{13}$$

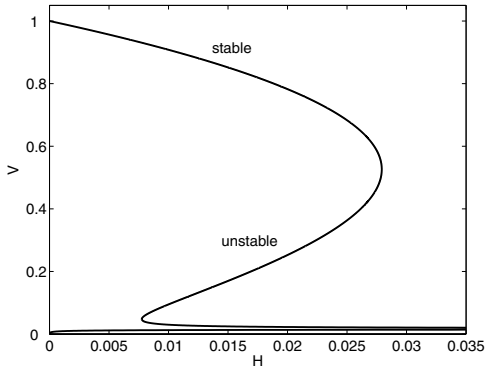


FIG. 1. *Plot of V vs. H based on* (14).

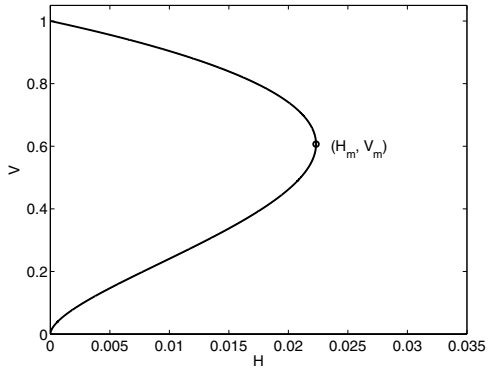


FIG. 2. *Plot of V vs. H based on* (15).

Solving (13) for H, we find

$$H = V^2 \left[\frac{1}{(1 + \ln V/Z)^2} - 1 \right]. \tag{14}$$

The expression on the right-hand side of (14) is positive in the interval $\exp(-2Z) < V < 1$. Only for sufficiently small Z is $H(V)$, given by (14), monotonic, so that $V(Z)$ is single valued. For $Z \gg 1$, which is typical for solid combustion, under the physically reasonable condition $V \gg \exp(-Z)$, formula (14) can be written to leading order as

$$H = -\frac{2V^2 \ln V}{Z}. \tag{15}$$

Equation (15) has *two* solution branches $V(H)$ if $H < H_m = \frac{1}{eZ}$ and *no* solutions if $H > H_m$ (see Figure 2). At the value $H = H_m$, which corresponds to the extinction limit, the two solution branches merge at $V = V_m = 1/\sqrt{e}$, corresponding to $H = H_m$, and disappear for larger H. It will be shown below that the solution with the higher value of V is monotonically stable, while the solution with the lower value of V is monotonically unstable, in agreement with the predictions of [6]. There is a third solution branch of (14), with exponentially small velocity V, as well as a yet smaller fourth solution branch (see Figure 1), and a still smaller fifth solution branch, which are unphysical and are to be ignored. Indeed, the fourth solution branch of (14) is not a solution of (13), since it corresponds to $1 + \frac{\ln V}{Z} < 0$, and the fifth solution branch corresponds to $H < 0$ and is also not a solution of (13) (this branch, which approaches the origin, is not plotted in Figure 1).

After finding the physically relevant value of the steadily propagating front velocity, we can further simplify the description of the problem by introducing the following nondimensional variables: $x = (x_* - \xi_*(0))v_f/\kappa$, $t = t_* v_f^2/\kappa$, $T = (T_* - T_0)/C_{*0}Q$. In these nondimensional variables, the problem reads

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} - \Gamma T - \frac{d\xi}{dt}\delta(x - \xi(t)), \quad -\infty < x < \infty, \quad 0 < t < \infty, \tag{16}$$

$$T(x, 0) = T_i(x), \quad -\infty < x < \infty, \tag{17}$$

$$T(-\infty, t) = T(\infty, t) = 0, \quad 0 \le t < \infty, \tag{18}$$

$$-\frac{d\xi}{dt} = \exp\{Z(T(\xi(t),t) - T_f)\}, \tag{19}$$

where

$$\Gamma = \frac{\gamma\kappa}{v_f^2} > 0 \tag{20}$$

can be considered as the dimensionless heat loss parameter and simultaneously as a dimensionless inverse square interface velocity, and T_f is the dimensionless interface temperature. Note that we have assumed that $\Gamma > 0$, since for $\Gamma = 0$ the boundary condition for $T(\infty, t)$ would not be that given in (18). The particular solution (7), (8), corresponding to an interface uniformly propagating with the velocity v_f, is given by

$$\xi(t) = -t,$$

$$T(x,t) = T_f e^{K_-(x+t)}, \quad x+t<0; \qquad T(x,t) = T_f e^{-K_+(x+t)}, \quad x+t>0, \tag{21}$$

where

$$K_\pm = (\mp 1 + \sqrt{1+4\Gamma})/2; \qquad T_f = \frac{1}{\sqrt{1+4\Gamma}}. \tag{22}$$

We refer to this solution as the basic solution. Equation (13), which reads

$$f(\Gamma) = \Gamma^{1/2} \exp\left(\frac{Z}{\sqrt{1+4\Gamma}}\right) = H^{1/2} \exp Z, \tag{23}$$

determines values of Γ corresponding to given values of Z and H. Extinction occurs at $\Gamma_m = H_m/4V_m^2 = 1/4Z$ (in the limit $Z \gg 1$). Note that the solution with $V > V_m$ has $\Gamma < \Gamma_m$, while the solution with $V < V_m$ has $\Gamma > \Gamma_m$.

3. Linear stability of the basic solution. We rewrite the system (16)–(18) in a reference frame moving with the interface as

$$\frac{\partial T}{\partial t} - \frac{d\xi}{dt}\frac{\partial T}{\partial X} = \frac{\partial^2 T}{\partial X^2} - \Gamma T - \frac{d\xi}{dt}\delta(X), \quad -\infty < X < \infty,\ 0 < t < \infty, \tag{24}$$

$$T(-\infty, t) = T(\infty, t) = 0, \quad 0 \le t < \infty, \tag{25}$$

where $X = x - \xi(t)$. Linearizing equations (24), (25) about the basic solution (21), we obtain the following eigenvalue problem for the infinitesimal temperature perturbation $\tilde{T}(X)\exp(\sigma t)$ and the interface velocity perturbation $\tilde{v}\exp(\sigma t)$

$$\sigma\tilde{T} - \tilde{v}T' + \tilde{T}' = \tilde{T}'' - \Gamma\tilde{T}, \quad X \neq 0, \tag{26}$$

$$\tilde{T}(0^-) = \tilde{T}(0^+) \equiv \tilde{T}(0), \quad \tilde{T}'(0^+) - \tilde{T}'(0^-) - \tilde{v} = 0, \tag{27}$$

$$\tilde{T}(\pm\infty) = 0. \tag{28}$$

Linearization of (19) provides a relation between the velocity perturbation $\tilde{v}$ and the interface temperature perturbation $\tilde{T}(0)$,

$$\tilde{v} = -Z\tilde{T}(0). \tag{29}$$

We are interested in two types of stability boundaries, the monotonic boundary corresponding to $\sigma = 0$ and the oscillatory boundary corresponding to $\sigma = i\omega$, with ω real.

If $\sigma = 0$, the temperature perturbations are given by

$$\tilde{T} = e^{K_- X} + \frac{ZT_f K_-}{\sqrt{1+4\Gamma}} X e^{K_- X}, \quad X < 0, \tag{30}$$

$$\tilde{T} = e^{-K_+ X} + \frac{ZT_f K_+}{\sqrt{1+4\Gamma}} X e^{-K_+ X}, \quad X > 0, \tag{31}$$

and the expression for the monotonic stability boundary $Z = Z_m(\Gamma)$, i.e., the curve which separates stable from unstable basic solutions, is given by

$$Z = Z_m(\Gamma) = \frac{(1+4\Gamma)^{3/2}}{4\Gamma}. \tag{32}$$

The function $Z_m(\Gamma)$ has a minimum, given by $Z_c = 3^{3/2}/8$. Note that our approach is valid only for Z sufficiently large and Γ sufficiently small, hence only the left branch of the curve (32) has physical meaning. The curve (32), which separates stable and unstable solutions, corresponds to the extinction limit [6]. Inverting $Z = Z_m(\Gamma)$ determines the extinction line $\Gamma = \Gamma_m(Z)$ discussed in the previous section. As mentioned above, for a fixed value of Z, solutions with a value of Γ smaller than the critical value $\Gamma_m(Z)$, i.e., those corresponding to the larger of the two interface velocities v_f, are stable (with respect to monotonic disturbances), while solutions with a value of Γ larger than the critical value, i.e., those corresponding to the smaller of the two interface velocities, are unstable, in agreement with the predictions of [6].

If $\sigma = i\omega$, the eigenfunction of the problem (26)–(29) has the form

$$\tilde{T} = e^{k_- X} - \frac{iZT_f K_+}{\omega}\left(e^{-K_+ X} - e^{-k_+ X}\right), \quad X > 0, \tag{33}$$

where

$$k_\pm = \frac{\mp 1 + \sqrt{1+4\Gamma+4i\omega}}{2}. \tag{34}$$

The oscillatory stability boundary $Z_o(\Gamma)$ and the frequency ω are obtained from the real and imaginary parts of the dispersion relation

$$Z_o\{\omega + iT_f[K_+(K_+ - k_+) - K_-(K_- - k_-)]\} - (k_+ + k_-)\omega = 0. \tag{35}$$

We find that

$$Z_o(\Gamma) = \frac{2 + \sqrt{4 + (1+4\Gamma)^2}}{\sqrt{1+4\Gamma}}, \tag{36}$$

$$\omega^2(\Gamma) = Z_o \frac{1 + 16\Gamma^2 - 4\Gamma\sqrt{4 + (1+4\Gamma)^2}}{(1+4\Gamma)^{3/2}}. \tag{37}$$

The expression for ω^2 is positive only for $\Gamma < \Gamma_*$, where Γ_* is the positive root of the cubic equation

$$1 - 48\Gamma_*^2 - 128\Gamma_*^3 = 0.$$

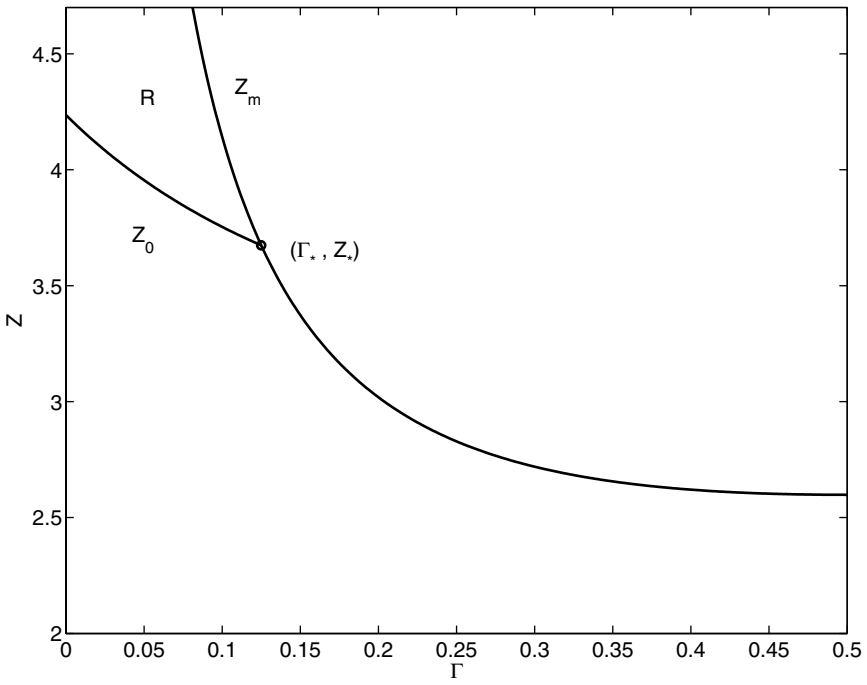


FIG. 3. *Region R in parameter space in which solution behavior is computed.*

The point $(\Gamma_*, Z_* = Z_o(\Gamma_*) = Z_m(\Gamma_*))$ is a codimension-two point where the monotonic and oscillatory boundaries merge. In the interval $0 < \Gamma < \Gamma_*$, $Z_o(\Gamma)$ is a monotonically decreasing function; the curve $Z = Z_o(\Gamma)$ is situated below and to the left of the curve $Z = Z_m(\Gamma)$. We conclude that for a fixed value Z in the interval $Z_* < Z < 2 + \sqrt{5}$, as Γ grows, pulsations precede extinction. The region in (Γ, Z) parameter space in which we compute solution behavior is shown in Figure 3.

4. Integrodifferential equation for the interfacial motion. We now turn to the nonlinear problem (16)–(19). The term with heat losses is eliminated by the transformation

$$T(x,t) = \Theta(x,t)e^{-\Gamma t}. \tag{38}$$

Substituting (38) into (16), we obtain the inhomogeneous heat equation for $\Theta(x,t)$,

$$\frac{\partial \Theta(x,t)}{\partial t} = \frac{\partial^2 \Theta(x,t)}{\partial x^2} - e^{\Gamma t}\frac{d\xi}{dt}\delta(x - \xi(t)), \quad -\infty < x < \infty,\ 0 < t < \infty \tag{39}$$

subject to the initial condition

$$\Theta(x,0) = T_i(x), \quad -\infty < x < \infty, \tag{40}$$

and the boundary conditions

$$\Theta(-\infty,t) = \Theta(\infty,t) = 0, \quad 0 \le t < \infty. \tag{41}$$

The solution of the problem (39)–(41) is given by

$$\Theta(x,t) = \int_{-\infty}^{\infty} \frac{1}{2\sqrt{\pi t}} e^{-(x-y)^2/4t} T_i(y)dy$$
$$+ \int_0^t d\tau \int_{-\infty}^{\infty} dy \frac{1}{2\sqrt{\pi(t-\tau)}} e^{\Gamma\tau} e^{-(x-y)^2/4(t-\tau)} \left[-\frac{d\xi}{d\tau}(\tau)\right] \delta(y-\xi(\tau)). \tag{42}$$

In terms of the variable $T(x,t)$, we obtain

$$T(x,t) = e^{-\Gamma t} \int_{-\infty}^{\infty} \frac{1}{2\sqrt{\pi t}} e^{-(x-y)^2/4t} T_i(y)dy$$
$$+ \int_0^t d\tau \int_{-\infty}^{\infty} dy \frac{1}{2\sqrt{\pi(t-\tau)}} e^{-\Gamma(t-\tau)} e^{-(x-y)^2/4(t-\tau)} \left[-\frac{d\xi}{d\tau}(\tau)\right] \delta(y-\xi(\tau)). \tag{43}$$

Evaluating the integral containing the δ-function and substituting into (19), we find the self-consistency condition for $\xi(t)$,

$$-\frac{d\xi}{dt}(t) = \exp\left\{ Z \left[e^{-\Gamma t} \int_{-\infty}^{\infty} \frac{1}{2\sqrt{\pi t}} e^{-(\xi(t)-y)^2/4t} T_i(y)dy \right.\right.$$
$$\left.\left. + \int_0^t d\tau \frac{e^{-\Gamma(t-\tau)}}{2\sqrt{\pi(t-\tau)}} e^{-(\xi(t)-\xi(\tau))^2/4(t-\tau)} \left[-\frac{d\xi}{d\tau}(\tau)\right] - T_f \right] \right\}, \tag{44}$$

where T_f is the undisturbed front temperature determined by (22). This is a closed integrodifferential equation for the function $\xi(t)$ defined for $t > 0$.

The first term on the right-hand side of (44), which depends on the initial conditions, decays in time. Thus, for large t it can be disregarded. Employing the change of variable $\tau = t - s$, the second integral is transformed to

$$\int_0^t ds \frac{e^{-\Gamma s}}{2\sqrt{\pi s}} e^{-(\xi(t)-\xi(t-s))^2/4s} \left[-\frac{d\xi(t-s)}{dt}\right].$$

Thus, we finally obtain that $\xi(t)$ is the solution of

$$\frac{d\xi}{dt}(t) = -\exp\left\{ -Z \left[T_f + \int_0^t ds \frac{e^{-\Gamma s}}{2\sqrt{\pi s}} e^{-(\xi(t)-\xi(t-s))^2/4s} \left[\frac{d\xi(t-s)}{dt}\right] \right] \right\}. \tag{45}$$

This integrodifferential equation describes the location of the interface $\xi(t)$ separating the fresh fuel from the burned products, for large times, after transients have died out.

5. Numerical method. We now describe the numerical method employed in the solution to (45). We note that we solve the initial value problem for (45) and march forward in time until steady state is achieved. As a result, all steady state solutions that we compute are necessarily stable. For conciseness, we introduce the symbol $\dot{\xi} = \frac{d\xi}{dt}$ to denote the front velocity. Note that we have assumed that the front propagates in the negative x direction so that $\dot{\xi} < 0$ for all t. We first take the natural logarithm of both sides of (45) to obtain

$$\ln(-\dot{\xi}(t)) = Z \left\{ -T_f + \int_0^t ds \frac{e^{-\Gamma s}}{2\sqrt{\pi s}} e^{-(\xi(t)-\xi(t-s))^2/4s} \left(-\dot{\xi}(t-s)\right) \right\}. \tag{46}$$

We next introduce the integration variable $\sigma = t - s$, so that (46) becomes

$$\ln(-\dot{\xi}(t)) = Z\left\{-T_f + \int_0^t d\sigma \frac{e^{-\Gamma(t-\sigma)}}{2\sqrt{\pi(t-\sigma)}} e^{-(\xi(t)-\xi(\sigma))^2/4(t-\sigma)} \left(-\dot{\xi}(\sigma)\right)\right\}. \tag{47}$$

Equation (47) will be solved on a nonuniform grid t_i with $t_0 = 0$. The grid will be chosen adaptively based on the size of $-\dot{\xi}(t)$ as described below. In order to solve for $\dot{\xi}(t_n)$ we split the integral in (47) into two parts,

$$\begin{aligned}\int_0^{t_n} d\sigma \frac{F(t_n,\sigma)}{2\sqrt{\pi(t_n-\sigma)}} &= \int_{t_{n-1}}^{t_n} d\sigma \frac{F(t_n,\sigma)}{2\sqrt{\pi(t_n-\sigma)}} + \int_0^{t_{n-1}} d\sigma \frac{F(t_n,\sigma)}{2\sqrt{\pi(t_n-\sigma)}} \\ &= I_n + S,\end{aligned} \tag{48}$$

where

$$F(t_n,\sigma) = e^{-\Gamma(t_n-\sigma)} e^{-(\xi(t_n)-\xi(\sigma))^2/4(t_n-\sigma)} \left(-\dot{\xi}(\sigma)\right),$$

and

$$I_n = \int_{t_{n-1}}^{t_n} d\sigma \frac{F(t_n,\sigma)}{2\sqrt{\pi(t_n-\sigma)}}, \qquad S = \int_0^{t_{n-1}} d\sigma \frac{F(t_n,\sigma)}{2\sqrt{\pi(t_n-\sigma)}}.$$

We approximate I_n by taking the average value of $F(t_n,\sigma)$ over the domain of integration ($t_{n-1} \le \sigma \le t_n$) to obtain

$$I_n \approx \frac{1}{2\sqrt{\pi}} \left[\frac{F(t_n,t_n) + F(t_n,t_{n-1})}{2}\right] \int_{t_{n-1}}^{t_n} \frac{d\sigma}{\sqrt{t_n-\sigma}}. \tag{49}$$

The computation of $F(t_n,t_{n-1})$ presents no difficulty but $F(t_n,t_n)$ must be treated in the limit as $\sigma \to t_n$,

$$\lim_{\sigma \to t_n} F(t_n,\sigma) = -\dot{\xi}(t_n)$$

so that

$$I_n \approx \frac{1}{2}\sqrt{\frac{h_n}{\pi}} \left\{-\dot{\xi}(t_n) + F(t_n,t_{n-1})\right\}, \tag{50}$$

where $h_n = t_n - t_{n-1}$. Substituting (50) into (47) yields

$$\ln\left(-\dot{\xi}(t_n)\right) = \frac{Z}{2}\sqrt{\frac{h_n}{\pi}} \left(-\dot{\xi}(t_n)\right) + Z\left\{-T_f + S_{n-1} + S\right\}, \tag{51}$$

where

$$S_{n-1} = \frac{1}{2}\sqrt{\frac{h_n}{\pi}} F(t_n,t_{n-1}).$$

Equation (51) is a transcendental equation of the form

$$\ln\left(-\dot{\xi}(t_n)\right) = m\left(-\dot{\xi}(t_n)\right) + b, \tag{52}$$

where

$$m = \frac{Z}{2}\sqrt{\frac{h_n}{\pi}}, \qquad b = Z\left\{-T_f + S_{n-1} + S\right\},$$

which must be solved at each timestep. We note that m is proportional to $\sqrt{h_n}$, the local timestep. For clarity we rewrite (52) as

$$H(w) = mw - \ln w + b = 0, \tag{53}$$

where $w = -\dot{\xi}(t_n)$. It is easy to see that $H(w)$ is minimized when $w = 1/m$ and that the value of the minimum is $1 + b + \ln m$. Thus, if

$$\ln m > -(1 + b),$$

there is no solution to (52) and the computation cannot continue. This typically occurs when $-\dot{\xi}$ is large and necessitates a reduction in the timestep.

Finally, we approximate S over the interval $0 \le \sigma \le t_{n-1}$ using the trapezoidal rule with nonuniform grid spacing,

$$S = \frac{1}{2\sqrt{\pi}} \int_0^{t_{n-1}} \frac{d\sigma}{\sqrt{t_n - \sigma}} F(t_n, \sigma) \approx \frac{1}{2\sqrt{\pi}} \sum_{i=1}^{n-1} h_i \frac{1}{2} \left(\frac{F(t_n, t_i)}{\sqrt{t_n - t_i}} + \frac{F(t_n, t_{i-1})}{\sqrt{t_n - t_{i-1}}} \right). \tag{54}$$

We have found that the exponentially decaying tail of the integral in S can be truncated with minimal loss of accuracy but significant savings in computational cost. We compute the sum in (54) starting with $i = n - 1$ and continuing to lower i until the relative change due to taking one additional term is less than 10^{-16}.

The solution is typically highly relaxational, with long periods of nearly quiescent behavior juxtaposed with brief periods of rapid variation, which are manifested by spikes in the front speed over very short time intervals. In order to address the relaxational behavior as well as to insure that the solvability condition described above for (52) will be satisfied, it is necessary to employ a timestep which varies in t according to the nature of the solution. Clearly, h_n should decrease as $|\dot{\xi}(t_n)|$ increases. We employed the following empirical formula to adjust the local timestep,

$$h_{n+1} = \min \left(\Delta t_0, \left(\frac{2\sqrt{\pi}}{Z \; A(-\dot{\xi}(t_n)^{3/2}} \right)^2 \right), \tag{55}$$

where A is a constant. The values $\Delta t_0 = 0.01$, as well as $A = 17$ for $Z = 4.12$ and $A = 12$ for $Z = 4.0$, were employed in the calculations presented in this paper. Finally, after solving for $\dot{\xi}$ at time t_n, ξ must be updated to t_{n+1}. This is done using the Adams–Bashforth method with nonuniform spacing,

$$\xi(t_{n+1}) = \xi(t_n) + h_{n+1} \left\{ \dot{\xi}(t_n) + \frac{1}{2} \frac{h_{n+1}}{h_n} \left(\dot{\xi}(t_n) - \dot{\xi}(t_{n-1}) \right) \right\}.$$

The initial condition for ξ is irrelevant since only differences in ξ appear in (45).

It is characteristic of the pulsating solutions in solid fuel combustion that they become increasingly relaxational in both space and time as parameters are pushed further into the nonlinear regime. Thus, a typical solution will propagate slowly for a long period of time exhibiting relatively gradual spatial and temporal variations in temperature, followed by a short period of rapid spatial and temporal variation. Therefore, computations require adaptive procedures in both space and time. The present method eliminates the spatial coordinate by restricting consideration to the motion of the interface. Thus, there is no longer a need for spatial adaptivity. Temporal adaptivity is still required as described above. Indeed, the relaxational behavior

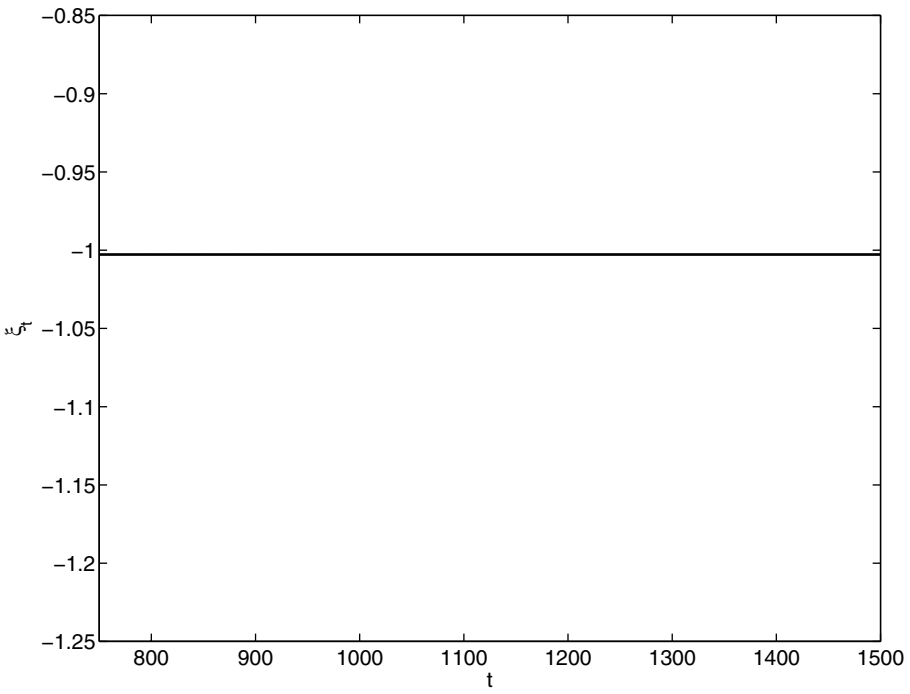


FIG. 4. *Uniformly propagating solution,* $Z = 4.12$, $\Gamma = 0.01$.

of the solution is enhanced, presumably due to the reaction sheet model employed for the kinetics rather than the distributed Arrhenius kinetics typically employed when the problem is solved as a system of partial differential equations. The cost of the present method is clearly in storage as the solution should in principle be stored for all values of time, due to the convolution nature of the integral in (45). We have addressed this issue to some extent by truncating the tail of the integral in S, as described above. Thus, employing the integrodifferential equation appears to be effective for the problem considered here. We note however, that the storage problem becomes more acute as Z is increased, since smaller timesteps are required. Finally, we note that our computations have been validated under timestep refinement.

6. Results. We employ the integrodifferential equation (45) governing the motion of the interface separating the fresh fuel and the burned products in a nonadiabatic solid flame to compute solutions in the form of a period doubling cascade as the heat loss Γ is increased toward extinction. We consider the Zeldovich number $Z = 4.12$, below the adiabatic ($\Gamma = 0$) pulsating stability limit $Z = 4.22$. Thus, for $\Gamma = 0$, the solution is uniformly propagating. In Figures 4–10 we plot the interface velocity $\dot{\xi}$ as a function of t for a sequence of increasing values of Γ.

Figure 4 shows a uniformly propagating solution for $\Gamma = 0.01$. Note that the computed value of ξ_t differs from the exact value $\xi_t = -1$ by less than 0.5%. In Figure 5 we show a $1T$ solution obtained for $\Gamma = 0.04$. In Figure 6 we show a $2T$ solution for $\Gamma = 0.053$. It is known that for period doubling cascades the intervals for each period doubling become progressively smaller. While we have not attempted to delineate these intervals, we show $4T$, $8T$ and $16T$ solutions in Figures 7–9. The corresponding values of Γ are $\Gamma = .05353$ ($4T$), $\Gamma = .053545$ ($8T$), and $\Gamma = .0535475$ ($16T$). We have not found $32T$ or higher period solutions. We assume that this is

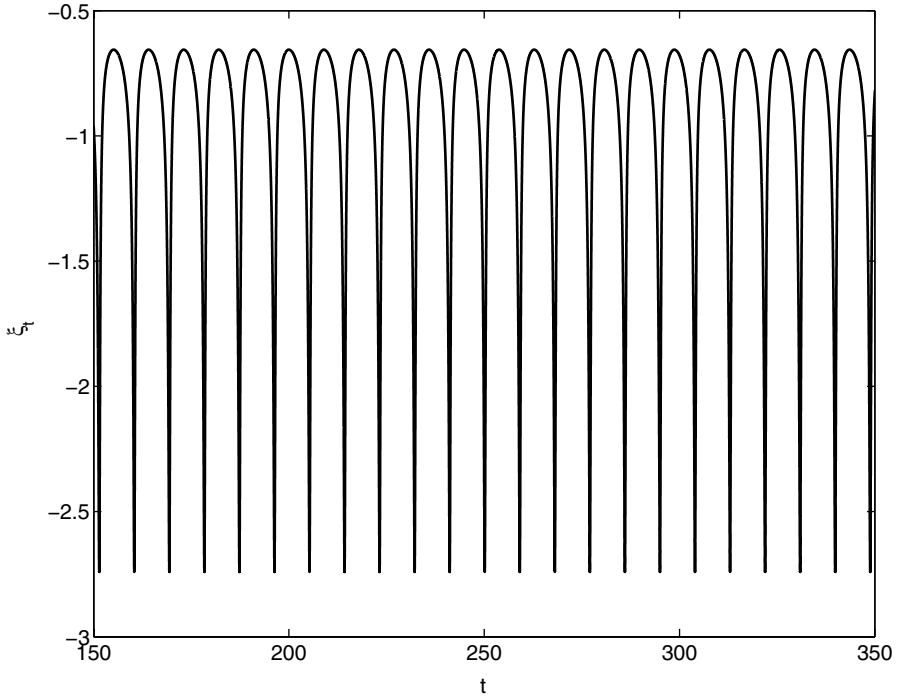


FIG. 5. $1T$ *solution,* $Z = 4.12$, $\Gamma = 0.04$.

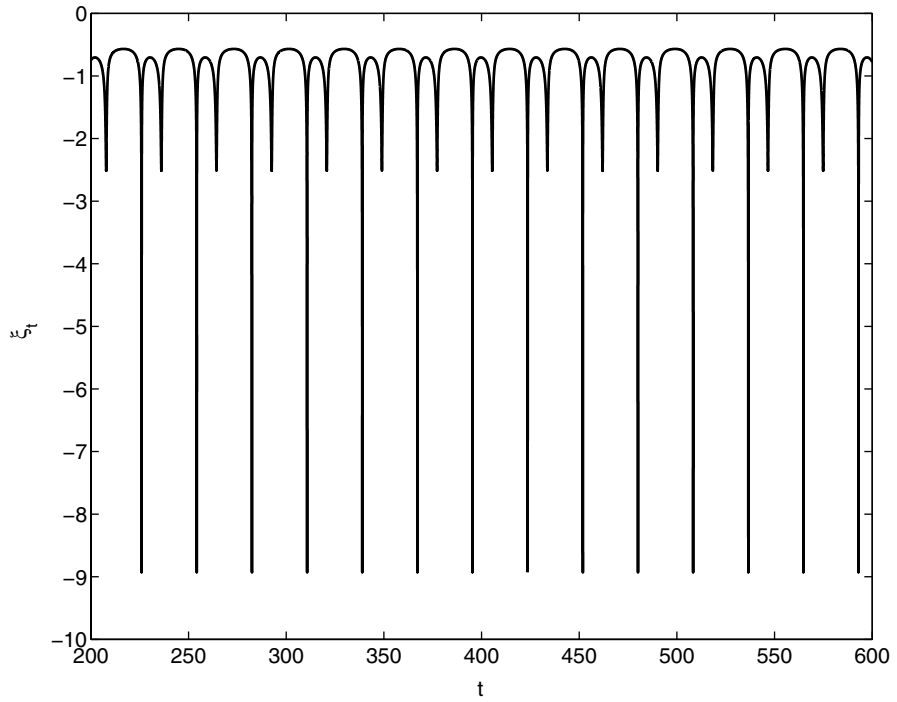


FIG. 6. $2T$ *solution,* $Z = 4.12$, $\Gamma = 0.053$.

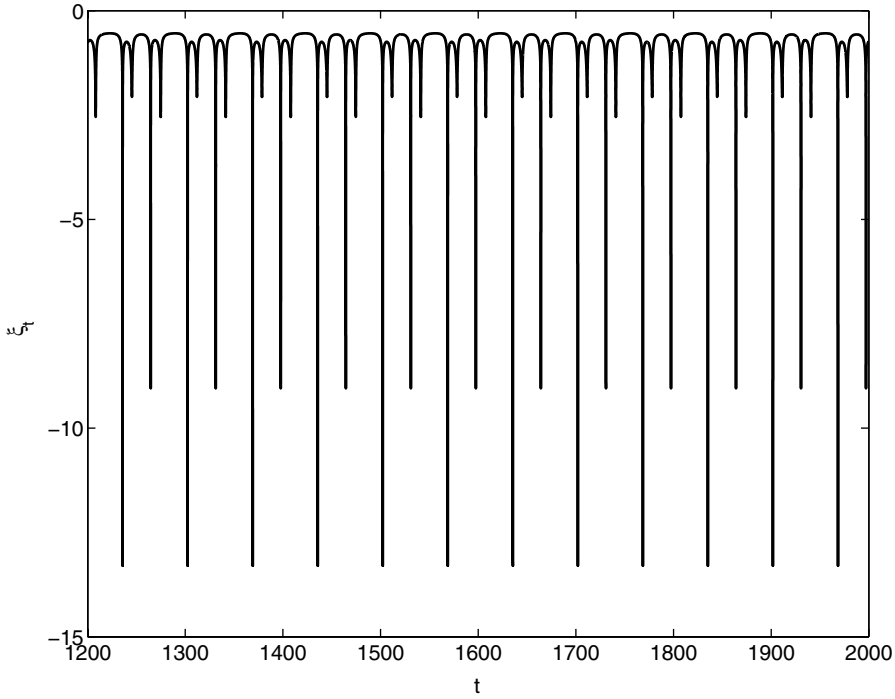


FIG. 7. $4T$ *solution,* $Z = 4.12$, $\Gamma = 0.05353$.

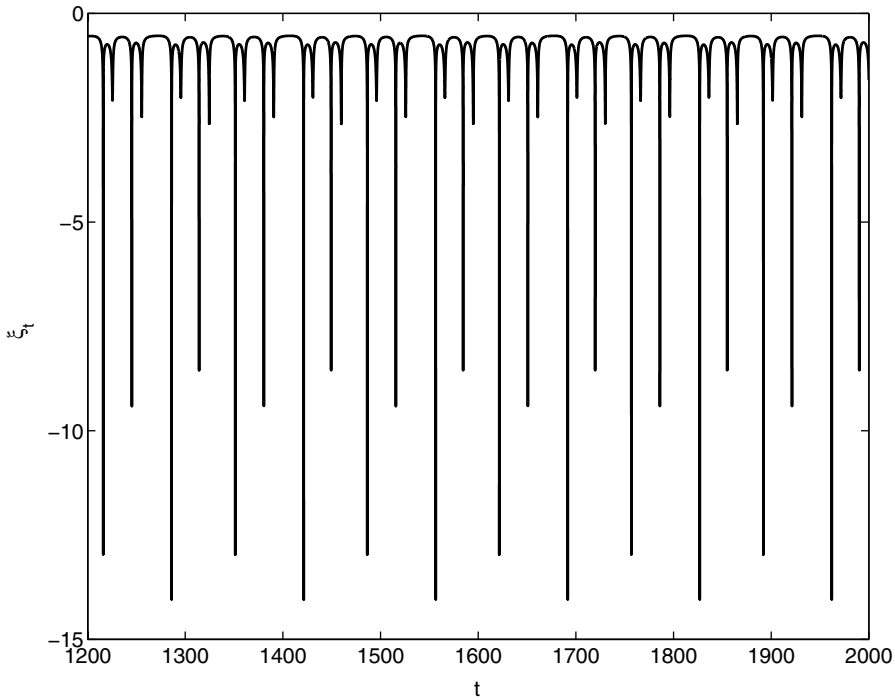


FIG. 8. $8T$ *solution,* $Z = 4.12$, $\Gamma = 0.053545$.

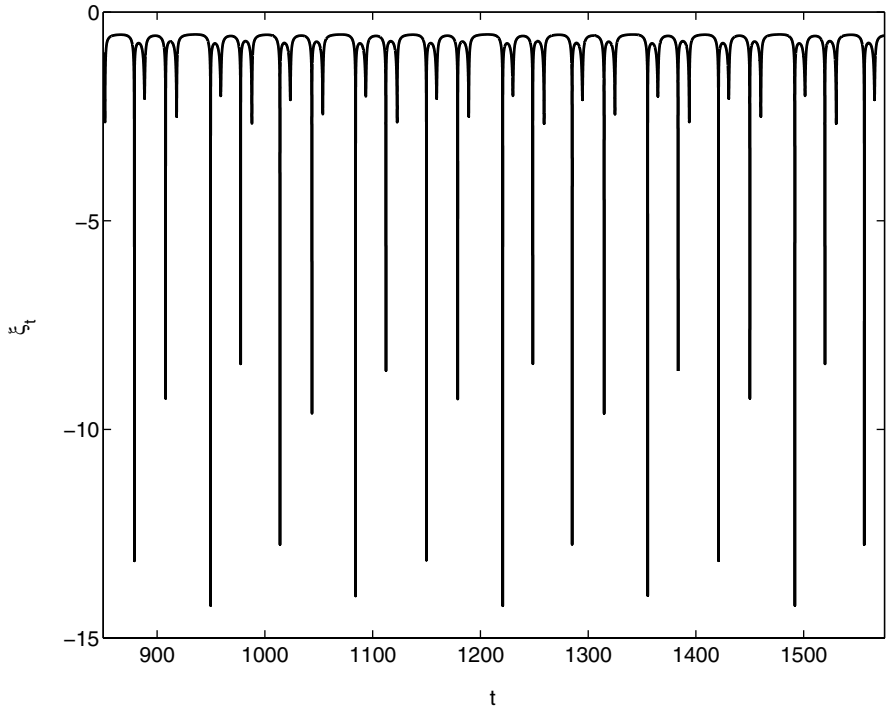


FIG. 9. 16T *solution,* $Z = 4.12$, $\Gamma = 0.0535475$.

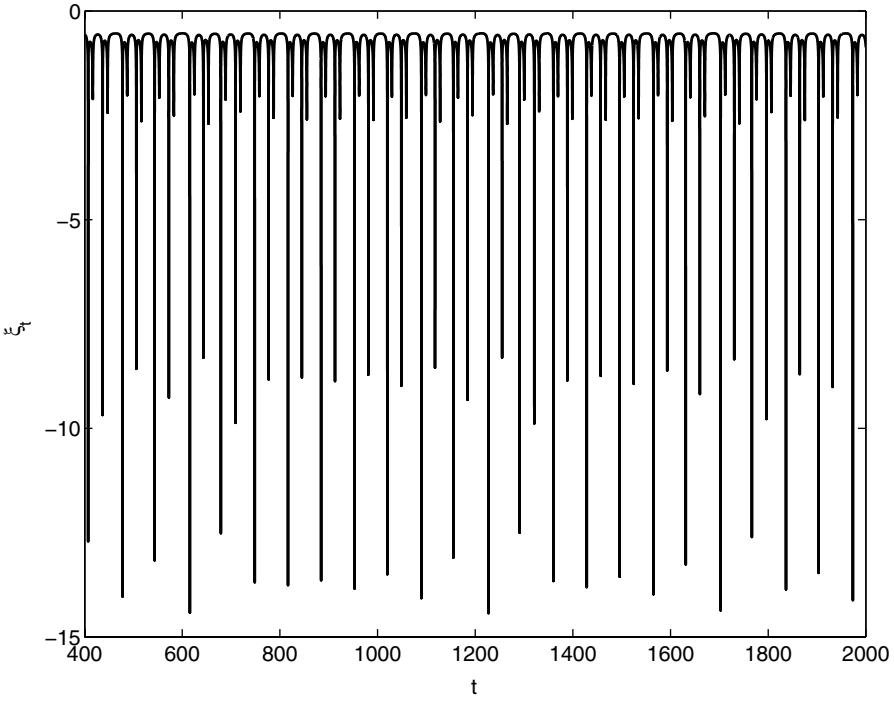


FIG. 10. *Apparently chaotic solution,* $Z = 4.12$, $\Gamma = 0.05355$.

due to the small intervals of existence of these solutions. We note the similarity of these figures to those obtained for adiabatic combustion in the pulsating regime as Z is increased (e.g., [3]). The solutions are highly relaxational with long periods of slow propagation interrupted by sharp spikes (in this case large negative velocities) corresponding to rapid propagation. The relaxational nature of the solution, i.e., the time duration of the spikes, becomes progressively smaller as Γ increases. We note that this relaxational behavior is not related to the period doubling cascade. Similar behavior was observed in [3] for a model with melting where the transition to chaos was via intermittency. The different dynamical behaviors are manifested by variations in the amplitudes of the spikes.

In Figure 10 we show an apparently chaotic solution for $\Gamma = 0.05355$ beyond the period doubling cascade. For period doubling cascades, windows of laminar behavior exist within the chaotic region, e.g., [18]. In Figures 11 and 12 we show the evolution in time and the return map, respectively, for the $12T$ solution found for $\Gamma = 0.0535522$. In Figure 13 we show the return map for the $24T$ solution found for $\Gamma = 0.05355236$. In Figures 14 and 15 we show blowups of the upper left and lower right regions of Figure 13. In Figure 16 we show the return map for the $48T$ solution, found for $\Gamma = 0.0535524$. In Figures 17 and 18 we show blowups of the upper left and lower right regions of Figure 16. Finally, in Figure 19 we show a chaotic solution for $\Gamma = 0.053553$, illustrating a return to chaos after the window of periodic solutions. We note that for this value of Z the extinction limit Γ_e lies in the interval $0.053555 \leq \Gamma_e \leq 0.053557$.

For smaller values of Z, though there are larger intervals in Γ prior to extinction (since $\Gamma_e = O(1/Z)$ is larger), the range of dynamical behavior is more limited since smaller Z is less unstable. Thus, we do not necessarily get a full period doubling

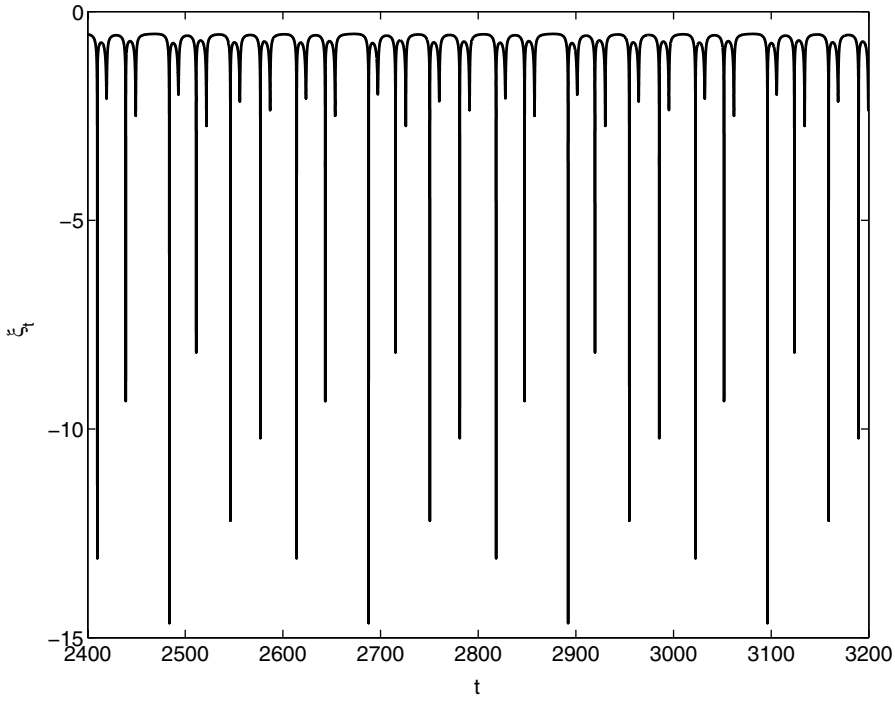


FIG. 11. $12T$ *solution,* $Z = 4.12$, $\Gamma = 0.0535522$.

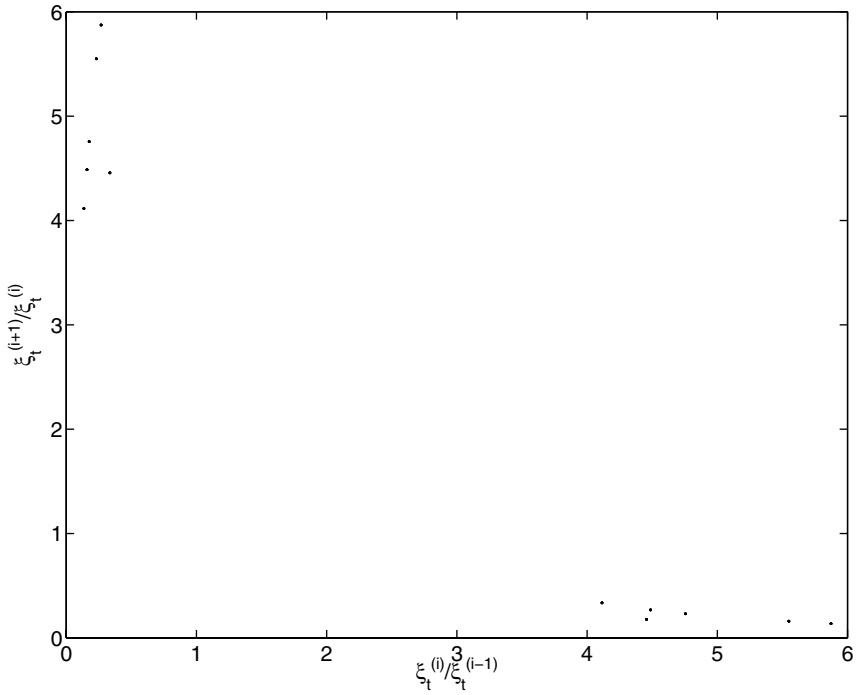


FIG. 12. *Return map for* $12T$ *solution.*

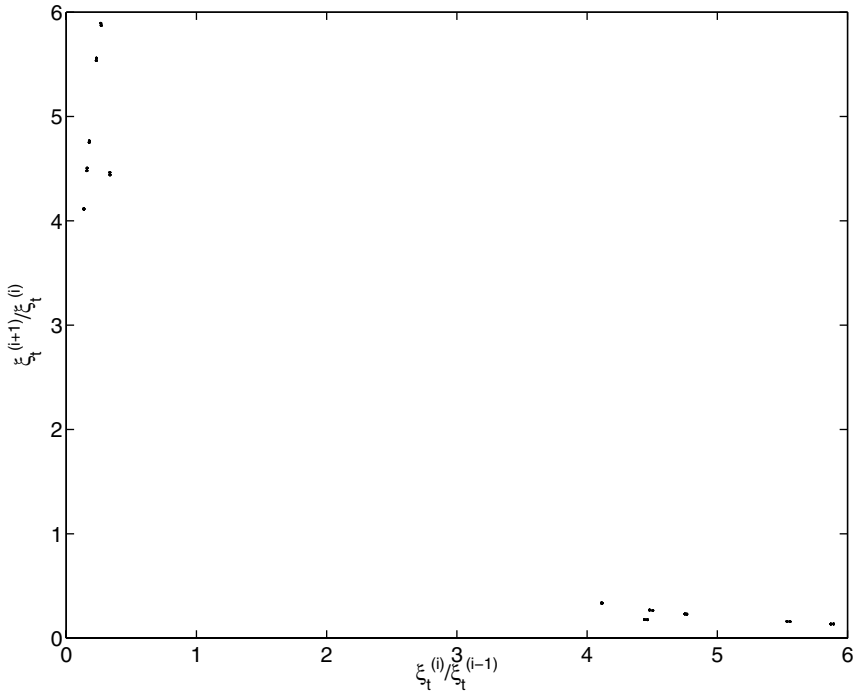


FIG. 13. *Return map for* $24T$ *solution,* $Z = 4.12$, $\Gamma = 0.05355236$.

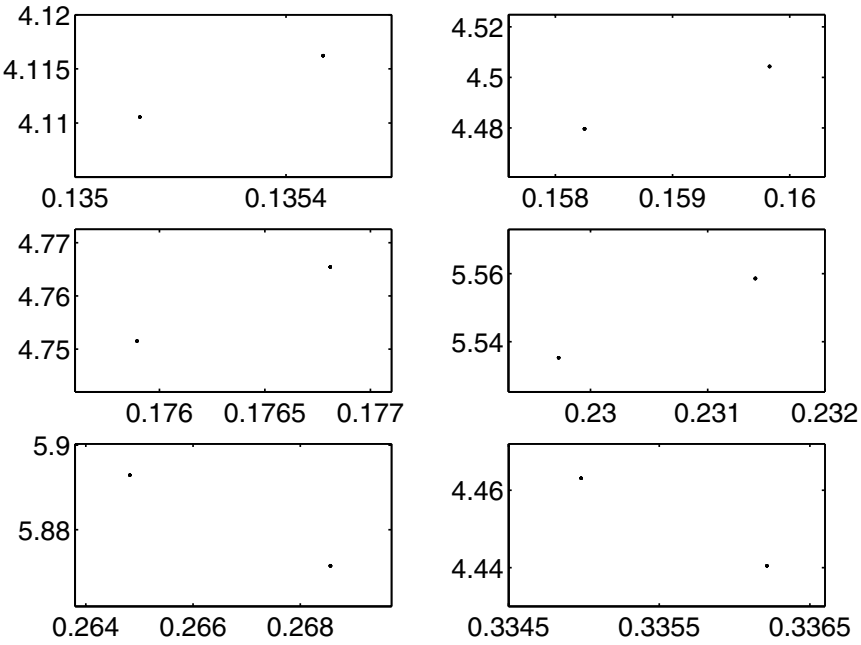


FIG. 14. *Blowup of upper left portion of return map for 24T solution.*

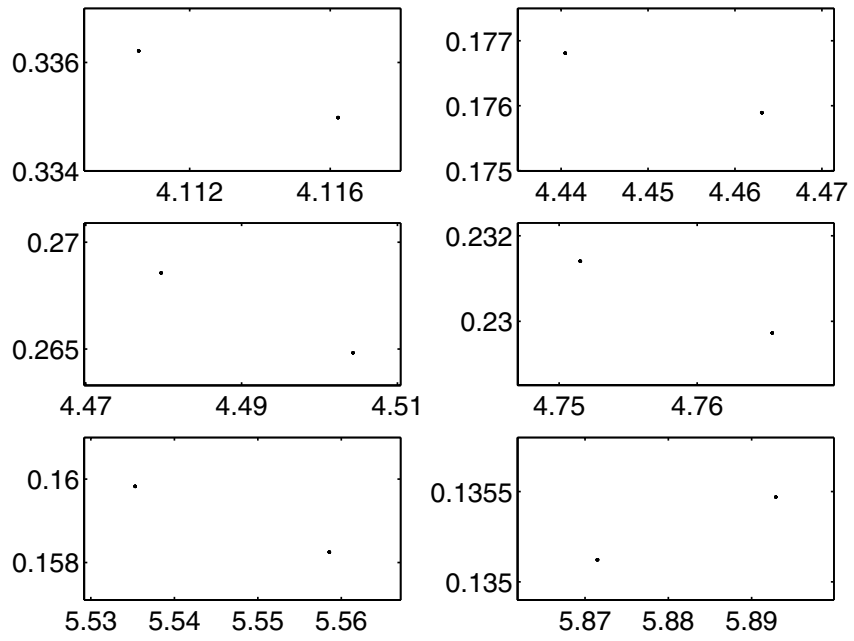


FIG. 15. *Blowup of lower right portion of return map for 24T solution.*

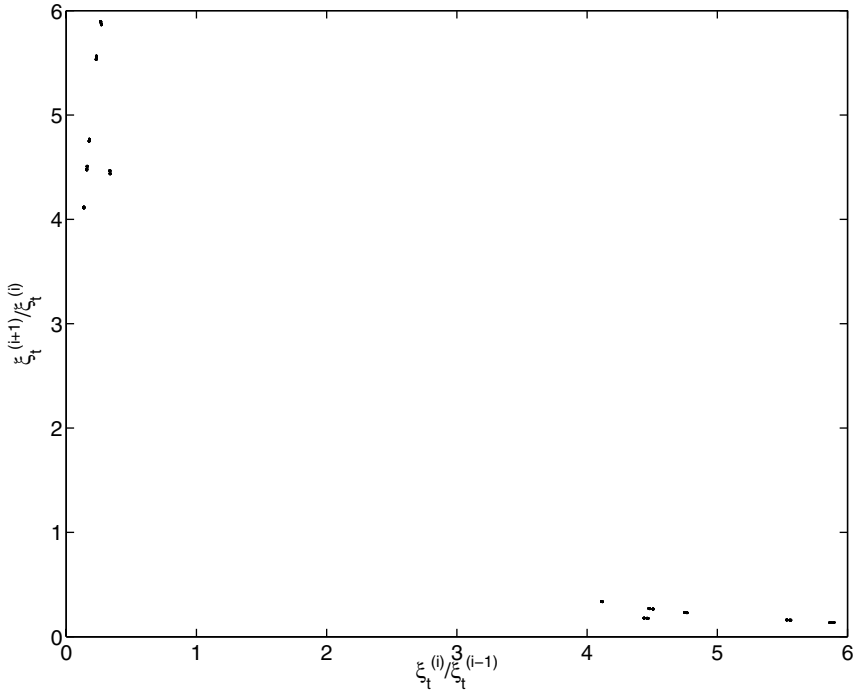


FIG. 16. *Return map for* $48T$ *solution,* $Z = 4.12$, $\Gamma = 0.0535524$.

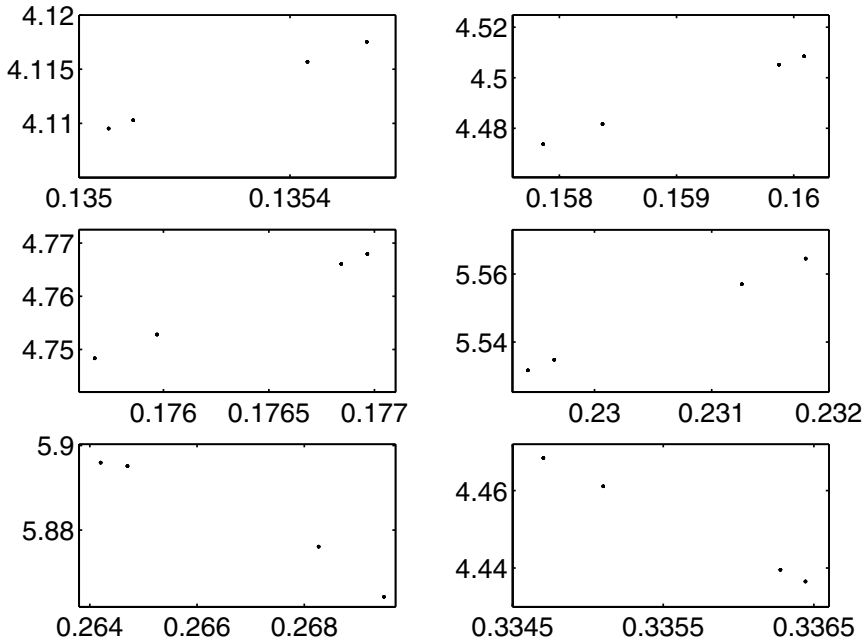


FIG. 17. *Blowup of upper left portion of return map for* $48T$ *solution.*

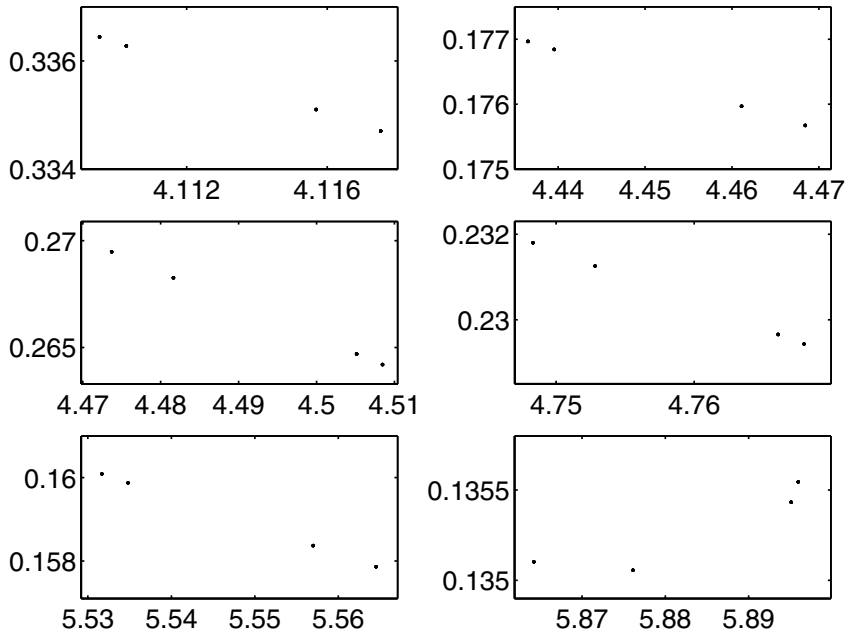


FIG. 18. *Blowup of lower right portion of return map for* $48T$ *solution.*

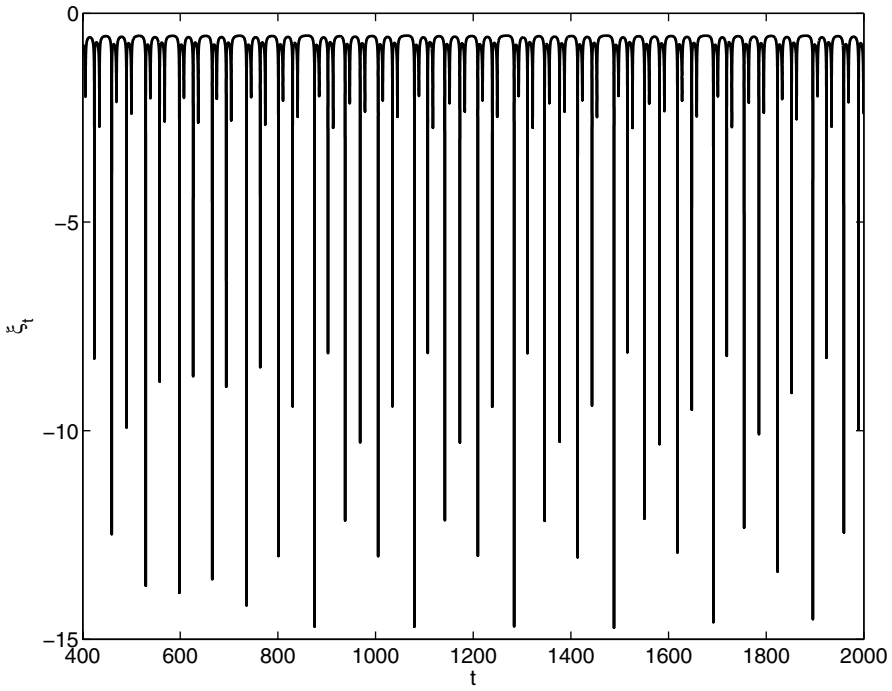


FIG. 19. *Apparently chaotic solution,* $Z = 4.12$, $\Gamma = 0.053553$.

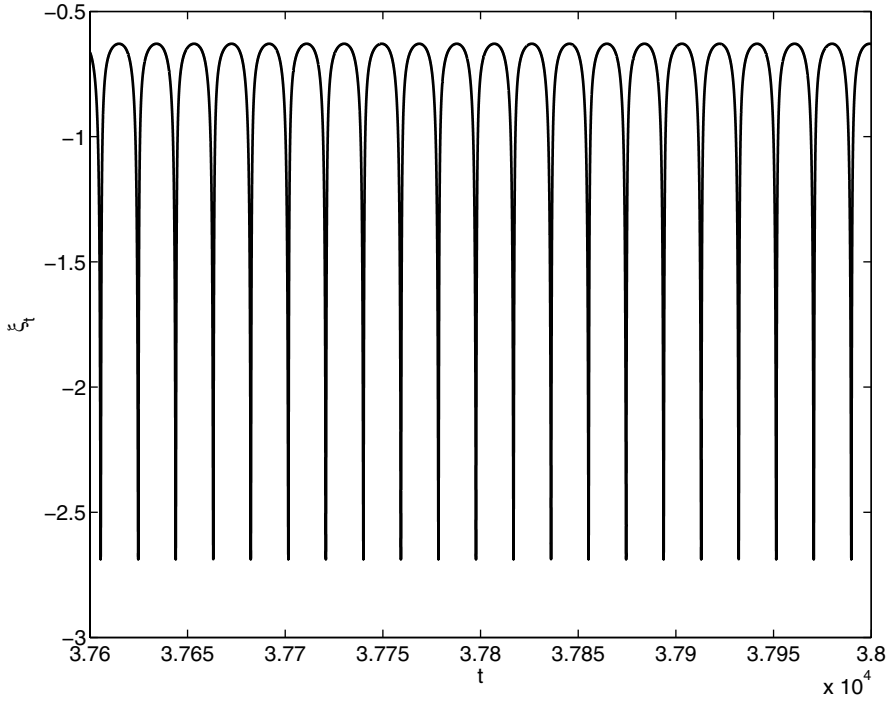


FIG. 20. $1T$ *solution,* $Z = 4.0$, $\Gamma = 0.063453$.

sequence followed by chaos. Rather, the sequence may be truncated after a finite number of periods. An extreme example of this occurs for $Z = 4.0$, where we find only $1T$ solutions for the full range of Γ that we have investigated up to extinction. In Figure 20 we plot the $1T$ solution for $\Gamma = 0.063453$, and in Figure 21 we show that for $\Gamma = 0.0634535$, the system attempts to establish a $2T$ solution. However, the $2T$ solution is apparently unstable, as the two spike amplitudes begin to diverge from one another and eventually the solution becomes extinct.

7. Summary. We derived an integrodifferential equation for the position of the interface in one dimensional nonadiabatic solid fuel combustion from the reaction sheet model. The equation permits computation of the dynamics of the reaction front without the necessity of spatial discretization for the temperature and mass fraction profiles thereby affording a savings in computational resources.

We computed solution behavior describing complex dynamics when the Zeldovich number Z is below the pulsating stability boundary. The control parameter employed is the heat loss coefficient Γ. For values of Z near the pulsating stability boundary we find that a cascade of period doublings occur as Γ increases, leading to chaotic behavior prior to extinction. Specifically, we compute T, $2T$, $4T$, $8T$ and $16T$ solutions as Γ increases. Beyond the $16T$ solution, we find a window of apparently chaotic behavior prior to extinction. Within the chaotic window, we find a subwindow corresponding to a period doubling sequence beginning with a $12T$ solution. Specifically, we compute $12T$, $24T$, and $48T$ solutions. Finally, for smaller values of Z, though there are larger intervals in Γ prior to extinction, the range of dynamical behavior is more limited since smaller Z is less unstable. Thus, we do not necessarily get a full period doubling sequence followed by chaos. Rather, the sequence may be truncated after a finite

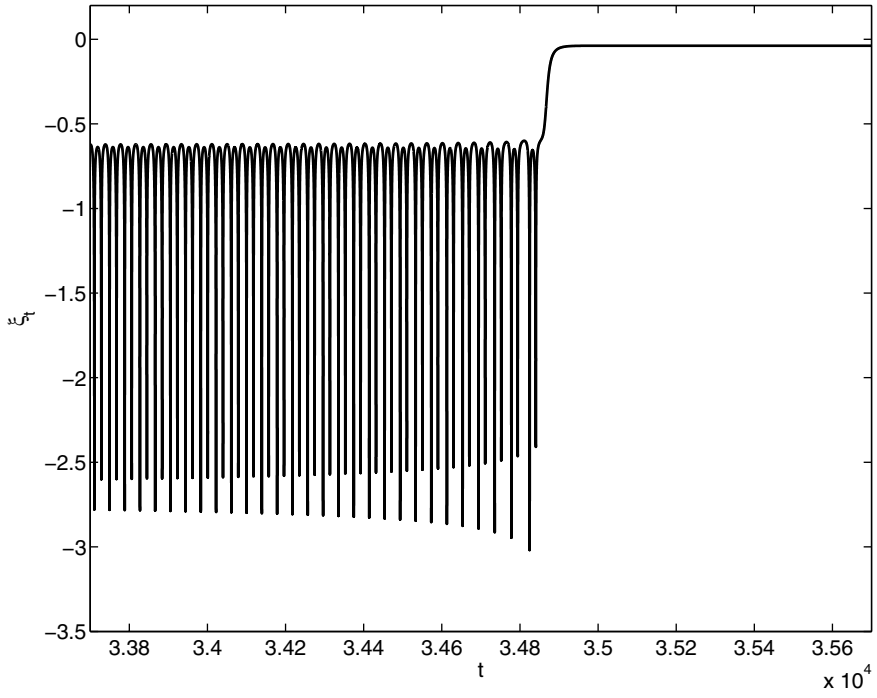


FIG. 21. *Evolution to extinction,* $Z = 4.0$, $\Gamma = 0.0634535$.

number of periods. In an extreme example of this, we find only $1T$ solutions for the full range of Γ that we have investigated up to extinction.

Acknowledgments. We are pleased to acknowledge useful conversations with Prof. R. Clark Robinson.

REFERENCES

[1] A. P. ALDUSHIN, T. M. MARTEMYANOVA, A. G. MERZHANOV, B. I. KHAIKIN, AND K. G. SHKADINSKY, *Autooscillatory propagation of combustion front in heterogeneous condensed media*, Combust. Explosion and Shock Waves, 9 (1973), pp. 531–539.

[2] A. BAYLISS AND B. J. MATKOWSKY, *Fronts, relaxation oscillations, and period doubling in solid fuel combustion*, J. Comput. Phys., 71 (1987), pp. 147–168.

[3] A. BAYLISS AND B. J. MATKOWSKY, *Two routes to chaos in condensed phase combustion*, SIAM J. Appl. Math., 50 (1990), pp. 437–459.

[4] P. DIMITRIOU, J. PUSZYNSKI, AND V. HLAVACEK, *On the dynamics of equations describing gasless combustion*, Combust. Sci. Tech., 68 (1989), pp. 101–111.

[5] M. FRANKEL, V. ROYTBURD, AND G. I. SIVASHINSKY, *Complex dynamics generated by a sharp interface model of self-propagating high-temperature synthesis*, Combust. Theory Model., 2 (1998), pp. 479–496.

[6] H. G. KAPER, G. K. LEAF, S. B. MARGOLIS, AND B. J. MATKOWSKY, *On nonadiabatic condensed phase combustion*, Combust. Sci. Tech., 53 (1987), pp. 289–314.

[7] V. B. LIBROVICH AND G. M. MAKHVILADZE, *One limiting scheme for the propagation of a pulsating exothermic reaction front in a condensed medium*, J. Appl. Mech. Tech. Phys., 15 (1974), p. 818.

[8] S. B. MARGOLIS, *An asymptotic theory of condensed two-phase flame propagation*, SIAM J. Appl. Math., 43 (1983), pp. 351–369.

[9] B. J. MATKOWSKY AND G. I. SIVASHINSKY, *Propagation of a pulsating reaction front in solid fuel combustion*, SIAM J. Appl. Math., 35 (1978), pp. 465–478.

[10] B. J. Matkowsky and G. I. Sivashinsky, *An asymptotic derivation of two models in flame theory associated with the constant density approximation*, SIAM J. Appl. Math., 37 (1979), pp. 686–699.

[11] G. N. Mercer, R. O. Weber, and H. S. Sidhu, *An oscillatory route to extinction for solid fuel combustion waves due to heat losses*, Proc. R. Soc. Lond. Ser. A Math. Phys. Engrg. Sci., 454 (1998), pp. 2015–2022.

[12] A. G. Merzhanov, *SHS processes: Combustion theory and practice*, Arch. Combustionis, 1 (1981), pp. 23–48.

[13] A. G. Merzhanov, *Self-propagating high-temperature synthesis: Twenty years of search and findings*, in Combustion and Plasma Synthesis of High-Temperature Materials, Z. A. Munir and J. B. Holt, eds., VCH, Weinheim, Berlin, (1990), pp. 1–53.

[14] A. G. Merzhanov, A. K. Filonenko, and I. P. Borovinskaya, *New phenomena in combustion of condensed waves*, Soviet Phys. Dokl., 208 (1973), pp. 122–125.

[15] Z. A. Munir and U. Anselmi-Tamburini, *Self-propagating exothermic reactions: The synthesis of high-temperature materials by combustion*, Mater. Sci. Reports, 3 (1989), pp. 277–365.

[16] C. Raymond, A. Bayliss, B. J. Matkowsky, and V. Volpert, *Transitions to chaos in condensed phase combustion with reactant melting*, Intl. J. Self-Propagating High-Temperature Synthesis, 10 (2001), pp. 133–139.

[17] K. G. Shkadinsky, B. I. Khaikin, and A. G. Merzhanov, *Propagation of a pulsating exothermic reaction front in the condensed phase*, Combust. Explosion and Shock Waves, 1 (1971), pp. 15–22.

[18] S. H. Strogatz, *Nonlinear Dynamics and Chaos*, Westview Press, Cambridge, MA, 1994.

SIAM J. APPL. MATH.
Vol. 66, No. 3, pp. 896–920

A MODEL FOR THE DYNAMICS OF LARGE QUEUING NETWORKS AND SUPPLY CHAINS*

D. ARMBRUSTER†, P. DEGOND‡, AND C. RINGHOFER†

Abstract. We consider a supply chain consisting of a sequence of buffer queues and processors with certain throughput times and capacities. Based on a simple rule for releasing parts, i.e., batches of product or individual product items, from the buffers into the processors, we derive a hyperbolic conservation law for the part density and flux in the supply chain. The conservation law will be asymptotically valid in regimes with a large number of parts in the supply chain. Solutions of this conservation law will in general develop concentrations corresponding to bottlenecks in the supply chain.

Key words. supply chains, conservation laws, asymptotics.

AMS subject classifications. 65N35, 65N05

DOI. 10.1137/040604625

1. Introduction. This paper is concerned with the development and analysis of continuum models for supply chains. We consider a chain of M suppliers or processors $S_0, \ldots, S_{M-1}$. In the generic picture of a supply chain (see cf. [12] for an overview) each supplier processes a certain good (measured in units of parts) and passes it on to the next supplier in the chain. Labeling the parts by the index n, we denote by $\tau(m, n)$ the time at which part number n passes from supplier number $m-1$ to supplier number m. The goal of supply chain modeling and control is to derive rules governing the evolution of the times $\tau(m, n)$ and, in further consequence, to design such rules to, in some predefined sense, optimally manage the supply chain. There is a hierarchy of models available for this purpose. If the times $\tau(m, n)$ are used as primary variables, and therefore each part is considered individually, this leads to so-called discrete event simulation models (see [7] for an overview), which represent the most exact, and computationally most expensive, simulation tool. On the other end of the spectrum lie so-called fluid models, which replace the individual parts by a continuum and use rate equations for the flow of product through a supplier (see [1], [8] for an overview). For a large number of parts, fluid models are much less expensive but necessarily represent an approximation to the actual situation. As a compromise between the two extremes, so-called traffic flow models have received a lot of attention recently. The name derives from the analogy of the parts moving like cars on a highway and the use of a large already developed body of theory for modeling traffic flows. This theory employs the methodology of an even older and better developed theory, namely that of gas dynamics. So, discrete event simulation takes the place of particle based (i.e., Monte Carlo-type) models for gases, which can be approximated by the equations of gas dynamics (see [10] for an overview) and so on. The analogy is of course not one to one since the basic rules governing the parts

*Received by the editors March 1, 2004; accepted for publication (in revised form) September 19, 2005; published electronically February 21, 2006. This work was supported by NSF grant DMS-0204543.

http://www.siam.org/journals/siap/66-3/60462.html

†Department of Mathematics, Arizona State University, Tempe, AZ 85287-1804 (dieter@math.la.asu.edu, ringhofer@asu.edu).

‡MIP, Laboratoire CNRS (UMR 5640), Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse Cedex 04, France (degond@mip.ups-tlse.fr).

in a supply chain, the cars on a highway and the molecules in a gas, will be different [2], [3], [4], [6], [11], [16].

This paper is concerned with the derivation of a type of traffic flow model, namely a conservation law for a partial differential equation, out of very simple principles governing the evolution of the times $\tau(m,n)$. Given the times $\tau(m,n)$ conservation of the number of parts is expressed via the introduction of so-called N-curves (originally defined by Newell [17]). The N-curve $U(t)$ at supplier S_m is given by the number of parts which have passed from processor S_{m-1} to processor S_m at time t, i.e., by

$$U(m,t) = \sum_{n=0}^{\infty} H(t - \tau(m,n)), \tag{1}$$

where H denotes the usual Heaviside function. The flux from processor S_{m-1} into processor S_m is given by the derivative of $U(m,t)$, i.e.,

$$F(m,t) = \frac{d}{dt}U(m,t) = \sum_{n=0}^{\infty} \delta(t - \tau(m,n)), \quad m = 0, \ldots, M, \tag{2}$$

which holds with $F(0,t)$ and $F(M,t)$ the total influx and outflux of the supply chain. So N-curves are just the antiderivatives of fluxes. The work in progress (WIP) $W(m,t)$ of processor S_m, the total number of parts currently at the supplier S_m at time t, is now given by the difference of two consecutive N-curves, i.e.,

$$W(m,t) = U(m,t) - U(m+1,t) + K(m), \quad m = 0, \ldots, M-1, \tag{3}$$

where the time independent constants $K(m)$ are determined from the initial situation. Combining (2) and (3) yields the conservation law

$$\frac{d}{dt}W(m,t) = F(m,t) - F(m+1,t) \tag{4}$$

for the WIP $W(m,t)$ and the flux $F(m,t)$, both given in terms of the transition times $\tau(m,n)$. Note that in (4) $W(m,t)$ is a step function in time while $F(m,t)$ is a superposition of δ-functions. Furthermore, if each of the suppliers S_m has a given minimal processing time $T(m)$, $\tau(m+1,n) \geq \tau(m,n)+T(m)$ will hold, which implies, amongst other things, that the WIP $W_m(t)$ can never become negative. Fluid and traffic models replace the WIP W and the flux F by continuous functions and eliminate the dependence on individual parts, either by ad hoc assumptions, constitutive relations derived from stochastic queuing theory in a quasi-steady state (see cf. [9]), or via asymptotic methodology borrowed from the theory of gas dynamics [4], [5], [14], [15], [18]. In the simplest fluid models, the fluxes $F(m,t)$ in (4) are prescribed and the WIP's $W(m,t)$ are computed from F. Constraints have to be placed on the fluxes in order to guarantee nonnegative WIPs. This is usually done in a linear programming framework [19].

The basic concept of the approach presented in this paper is somewhat different. Rather than artificially constraining the fluxes, we will derive a continuum model which contains as an input parameter a service rate μ, and in which the WIP's $W(m,t)$ will always be nonnegative. The model is based on very simple assumptions, namely that each supplier functions as a single processor with a processing time T and a buffer queue in front of it. Based on this assumption, we derive, in a continuum limit,

a conservation law of the form

$$\partial_t \rho + \partial_x \min\left\{\mu, \frac{W}{T}\right\} = 0, \tag{5}$$

where the artificial continuous variable x indexes the suppliers and the $\rho(x,t)$ denotes the product density over x, i.e., $W = \int \rho\, dx$ holds. If the number of parts considered is very large, then solving the conservation law (5) is obviously much more effective than to directly compute the $\tau(m,n)$.

In contrast to previously presented approaches [2], [3], [4], [5], the approach in this paper is based on first principles. While purely fluid dynamic approaches rely on constitutive laws (usually for the equivalent of the pressure tensor [4]), we derive the conservation law (5) rigorously from a simple recursion of the arrival times $\tau(m,n)$.

This paper is organized as follows. In section 2 we define the basic rule governing the transition times $\tau(m,n)$, modeling one supplier in the chain as a processor with a given throughput time and a linear buffer queue in front of it. We heuristically derive simplified formulas to compute the WIP density ρ and the flux from the transition times. These formulas are simpler than (2) and (3), in the sense that they depend only locally on the $\tau(m,n)$. This will allow us to derive simple constitutive relations for the flux and WIP density leading to the conservation law (5). However, with this simple constitutive relation, the conservation law (5) will be satisfied only approximately. In section 3 we show that (5) is satisfied asymptotically in the limit for a large number of suppliers. The main difficulty here is that, as it turns out, the conservation law (5) will in general have only distributional solutions. $\rho(x,t)$ will develop δ-function concentrations, corresponding to bottlenecks in the supply chain. We will resolve this problem by instead deriving the corresponding hyperbolic equation for the N-function U in (1). This will also allow us to numerically compute the distributional solutions of (5) in a reasonable way. The assumption of a large number of nodes in the supply chain is actually unreasonable for many applications. In section 4 we remove this assumption by replacing one individual supplier with an arbitrary number of virtual suppliers, allowing us to pass to a continuum limit in almost every situation. Section 5 is devoted to numerical experiments. We demonstrate the asymptotic validity of the continuum model on two examples, one with only a few nodes in the supply chain where we utilize the concept of virtual suppliers, and one example of a long supply chain with randomly generated processing times and capacities.

2. The basic model. In this section we first define the basic rules governing the supply chain. We then give a more or less heuristic reasoning for a formula which expresses the flux and the density of parts locally in time, i.e., it is dependent only on differences of neighboring transition times τ, in what is essentially a large time regime. With these local formulas we derive in Theorem 1 a constitutive relation which expresses the flux in terms of the density. We first present the basic model for a single node in the supply chain. We assume that the node consists of a processor which processes parts at a rate μ. In front of this "machine" we assume a buffer queue, i.e., parts arrive at the end of the queue, wait until they reach the front, and then are fed into the processor. We denote by a_n, $n = 0, 1, \ldots$, the time part number n arrives at the end of the queue and by b_n the "release time," i.e., the time part number n, reaches the front of the queue and is fed into the processor. If the queue is full, the interval between two consecutive release times b_n will be given by the processing rate μ, i.e.,

$$b_n = b_{n-1} + \frac{1}{\mu}$$

will hold as long as $a_n \leq b_{n-1} + \frac{1}{\mu}$ holds, meaning that part number n has already arrived when we want to feed it into the processor. If, on the other hand, the queue is empty, i.e., if at the desired release time $b_{n-1}+\frac{1}{\mu}$ part number n has not arrived yet at the end of the queue, then we wait for its arrival and then immediately feed it into the processor. So, $a_n > b_{n-1} + \frac{1}{\mu}$ will imply $b_n = a_n$. This gives altogether the relation

$$b_n = \max\left\{a_n, b_{n-1} + \frac{1}{\mu}\right\}. \tag{6}$$

We assume that the processor takes a time T to finish the part and denote by $e_n = b_n + T$ the time the part leaves the processor (and enters the next queue). Inserting this relation into (6) gives

$$e_n = \max\left\{a_n + T, e_{n-1} + \frac{1}{\mu}\right\} \tag{7}$$

as the basic law relating the arrival times a_n to the exit times e_n. We now consider a chain of M suppliers $S_0, \ldots, S_{M-1}$ and denote with $\tau(m, n)$ the time part number n arrives at supplier S_m. Using the obvious change of notation $a_n \to \tau(m, n)$ and $e_n \to \tau(m + 1, n)$, we obtain from (7)

$$\begin{aligned}\tau(m+1,n) =& \max\left\{\tau(m,n) + T(m), \tau(m+1,n-1)\right.\\ &\left. + \frac{1}{\mu(m,n-1)}\right\}, \quad m = 0, \ldots, M_1,\ n \geq 1.\end{aligned} \tag{8}$$

Here $T(m)$ denotes the processing time of processor S_m, and we have made the processing rates μ time dependent, i.e., dependent on the part index n as well. The reason for the latter is that the service rates μ are used to control the supply chain. So, after part number $n - 1$ has been fed into processor S_m, we wait a time interval $\frac{1}{\mu(m,n-1)}$ before feeding in the next part. We assume that the processor belonging to the node S_m has a finite capacity $C(m)$, so

$$\mu(m, n) \leq C(m), \quad m = 0, \ldots, M - 1,\ n \geq 0 \tag{9}$$

has to hold, but otherwise the μ's can be chosen arbitrarily. The recursion (8) still needs initial and boundary conditions. They are of the form

$$\tau(0, n) = \tau^A(n), \quad n \geq 0, \quad \tau(m, 0) = \tau^I(m), \quad m = 0, \ldots, M.$$

$\tau^A(n)$ simply denotes the arrival time of part n in the first processor of the chain. The interpretation of $\tau^I(m)$ is somewhat more subtle. Obviously, $\tau^I(m)$ denotes the time the first part has arrived at supplier S_m. So, $\tau^I(m + 1) - \tau^I(m) - T(m)$ denotes the time the first part has waited in the buffer in front of the processor at S_m. Assuming a constant service rate μ in the past, $\mu(m, 0)[\tau^I(m + 1) - \tau^I(m) - T(m)]$ would be the number of parts in the queue at the time part number 0 arrives. This, in a sense, records the history of what has happened in the system before the first part went through and determines the queue length at the initial time. This somewhat awkward definition is necessitated by the fact that, for an actual simulation, we have to start somewhere. This issue will be resolved once the problem is formulated in terms of an approximate conservation law.

The goal of this paper is to asymptotically replace (8) by a conservation law with a simple constitutive relation. The rest of this section is devoted to considerations of what the appropriate form of the constitutive relation $F = F(W)$ in (4) should be. In the next section we will then show that with this relation, an equivalent of (5), holds in a weak sense. We start by redefining the flux. First, we map (4) onto a grid in an artificial spatial variable x, called the "degree of completion" (DOC). We define a mesh $0 = x_0 < \cdots < x_M = X$ and replace $F(m,t)$ by $F(x_m,t)$. So, parts enter the supply chain at the DOC $x = 0$ and leave at the DOC $x = X$. Next, we observe that, for an arbitrary test function $\psi(t)$,

$$\int_{\tau^I(m)}^{\infty} \psi(t)F(x_m,t)\ dt = \sum_{n=0}^{\infty} \psi(\tau(m,n))$$

holds. We rewrite this into a Riemann sum for an integral as

$$\int_{\tau^I(m)}^{\infty} \psi(t)F(x_m,t)\ dt = \sum_{n=0}^{\infty} \psi(\tau(m,n))\Delta_n\tau(m,n)f(x_m,\tau(m,n)), \tag{10}$$

where $\Delta_n\tau$ denotes the difference of $\tau(m,n)$ in the index n and the function $f(x,t)$ is given at $x = x_m$ and $t = \tau(m,n)$ as the reciprocal difference, i.e.,

$$\text{(a) } \Delta_n\tau(m,n) := \tau(m,n+1) - \tau(m,n), \quad \text{(b) } f(x_m,\tau(m,n)) := \frac{1}{\Delta_n\tau(m,n)} \tag{11}$$

holds. On a time scale, where $\Delta_n\tau$ is small, (10) becomes

$$\int_{\tau^I(m)}^{\infty} \psi(t)F(x_m,t)\ dt \approx \int_{\tau^I(m)}^{\infty} \psi(t)f(x_m,t)\ dt.$$

So (11)(b) will be the definition of our approximate flux f, which is given on the grid $\tau(m,n)$ for $x = x_m$. To find an approximate expression for the density ρ of parts per unit DOC, we consider the case when the arrival times τ would be distributed continuously, i.e., if they were given as a function $\tau(x,y)$. In this case (11)(b) would become $f(x,\tau(x,y)) = \frac{1}{\partial_y\tau(x,y)}$. The N-function $U(x,t)$, the antiderivative of the flux, would then satisfy the relations

$$\begin{aligned} &\text{(a) } \frac{d}{dy}U(x,\tau(x,y)) = \partial_t U(x,\tau)\partial_y\tau = 1, \\ &\text{(b) } \frac{d}{dx}U(x,\tau(x,y)) = \partial_x U(x,\tau) + \partial_t U(x,\tau)\partial_x\tau. \end{aligned} \tag{12}$$

Setting $\rho(x,t) = K(x) - \partial_x U(x,t)$ for an arbitrary function K, in analogy to (3), (12) becomes

$$\frac{d}{dx}U(x,\tau(x,y)) = K(x) - \rho(x,\tau) + f(x,\tau)\partial_x\tau.$$

Now (12)(a) implies that $\frac{d}{dx}U(x,\tau(x,y))$ is a function of the DOC variable x only, which we set equal to the arbitrarily chosen function $K(x)$. So, for a continuum $\tau(x,y)$ of arrival times, we set $f(x,\tau) = \frac{1}{\partial_y\tau}$ and $\rho(x,\tau) = \frac{\partial_x\tau}{\partial_y\tau}$. Direct calculus yields that, if so defined, ρ and f satisfy a conservation law of the form $\partial_t\rho + \partial_x f = 0$.

Motivated by this, we define the approximate density ρ and the approximate flux f from the arrival times τ by

$$\begin{aligned} &\text{(a) } \mathrm{f}(\mathrm{x_m}, \tau(\mathrm{m}, \mathrm{n})) = \frac{1}{\Delta_n \tau(m,n)}, \quad m = 0, \dots, M,\ n = 0, 1, \dots, \\ &\text{(13) (b) } \rho(x_m, \tau(m+1,n)) = \frac{\Delta_m \tau(m, n+1)}{h_m \Delta_n \tau(m+1, n)}, \quad m = 0, \dots, M-1,\ n = 0, 1, \dots, \\ &\text{(c) } \Delta_m \tau(m,n) := \tau(m+1,n) - \tau(m,n), \quad h_m := x_{m+1} - x_m. \end{aligned}$$

The density and flux defined by (13) are approximate in the sense that they will, as will be seen in section 3, satisfy an approximate or discretized version of the conservation law. However, the definition (13) allows us to derive a simple constitutive relation of the form $f = f(\rho)$. Under what circumstances ρ and f satisfy an approximate conservation law will be the subject of the next section. We have the following.

THEOREM 1. *Let the arrival times $\tau(m,n)$ satisfy the recursion* (8). *Let the approximate density ρ and flux f be defined by* (13). *Then the approximate flux can be written in terms of the approximate density via a constitutive relation of the form*

$$f(x_m, \tau(m,n)) = \phi_{mn}(\rho(x_{m-1}, \tau(m,n))), \quad m = 1, \dots, M,\ n \geq 0,$$

with the flux function ϕ_m given by

$$\phi_{mn}(\rho) = \min\left\{\mu(m-1,n), \frac{h_{m-1}\rho}{T(m-1)}\right\}. \tag{14}$$

The proof of Theorem 1 is rather lengthy and therefore deferred to the appendix.

The advantage of the approximative constitutive law (14) over the exact law given by (2) and (3) lies in the fact that it does not involve the transition times $\tau(m,n)$ anymore. The subject of the next two sections will be if, and in what sense, ρ and f will still satisfy a conservation law of the form $\partial_t \rho + \partial_x f = 0$.

3. Asymptotic validity of the conservation law. In this section we show that the approximate density ρ and flux f defined by (13) satisfy, in a certain sense, a conservation law of the form $\partial_t \rho + \partial_x f = 0$ asymptotically. The asymptotic regime we consider is one for a large number of nodes in the supply chain and for large time scales. The assumption of a large number of nodes is to some extent artificial and will be removed in section 4. As it turns out the limiting density ρ will in general not be a classical function but a distribution. We therefore show the asymptotic validity for the corresponding hyperbolic differential equation for the limiting N-curve U in (1).

3.1. Scaling and dimensionless formulation. We define by T_0 the average processing time, i.e.,

$$T_0 = \frac{1}{M} \sum_{m=0}^{M-1} T(m)$$

holds. So MT_0 would be the time for a part to be processed in the empty system, without waiting in any queue. This is chosen as the overall time scale, whereas we scale the individual processing times $T(m)$ and service rates $\mu(m,n)$ by T_0. Denoting scaled variables with the subindex s, this gives

$$\tau(m,n) = MT_0 \tau_s(m,,n), \quad T(m) = T_0 T_s(x_m), \quad \mu(m,n) = \frac{\mu_s(x_m, \tau_s(m+1,n))}{T_0}. \tag{15}$$

We will consider a regime where $M >> 1$ holds and will set $\varepsilon = \frac{1}{M} << 1$ from here on. With this scaling the recursion (8) becomes

(16)

$$\text{(a) } \tau_s(m+1,n+1) = \max\left\{\tau_s(m,n+1) + \varepsilon T_s(x_m), \tau_s(m+1,n) + \frac{\varepsilon}{\mu_s(x_m, \tau_s(m+1,n))}\right\}, m = 0,\dots,M-1,\ n = 0,1,\dots,$$

$$\text{(b) } \tau_s(0,n) = \tau_s^A(n), \quad n \geq 0, \quad \tau_s(m,0) = \tau_s^I(m), \quad m = 0,\dots,M,$$

where we have scaled τ^A and τ^I in the same way as $\tau(m,n)$. Also we have made grid functions out of the throughput times and processing times. We assume that the differences between two consecutive arrival times τ are of the same order as the average processing time T_0. This is reasonable since otherwise the total WIP would either go to zero or infinity. So we set

$$\Delta_n \tau(m,n) = \tau(m,n+1) - \tau(m,n) = T_0 \Delta_{ns} \tau_s(m,n),$$
$$\Delta_m \tau(m,n) = \tau(m+1,n) - \tau(m,n) = T_0 \Delta_{ms} \tau_s(m,n),$$

giving

$$\tau_s(m+1,n) = \tau_s(m,n) + \varepsilon \Delta_{ms} \tau_s(m,n), \quad \tau_s(m,n+1) = \tau_s(m,n) + \varepsilon \Delta_{ns} \tau_s(m,n).$$

In accordance with (13), we scale the density ρ and the flux f by

$$f(x,t) = \frac{1}{T_0} f_s\left(x, \frac{t}{MT_0}\right), \quad \rho(x,t) = \frac{M}{X} \rho_s\left(x, \frac{t}{MT_0}\right),$$

where X is the length of the DOC interval. This gives

(17)

$$\text{(a) } f_s(x_m, \tau_s(m,n)) = \frac{1}{\Delta_{ns}\tau_s(m,n)}, \quad m = 0,\dots,M,\ n = 0,1,\dots,$$

$$\text{(b) } \rho_s(x_m, \tau_s(m+1,n)) = \frac{\varepsilon X \Delta_{ms}\tau_s(m,n+1)}{h_m \Delta_{ns}\tau_s(m+1,n)}, \quad m = 0,\dots,M-1,\ n = 0,1,\dots,$$

as a definition for the scaled flux and density with $f_s, \rho_s = O(1)$. The scaled version of the constitutive relation (14) then reads

(18)

$$\text{(a) } f_s(x_m, \tau_s(m,n)) = \phi_s(x_{m-1}, \tau_s(m,n), \rho_s(x_{m-1}, \tau_s(m,n))),$$

$$\text{(b) } \phi_s(x_{m-1}, t, \rho_s) = \min\left\{\mu_s(x_{m-1}, t), \frac{h_{m-1}\rho_s}{\varepsilon X T_s(x_{m-1})}\right\}.$$

(18)(b) suggests a natural choice for the grid in x-direction, namely

(19)

$$h_m = \varepsilon X T_s(x_m) = \frac{XT(m)}{\sum_{m'=0}^{M-1} T(m')}, \quad m = 0,\dots,M-1,$$

which makes the propagation velocity in (18) equal to unity, i.e., we assign an interval in the DOC variable x to processor S_m which is proportional to its processing time. We will use this choice of the mesh from here on. Also, from now on we will drop the subscript s for simplicity.

3.2. Interpolation and weak formulation. We now proceed to show the asymptotic validity of a conservation law in the limit $\varepsilon \to 0$. The goal is an initial boundary value problem for a conservation law of the form

$$\partial_t \rho + \partial_x f = 0, \quad f = \min\{\mu(x,t), \rho\}, \quad f(0,t) = f^A(t), \tag{20}$$

together with some initial condition. There are several complications in this approach.

- First, the resulting initial boundary value problem cannot be defined on a strip in (x,t) plane but on a domain bounded by $t > \tau^I$. This is more of a notational inconvenience, but impacts the definition of initial conditions.
- From the original definition of the problem we cannot assume any kind of smooth relation between two consecutive processors, i.e., we cannot assume that the throughput times $T(x_m)$ and service rates $\mu(x_m,t)$, defined by (15), will converge to a smooth function in the limit $\varepsilon \to 0$. The limiting problem therefore has to be defined weakly.
- The most severe problem is that the flux function f can become discontinuous. This can be seen from the following consideration. Since we cannot assume any smooth relation between consecutive processors, we have to allow for the possibility of a sharp drop in the service rate μ, i.e., $\mu(x_m,t) > \mu(x_{m+1},t)$, which does not vanish in the limit $M \to \infty$. At this point we can easily construct a situation where $f(x_m,t) > \mu(x_{m+1},t)$ holds. Since $f(x_{m+1},t)$ is cut off by the min-function, the limiting flux f will have to be discontinuous. Because mass still has to be conserved, this discontinuity has to be compensated by a δ-function concentration in the density ρ at this point. This corresponds to a bottleneck situation, where we feed into a processor at a rate higher than its capacity over a significant period of time. Consequently, the queues will grow, which is expressed as a δ-function in the limit. This situation will actually occur right in the beginning of the supply chain if the boundary flux $f^B(t)$ is chosen larger than the capacity of the first processor.

We deal with the above problem by redefining our concept of a solution. Instead of deriving a conservation law for the density ρ we derive a hyperbolic equation for the limiting N-function U in (2). We denote its approximation by u, set $\rho(x,t) = -\partial_x u(x,t)$, and integrate (20) once with respect to x. This gives

$$\partial_t u = \min\{\mu(x,t), -\partial_x u\}, \quad \lim_{x \to 0-} u(x,t) = g^A(t), \quad \frac{d}{dt} g^A(t) = f^A(t). \tag{21}$$

Clearly, if the solution $u(x,t)$ is continuous and has a bounded x-derivative, we obtain a solution $\rho(x,t)$ of (20) by differentiating u with respect to x. However, (21) allows for shock solutions which result in δ-functions in the variable ρ. Although the x-derivative of u in this case becomes unbounded, the flux will remain bounded because of the min-function. ($-\partial_x u = \rho$ will always be bounded from below by zero.) We will therefore show that, in the limit $\varepsilon \to 0$, the N-function u satisfies a hyperbolic problem of the form (21) weakly in x and t. To do so, we first have to define the variables given on the nonuniform and nonrectangular mesh in (x,t) for continuous arguments by piecewise constant interpolation. For a given gridpoint x_m we first interpolate the

grid functions ρ and f defined by (17) in time direction by

$$\begin{aligned}
&\text{(a) } f_1(x_m,t) = f(x_m,\tau(m,n)), \quad \tau(m,n) \le t < \tau(m,n+1),\\
&\qquad\qquad m = 0,\dots,M-1,\ n \ge 0\\
&\text{(b) } \rho_1(x_m,t) = \rho(x_m,\tau(m+1,n-1)), \quad \tau(m+1,n-1) \le t < \tau(m+1,n),\\
&\qquad\qquad m = 0,\dots,M-1,\ n \ge 1,\\
&\text{(c) } f^A(t) = \frac{1}{\Delta_n \tau^A(n)}, \quad \tau^A(n) \le t < \tau^A(n+1).
\end{aligned} \tag{22}$$

Next we define the N-function $u(x_m,t)$ by

$$u_1(x_{m+1},t) = u_1(x_m,t) - \frac{h_m}{X}\rho_1(x_m,t), \quad m = 0,\dots,M-1, \quad u_1(x_0,t) = \int_{\tau(0,0)}^{t} f^A(s)\ ds. \tag{23}$$

Given the functions ϕ_1, u_1 which are now defined for continuous time and discrete space, we define the functions

$$\begin{aligned}
&\text{(a) } f_2(x,t) = f_1(x_{m+1},t), \quad x_m \le x < x_{m+1},\ m = 0,\dots,M-1,\\
&\text{(b) } \tau_2^I(x) = \tau^I(m+1), \quad x_m \le x < x_{m+1},\ m = 0,\dots,M-1,\\
&\text{(c) } u_2(x,t) = u_1(x_{m+1},t), \quad x_m \le x < x_{m+1},\ m = 0,\dots,M-1,
\end{aligned} \tag{24}$$

as functions of continuous space and time.

3.3. The limit $\varepsilon \to 0$. We can now show that the so defined interpolant u_2, f_2 satisfies a weak version of (21). We have the following theorem.

THEOREM 2. *Given the scaled density and flux at the discrete points $x_m, \tau(m,n)$, as defined in* (16)*, let the piecewise constant interpolant u_2 and f_2 be defined as in* (22)*. Let the scaled throughput times $T(x_m)$ stay uniformly bounded, i.e., $h_m = O(\varepsilon)$ holds uniformly in m. Assume finitely many bottlenecks for a finite amount of time, i.e., let $\Delta_m\tau(m,n)$ be bounded for $\varepsilon \to 0$ except for a certain number of nodes m and a finite number of parts n, which stays bounded as $\varepsilon \to 0$. Then, for $\varepsilon \to 0$ and $\max h_m \to 0$ the interpolated N-function and flux u_2, f_2 satisfy the initial boundary value problem*

$$\begin{aligned}
&\text{(a) } \partial_t u_2 = f_2, \quad t > \tau_2^I(x),\ 0 < x < X,\\
&\text{(b) } u_2(x,\tau^I(x)) = 0, \quad \lim_{x\to 0-} u_2(x,t) = \int_{\tau_2(0,0)}^{t} f^A(s)\ ds,
\end{aligned} \tag{25}$$

in the limit $\varepsilon \to 0$, weakly in x and t.

The proof of Theorem 2 is deferred to the appendix.

Remark. Theorem 2 establishes the asymptotic validity of the integrated conservation law (25)(a) for any N-curve u and any flux function f, derived from an arbitrary sequence τ via the definition (13) and the interpolation formulas (22) and (24). The constitutive relation $f_2 = \min\{\mu, -\partial_x u_2\}$ is a consequence of the recursion relation satisfied by the sequence $\{\tau(m,n)\}$ and, consequently, of Theorem 1.

Remark. In unscaled variables Theorem 2 implies that the density $\rho(x,t)$ can be approximately computed as $\rho = -\partial_x u$ where the unscaled N-function $u(x,t)$ is the solution of

$$\partial_t u = \min\left\{\mu_-, -\frac{X}{MT_0}\partial_x u\right\}, \quad 0 < x < X, \quad \lim_{x\to 0-} u(x,t) = \int_{\tau(0,0)}^{t} f^A(s)\ ds, \tag{26}$$

$$\mu_-(x,t) := \lim_{y \to x-0} \mu(y,t).$$

Remark. The assumptions of Theorem 2 state that the number of nodes in the supply chain is large, that the number of bottlenecks is small compared to the number of processors, and that each of the processing times is small compared to the overall throughput time, i.e., $T(m) << \sum_{m'=0}^{M-1} T(m')$ holds. At first glance, these assumptions might seem rather restrictive. We will remove these restrictions in the next section by introducing the concept of virtual processors, which will allow us to arbitrarily increase M.

3.4. An exact solution for a single bottleneck. To illustrate the dynamics induced by the conservation law (20), we compute an exact solution for the special case of a single bottleneck. Suppose (20) is posed on the interval $x \in [0,1]$, with a bottleneck at $x = \frac{1}{2}$, and we prescribe an arrival rate f^A which can be processed by the processors in front of the bottleneck but is larger than the capacity of the processors behind the bottleneck. So we have

$$\mu(x) = \begin{cases} \mu_1 & \text{for} \quad 0 < x < \frac{1}{2}, \\ \mu_2 & \text{for} \quad \frac{1}{2} < x < 1, \end{cases} \qquad \mu_2 < f^A(t) < \mu_1.$$

The solution $\rho(x,t)$ will then be given by a classical part $\rho_c(x,t)$, with a jump discontinuity at $x = \frac{1}{2}$, and a δ-function of the form $q(t)\delta(x - \frac{1}{2})$, compensating the jump in the fluxes. The classical part ρ_c will just satisfy a one way wave equation with constant velocity. So, we have

$$\partial_t \rho_c + \partial_x \rho_c = 0, \ x \in \left(0, \frac{1}{2}\right) \cup \left(\frac{1}{2}, 1\right), \quad \rho_c(0,t) = f^A(t), \quad \rho_c\left(\frac{1}{2}+, t\right) = \mu_2,$$

whose solution is given via characteristics by

$$\rho_c(x,t) = \begin{cases} f^A(t-x) & \text{for} \quad 0 < x < \frac{1}{2}, \\ \mu_2 & \text{for} \quad \frac{1}{2} < x < 1. \end{cases} \tag{27}$$

In order for the whole solution $\rho(x,t) = \rho_c(x,t) + q(t)\delta(x - \frac{1}{2})$ to be a spatially weak solution of the conservation law (20), we have to satisfy

$$\int_0^1 \phi(x)\partial_t \rho(x,t) - \min\{\mu(x), \rho(x,t)\}\partial_x \phi(x)\ dx = \phi(0) f^A(t) - \phi(1)\min\{\mu_2, \rho(1,t)\}$$

for any arbitrarily smooth test function $\phi(x)$. Integrating by parts separately on the intervals $(0, \frac{1}{2})$ and $(\frac{1}{2}, 1)$ gives

$$\begin{aligned} &\int_0^1 \phi(x)\partial_t \rho_c(x,t)\ dx + q'(t)\phi\left(\frac{1}{2}\right) + \int \phi(x)\partial_x \min\{\mu(x), \rho_c(x,t)\}\ dx \\ &- \min\left\{\mu_1, \rho_c\left(\frac{1}{2}-, t\right)\right\}\phi\left(\frac{1}{2}\right) + \min\{\mu_1, \rho_c(0,t)\}\phi(0) + \min\left\{\mu_2, \rho_c\left(\frac{1}{2}+, t\right)\right\}\phi\left(\frac{1}{2}\right) \\ &= \phi(0) f^A(t). \end{aligned}$$

Since $\rho_c(x,t) < \mu(x)$ will hold everywhere and $\rho_c(0,t) = f^A(t)$ holds, this reduces to

$$q'(t) = f^A\left(t - \frac{1}{2}\right) - \mu_2 = 0. \tag{28}$$

Thus, away from the bottleneck at $x = \frac{1}{2}$ the solution is given by (27), and the bottleneck produces a buildup of the queue (a δ-function in this framework) with strength (or queue length) q, which is governed by (28).

4. Virtual processors. As pointed out in section 3, the asymptotic validity of the differential equation (26) is given only for the case when the number M of processors is large and each of the individual processing times $T(m)$ is small compared to the total processing time $MT_0 = \sum_{m=0}^{M-1} T(m)$. So, it excludes cf. the situation where one processor takes up half of the overall processing time. In this section we will relax this restriction by introducing the concept of virtual processors. The basic idea is that one processor with a processing time T and a service rate μ can be replaced by K virtual processors with the same service rate μ and processing times $\frac{T}{K}$. Thus, we can make the total number of processors as large as we like, and the relative processing times as small as we like, by introducing enough virtual processors. The purpose of this section is to make this statement precise. Since, in doing so, we will keep the service rates μ constant but decrease the processing times T, eventual bottlenecks will occur only in the first virtual processor, and the queues of the additional virtual processors will always remain empty. Given the recursion formula (8), we therefore derive a condition for queues being always empty.

LEMMA 1. *Given the recursion* (8) *for the arrival times* $\tau(m,n)$*, let the arrival rate in node* S_m *be below the service rate* μ*, i.e., let*

$$\tau(m,n+1) - \tau(m,n) \geq \frac{1}{\mu(m,n)}, \quad n = 0, \ldots, \tag{29}$$

hold. Furthermore let the queue be empty at the arrival of the first part, i.e., let

$$\tau(m,1) + T(m) \geq \tau(m+1,0) + \frac{1}{\mu(m,0)} \tag{30}$$

hold. Then

$$\tau(m+1,n) = \tau(m,n) + T(m), \quad n = 1, \ldots,$$

holds.

Proof of Lemma 1. Define waiting time in the queue number m as $Q(m,n) = \tau(m+1,n) - \tau(m,n) - T(m)$. Inserting this into (8) gives

$$Q(m,n+1) = \max\left\{0, Q(m,n) + \tau(m,n) - \tau(m,n+1) + \frac{1}{\mu(m,n)}\right\} \tag{31}$$

as a recursion for the waiting times $Q(m,n)$. In particular,

$$Q(m,1) = \max\left\{0, \tau(m+1,0) - T(m) - \tau(m,1) + \frac{1}{\mu(m,0)}\right\} = 0$$

holds because of (30). Because of (29), the term $\tau(m,n) - \tau(m,n+1) + \frac{1}{\mu(m,n)}$ is always nonpositive and therefore the recursion (31) has the trivial solution $Q(m,n) = 0$ for $n \geq 1$. ∎

Lemma 1 will provide the basic tool to split a processor into K virtual processors. The basic building block of the underlying idea is to split one processor into two. Without loss of generality we perform this split on the first node in the supply chain. We have the following lemma.

LEMMA 2. *Let the flow of parts in processor* S_0 *be governed by*

$$\tau(1,n+1) = \max\left\{\tau(0,n+1) + T(0), \tau(1,n) + \frac{1}{\mu(0,n)}\right\}, \tag{32}$$

with $\tau(0,n),\ n \geq 0$ *and* $\tau(1,0)$ *given and satisfying the compatibility condition* $\tau(1,0) \geq \tau(0,0) + T(0)$. *We replace* (32) *by two virtual nodes with the same processing rates and the same total throughput time, i.e.,*

(33)

$$\text{(a) } \hat{\tau}(1,n+1) = \max\left\{\hat{\tau}(0,n+1) + \hat{T}(0), \hat{\tau}(1,n) + \frac{1}{\mu(0,n)}\right\},$$

$$\text{(b) } \hat{\tau}(2,n+1) = \max\left\{\hat{\tau}(1,n+1) + \hat{T}(1), \hat{\tau}(2,n) + \frac{1}{\mu(0,n)}\right\},$$

$$\text{(c) } \hat{\tau}(0,n) = \tau(0,n), \quad n = 0,1\ldots, \qquad \hat{\tau}(1,0) = \tau(1,0) - \hat{T}(1), \quad \hat{\tau}(2,0) = \tau(1,0),$$

holds with $\hat{T}(0) + \hat{T}(1) = T(0)$. *Then the system* (33) *produces the same outflux as the system* (32), *i.e.,* $\hat{\tau}(2,n) = \tau(1,n)\ n \geq 0$ *holds.*

Proof of Lemma 2. We show that the second virtual processor, i.e., the times $\hat{\tau}(1,n)$ and $\hat{\tau}(2,n)$, satisfy the assumptions of Lemma 1. Because of (33)(a)

$$\hat{\tau}(1,n+1) \geq \hat{\tau}(1,n) + \frac{1}{\mu(0,n)}, \quad n \geq 0,$$

holds, giving (29). To show (30) we note that

$$\hat{\tau}(1,1) + \hat{T}(1) \geq \hat{\tau}(1,0) + \hat{T}(1) + \frac{1}{\mu(0,0)} = \tau(1,0) + \frac{1}{\mu(0,0)} = \hat{\tau}(2,0) + \frac{1}{\mu(0,0)}$$

holds. Because of Lemma 1

$$\hat{\tau}(2,n) = \hat{\tau}(1,n) + \hat{T}(1), \quad n = 1,\ldots, \tag{34}$$

holds, and (34) trivially holds for $n = 0$ as well because of the initial condition (33)(c). We now eliminate $\hat{\tau}(1,n)$ by inserting $\hat{\tau}(1,n) = \hat{\tau}(2,n) - \hat{T}(1)$ into (33)(a) and obtain

$$\hat{\tau}(2,n+1) = \max\left\{\hat{\tau}(0,n+1) + T, \hat{\tau}(2,n) + \frac{1}{\mu(0,n)}\right\},$$

i.e., $\hat{\tau}(0,n), \hat{\tau}(2,n)$ satisfy the same difference equation and initial and boundary conditions as $\tau(0,n), \tau(1,n)$. □

By repeatedly using Lemma 2, we immediately obtain the following theorem, as a corollary.

THEOREM 3. *Let the first processor* S_0 *in the chain be governed by* (8). *If we replace the single processor by* K *virtual processors with the same processing rates and the same total throughput time, i.e., by*

$$\hat{\tau}(m+1,n+1) = \max\left\{\hat{\tau}(m,n+1) + \frac{T(0)}{K}, \hat{\tau}(m+1,n) + \frac{1}{\mu(m,n)}\right\}, \quad m = 0,\ldots,K-1,$$

$$\hat{\tau}(0,n) = \tau(0,n), \quad \hat{\tau}(m,0) = \tau(1,0) - \left(1 - \frac{m}{K}\right)T(0), \quad m = 1,\ldots,K,$$

then we obtain the same outflux, i.e.,

$$\hat{\tau}(K,n) = \tau(1,n),\ n \geq 0,$$

holds.

So, in order to create the conditions appropriate for the application of Theorem 2, we would proceed as follows:

1. Given the processing times $T(m)$, $m = 0, \dots, M-1$, cut each of the processors S_m, $m = 0, \dots, M-1$, into $K(m)$ virtual processors, such that $T_1 = \frac{T}{K}$ is roughly equidistributed, giving $M_1 = \sum_{m=0}^{M-1} K(m)$ virtual processors.
2. If the number of virtual processors M_1 is still too small for the asymptotic regime in Theorem 2 to be valid, cut each of the virtual processors into additional L subprocessors to arrive at $M_2 = LM_1$ total processors.

Clearly, the number M_2 of virtual processors can be made as large as we like. No additional bottlenecks are created by this procedure since μ remains constant within each virtual processor belonging to one real processor, and a bottleneck can occur only if there is a drop in the processing rate μ. So, the number of bottlenecks remains finite as $M_2 \to \infty$. There is, however, a limit to this process since we have used the average processing time T_0 also to scale the service rates μ in (15). So, sending $T_0 \to 0$ would result in the scaled service rates μ, and therefore also the fluxes, going to zero. To obtain a reasonable limiting problem we should choose M_2 and T_0 in such a way that $T_0 C_0 = O(1)$ holds, where C_0 is some characteristic value for the capacities, the bounds on μ in (9). So, with the introduction of virtual nodes in the supply chain, the results of section 3 really apply to the case when $C_0 M_2 T_0 = C_0 \sum_{m=0}^{M-1} T(m) >> 1$ holds. For a stochastic queuing model in a steady state, this is, according to Little's law (see cf. [13]), a measure of the number of parts in the system. So the hyperbolic equation (25) in Theorem 2 is asymptotically valid for a large number of individual parts, i.e., precisely in situations where continuum models are computationally more efficient than discrete event simulators.

5. Numerical experiments. In this section we conduct two numerical experiments to verify Theorem 2 by comparing the solution of the hyperbolic problem (26) with the direct solution of the recursion (8) for the transition times τ. In both cases we solve (8); compute the WIP W, the N-curve U, and the flux F according to (1), (2), and (3); and compare it to ρ, u, and f computed from the solution of the hyperbolic equation (26). The hyperbolic problem for the approximate N-curve u is solved via a standard finite difference scheme of the form

(35)

(a) $$u(x_m, t_{n+1}) = u(x_m, t_n) + \Delta t f(x_m, t_n), \quad m = 0, \dots, M, \quad \Delta t = t_{n+1} - t_n,$$

(b) $$f(x_m, t) = \begin{pmatrix} \min\{\mu(x_{m-1}, t), -\frac{X}{MT_0 \Delta x_{m-1}}[u(x_m, t) - u(x_{m-1}, t)]\} & m = 1, \dots, M \\ f^A(t_n) & m = 0 \end{pmatrix}.$$

For simplicity, we use constant time steps satisfying a CFL condition of the form $\Delta t \le \frac{MT_0}{X} \min\{\Delta x_m\}$. If the spatial meshsizes Δx_m of the discretization of the conservation law are chosen equal to the h_m in (19), i.e., if we assign one gridpoint to one node in the supply chain, this would give $\Delta x_m = \frac{XT(m)}{MT_0}$, $m = 0, \dots, M-1$, and a CFL condition $\Delta t \le \min\{T(m)\}$. While this seems a natural choice it is not a necessary one. In particular, in regions where the service rates μ vary slowly, a larger spatial meshsize might be appropriate. Regardless of the choice of the spatial mesh the node S_m in the supply chain will always occupy an interval of length h_m. The influx f^A is computed according (11)(b) by

$$f^A(\tau^A(n)) = \frac{1}{\tau^A(n+1) - \tau^A(n)}$$

and piecewise linear interpolation. Note, that the discretization (35) is equivalent to discretizing the conservation law directly, i.e., if we define the discretized density ρ by

$$\rho(x_m, t_n) = -\frac{u(x_{m+1}, t_n) - u(x_m, t_n)}{\Delta x_m}, \quad m = 0, \ldots, M-1,$$

the discrete equation (35) becomes

(36)

(a) $$\rho(x_m, t_{n+1}) = \rho(x_m, t_n) - \frac{\Delta t}{\Delta x_m}[f(x_{m+1}, t_n) - f(x_m, t_n)], \quad m = 0, \ldots, M-1,$$

(b) $$f(x_m, t) = \begin{pmatrix} \min\{\mu(x_{m-1}, t), \frac{X\rho(x_{m-1}, t)}{MT_0}\} & m = 1, \ldots, M \\ f^A(t_n) & m = 0 \end{pmatrix}.$$

So, the discretization (35) is equivalent to directly discretizing the conservation law for the density ρ, ignoring the issue of distributional solutions. Of course u in (35) will still be discontinuous at bottlenecks, and ρ in (36) will grow like $\frac{1}{\Delta x}$ at these gridpoints. The discretization (36) represents only the simplest first order upwinding scheme for the hyperbolic conservation law. One could of course solve the hyperbolic problem (26) by more sophisticated high resolution methods on a correspondingly coarser mesh. Since this paper is concerned with the model per se, we felt that using a higher order method would somehow cloud the issue of model properties by introducing the artifacts of the numerical method.

In the first example we consider a supply chain of 3 suppliers with throughput times $T(0) = 1, T(1) = 3, T(2) = 1$ time units and capacities $C(0) = 15, C(1) = 10, C(2) = 15$ parts per time unit. Setting the characteristic value for the capacity $C_0 = 10$, this gives a value of $C_0 M T_0 = 50 >> 1$ for the average number of parts in a steady state. Thus, we can create the regime of Theorem 2 by introducing virtual processors according to section 4. We split the nodes S_0 and S_2 into 10 virtual nodes each with capacities of 15 parts per unit time and node S_1 into 30 virtual nodes with capacities of 10 parts per unit time. All of the 50 virtual nodes now have a throughput time of 0.1 time units and, setting the length X of the DOC interval equal to unity, the original suppliers will occupy the intervals $[0, 0.2], [0.2, 0.8], [0.8, 1]$. We simply set the service rates μ equal to the capacities, giving

$$\mu(x, t) = \begin{pmatrix} 15 & \text{for} & 0 < x < 0.2 \\ 10 & \text{for} & 0.2 < x < 0.8 \\ 15 & \text{for} & 0.8 < x < 1 \end{pmatrix}.$$

We expect 2 possible bottlenecks, namely at $x = 0.2$, where the capacity drops, and possibly at $x = 0$ if the influx exceeds 15 parts per unit time. We first solve the recursion (8) for the transition times τ, starting with all empty queues, i.e., $\tau^I(m+1) - \tau^I(m) = T(m) = 0.1$ holds, and set $\tau^I(M) = 0$. We compute the arrival times randomly according to $\tau^A(n+1) - \tau^A(n) = \frac{1}{f^A(\tau^A(n))}$, $\tau^A(0) = \tau^I(0)$. To study the development of bottlenecks, we choose a function $f^A(t)$ as the influx rate, which is first below the minimum capacity $C(1) = 10$, then between the minimum and the maximum capacity 10 and 15, then above the maximum capacity, and finally drops back to its original value. We add a random perturbation to a piecewise constant function. The influx rate f^A is shown in Figure 1. We compute fluxes and densities from the recursion (8) and the discretized conservation law (36). Figure 2 shows the

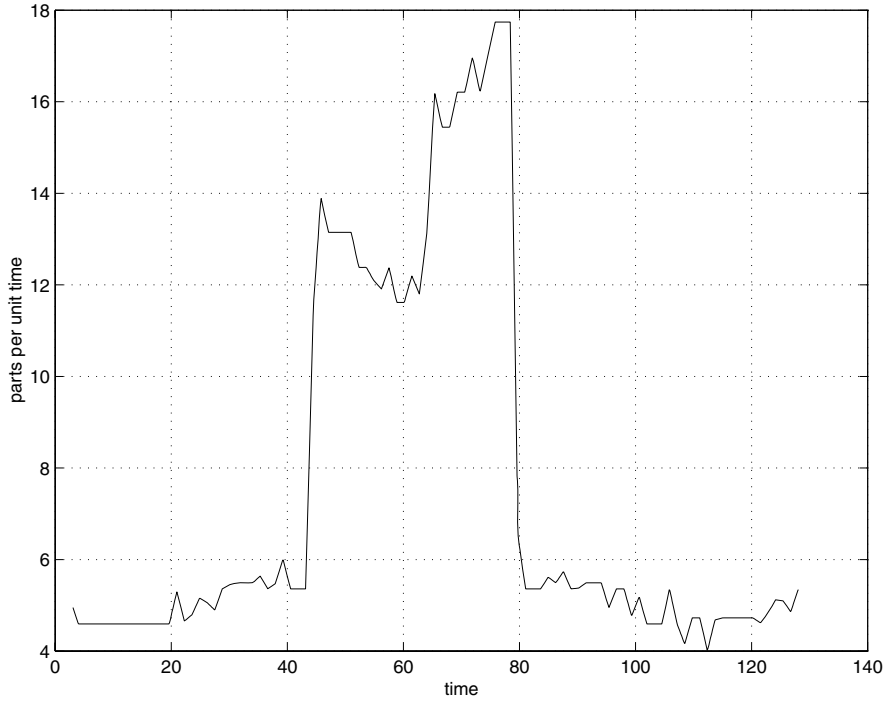


FIG. 1. *Influx.*

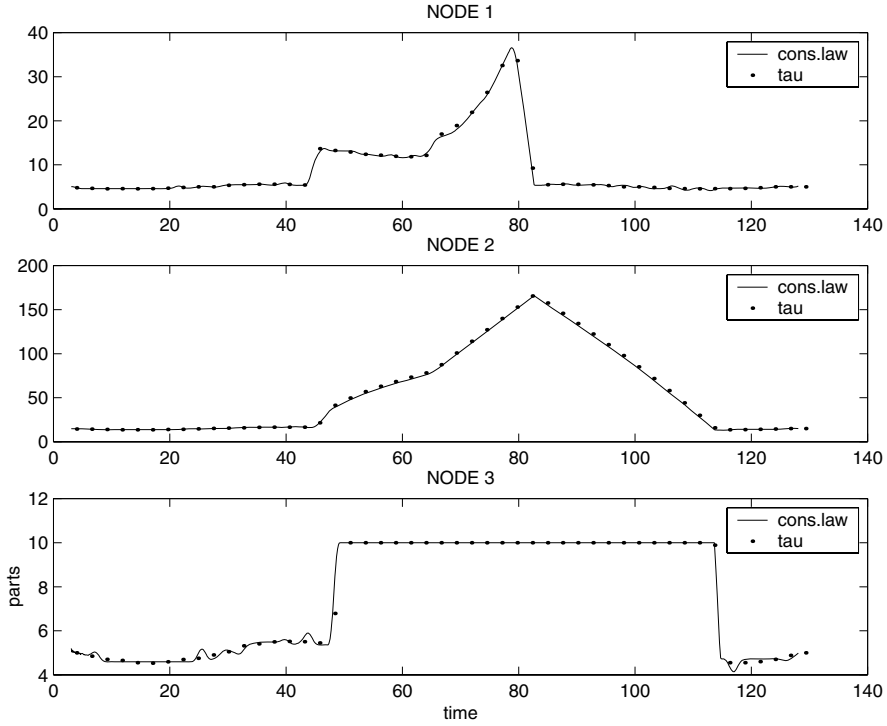


FIG. 2. *Work in progress.*

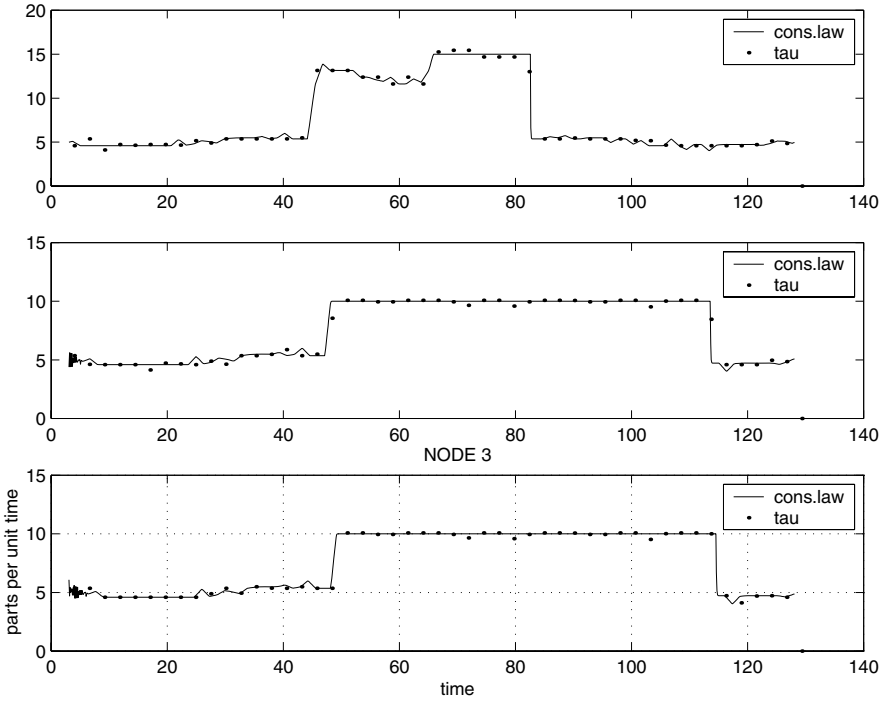


FIG. 3. *Outflux.*

corresponding WIP and Figure 3 shows the outflux of each node in the supply chain. The dots are computed from a time averaging of the solution of the recursion (8) for the transition times τ, and the solid line is computed from the conservation law (36). So, the WIP of node $S_0 m$ in the chain is computed as $\int_0^{0.2} \rho(x,t)\, dx$.

Figures 4, 5, and 6 depict the solution of the hyperbolic problem. Figure 4 shows the antiderivative of the density ρ in the DOC direction, i.e., $-u(x,t) + u(0,t)$, and Figure 5 shows the flux. As expected, we see bottlenecks, i.e., discontinuities, developing and vanishing again at $x = 0$ and $x = 0.2$. As long as the nodes in the supply chain work below capacity, i.e., as long as the governing equation is $\partial_t \rho + \frac{X}{MT_0} \partial_x \rho = 0$, we see the propagation of the fluctuations in the influx f^A through the system. As soon as the nodes go into saturation, i.e., as soon as $\partial_t \rho + \partial_x \mu$ holds, the solution becomes constant but develops discontinuities at the bottlenecks. Figure 6 shows the density ρ, which develops concentrations at $x = 0$ and $x = 0.2$, on a logarithmic scale.

As a second example, we consider a "long" supply chain with unstructured throughput times and capacities. We choose $M = 80$ and choose 80 random throughput times between $T = 1$ and $T = 5$ time units. For simplicity, we set $C(m) = \mu(m) = \frac{1}{T(m)}$, $m = 0, \ldots, M-1$. So each processor handles only 1 part per unit time, and we use the maximally possible relase rates μ. Figure 7 shows the corresponding mesh in the DOC variable $0 \le x \le X = 1$ and the capacities. So the meshsizes h_m are according to (19) randomly distributed. All the assumptions of Theorem 2 are satisfied, except that we cannot guarantee a relatively small number of bottlenecks, since

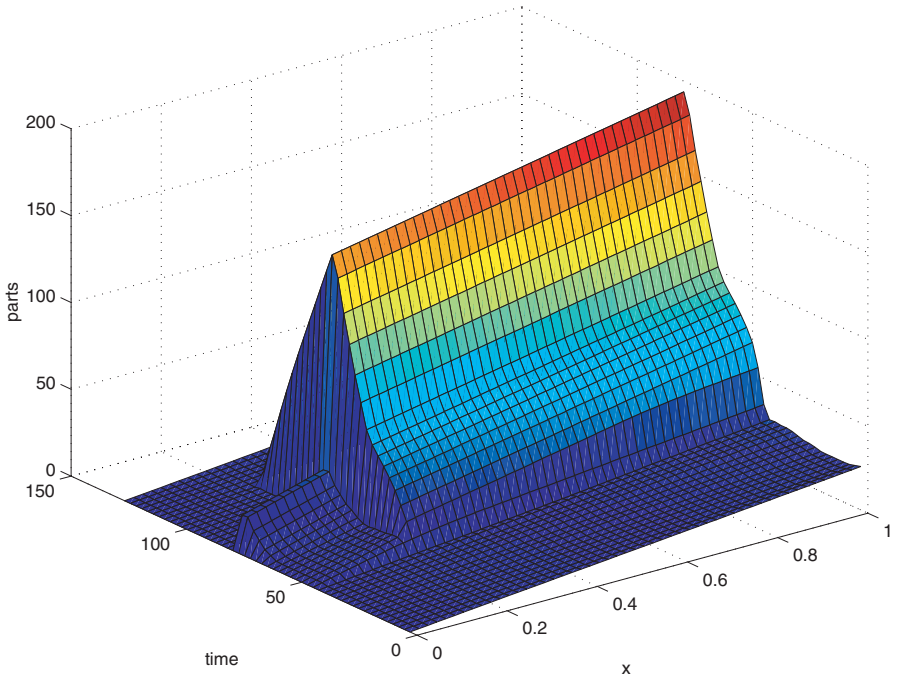


FIG. 4. N-*curve:* $u(0,t) - u(x,t)$.

the service rates μ now have an arbitrary number of significant drops. We choose $\Delta x_m = h_m$, i.e., we assign precisely one gridpoint per node for the numerical solution of the partial differential equation. Again, this goes beyond standard convergence theory since we do not resolve the rapidly varying function $\mu(x)$ in the continuous formulation. Figure 8 shows the influx and the outflux of the last node in the supply chain. Again, the dots denote the time averaged results computed from the recursion formula (8). The influx is chosen at and below the minimum capacity $C_{min} = 0.2$, with a spike at $t \approx 1500$. Figure 9 shows the corresponding total WIP of the whole supply chain. We observe almost perfect agreement although we have not resolved the service rate function $\mu(x)$ on the computational mesh. Figure 10 shows the density ρ on a logarithmic scale. We see the development of six bottlenecks. So, although the relationship between capacities of neighboring processors is completely random, the supply chain organizes itself to produce only a few bottlenecks and the assumptions of Theorem 2 are still satisfied.

6. Conclusions. We have derived a partial differential equation modeling a supply chain of arbitrary length with a large number of parts. Other than in similar approaches, this model is not based on some quasi-steady-state assumptions about the stochastic behavior of the involved queues, but rather on a simple deterministic rule for releasing parts from the buffer queues into the processors. The presented model incorporates the concept of the capacity of a processor in a natural way in a transient setting, while models based on queuing theory have to achieve this through a relation between throughput time and work in progress which is somehow extrapolated from the steady state situation. The model contains a distributed parameter (the service rates), which is constrained by the capacities, and can be used to control the behavior of the supply chain. It can be expected that relatively simple rules can be found governing these service rates which guarantee a certain behavior of the supply chain,

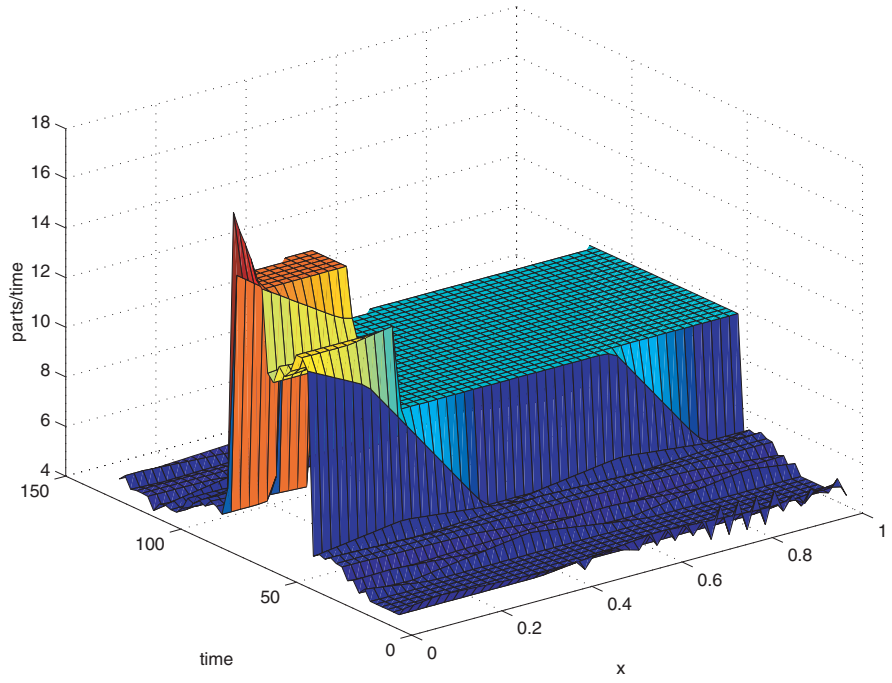


FIG. 5. *Flux* f.

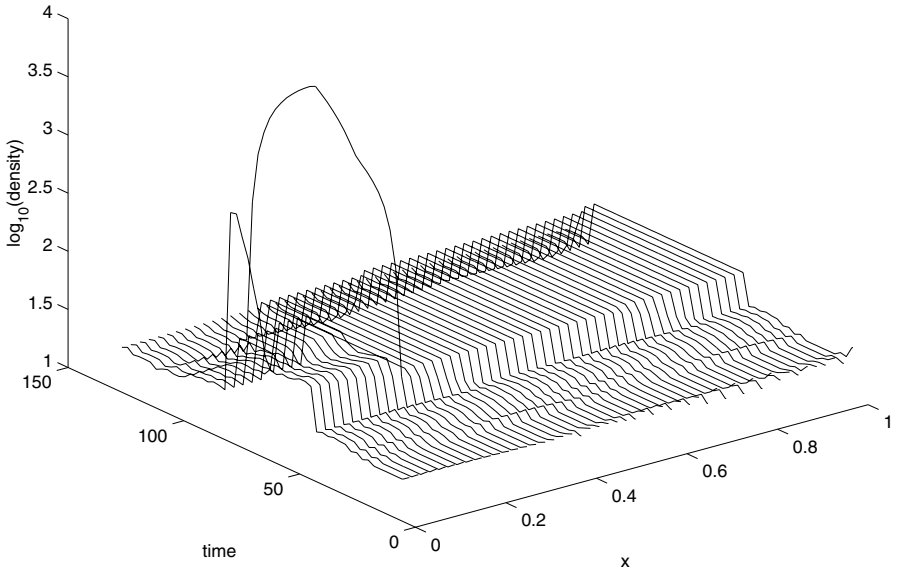


FIG. 6. *Density* ρ.

cf. the avoidance of bottlenecks.

7. Appendix.

Proof of Theorem 1. We first rewrite (8). Defining

$$\Delta_n \tau(m,n) := \tau(m,n+1) - \tau(m,n), \quad \Delta_m \tau(m,n) := \tau(m+1,n) - \tau(m,n),$$

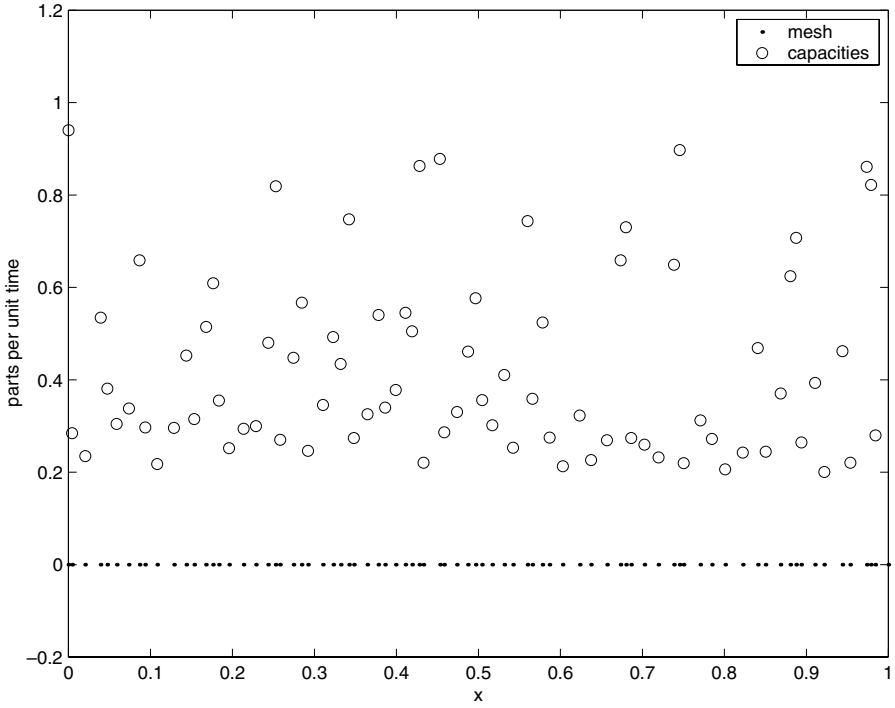


FIG. 7. *Mesh and capacities.*

(8) can be written as

$$0 = \min\left\{\Delta_m \tau(m-1,n+1) - T(m-1), \Delta_n \tau(m,n) - \frac{1}{\mu(m-1,n)}\right\}.$$

Using the definition (13) of ρ and f, this is equivalent to

$$0 = \min\left\{\frac{h_{m-1}\rho(x_{m-1},\tau(m,n))}{f(x_m,\tau(m,n))} - T(m-1), \frac{1}{f(x_m,\tau(m,n))} - \frac{1}{\mu(m-1,n)}\right\}.$$

To simplify the notation, we will drop the indices m and n and write the above as $\min\{\frac{h\rho}{f} - T, \frac{1}{f} - \frac{1}{\mu}\} = 0$. Furthermore, we will write this relation in the variable $z = \frac{1}{f}$. So we have to invert the function $\alpha(z,\rho)$, given by

$$y = \alpha(z,\rho) = \min\left\{h\rho z - T, z - \frac{1}{\mu}\right\}$$

as a function of z for any given parameter ρ, i.e., find a function $\beta(y,\rho)$ satisfying

$$y = \alpha(z,\rho) \quad \Longleftrightarrow \quad z = \beta(y,\rho).$$

If this is possible, then f is given in terms of ρ as $f = \frac{1}{\beta(0,\rho)} = \phi(\rho)$. There are two different cases to consider, namely the case $0 < h\rho < 1$ and the case $h\rho \geq 1$.

Case 1: $h\rho \geq 1$. In this case α is piecewise defined as

$$y = \alpha(z,\rho) = \begin{pmatrix} zh\rho - T & \text{for} & z < z_0 = \frac{T - \frac{1}{\mu}}{h\rho - 1} \\ z - \frac{1}{\mu} & \text{for} & z > z_0 \end{pmatrix}. \tag{37}$$

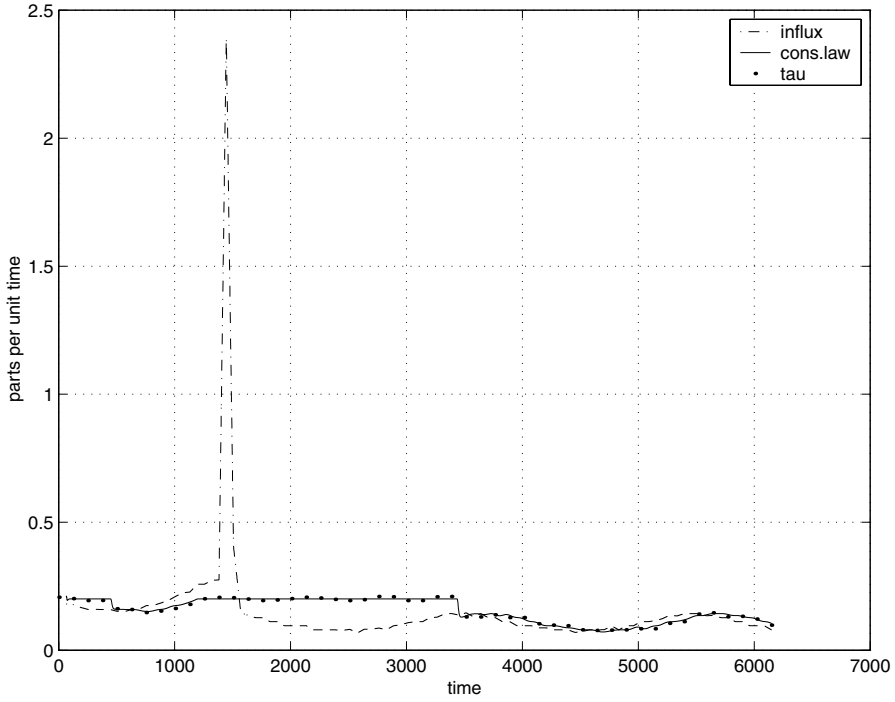


FIG. 8. *Influx and outflux.*

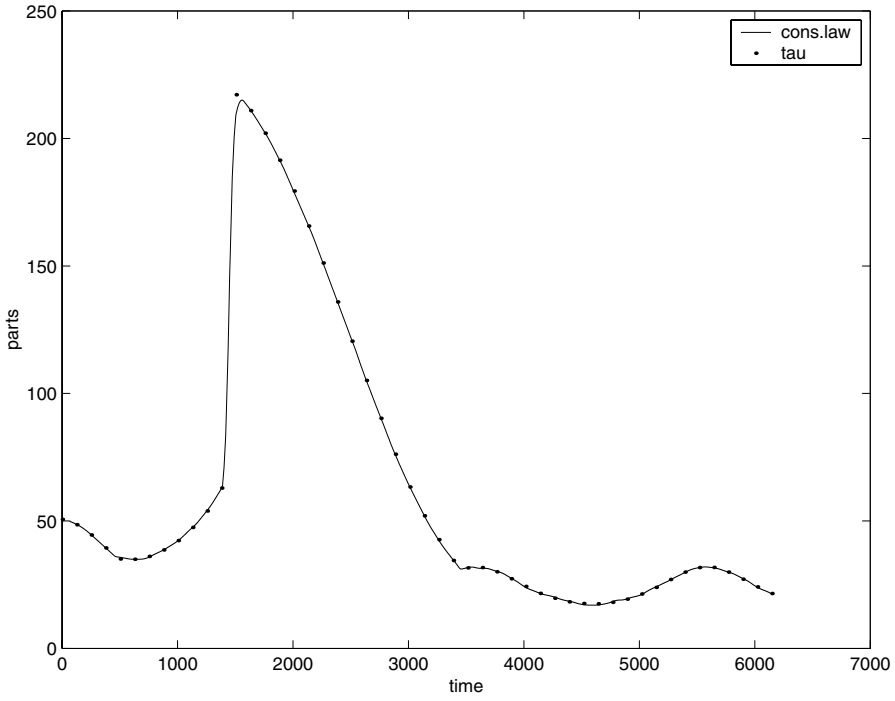


FIG. 9. *Total work in progress.*

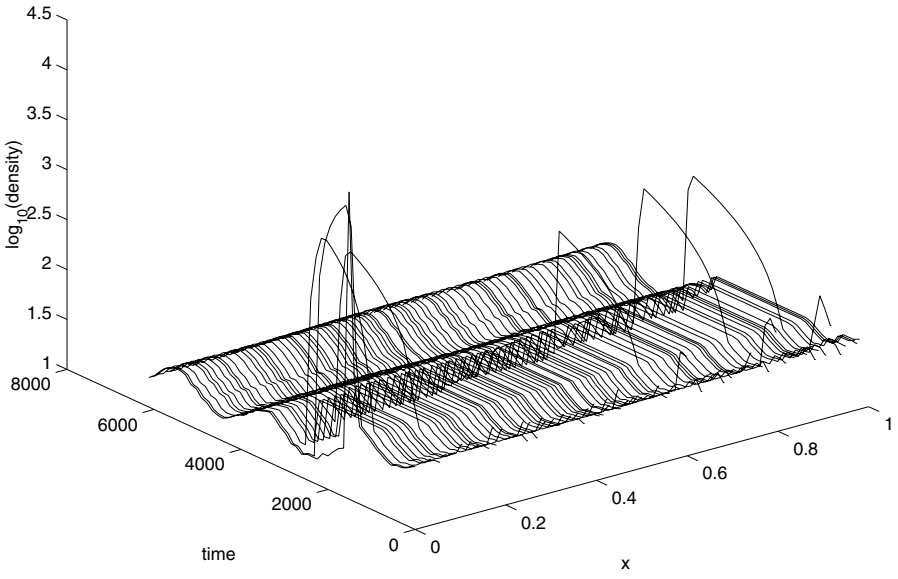


FIG. 10. *Part density* ρ.

(In the case $h\rho = 1$ the second case in (37) simply never occurs.) This monotonically increasing and piecewise linear function of z can be inverted as

$$z = \beta(y,\rho) = \begin{pmatrix} \frac{y+T}{h\rho} & \text{for} & y < y_0 := z_0 - \frac{1}{\mu} \\ y + \frac{1}{\mu} & \text{for} & y > y_0 \end{pmatrix}.$$

Evaluating β at $y = 0$ gives

$$\beta(0,\rho) = \begin{pmatrix} \frac{T}{h\rho} & \text{for} & 0 < y_0 = \frac{T - \frac{h\rho}{\mu}}{h\rho - 1} \\ \frac{1}{\mu} & \text{for} & 0 > y_0 \end{pmatrix} = \begin{pmatrix} \frac{T}{h\rho} & \text{for} & \frac{1}{\mu} < \frac{T}{h\rho} \\ \frac{1}{\mu} & \text{for} & \frac{1}{\mu} > \frac{T}{h\rho} \end{pmatrix} = \max\left\{\frac{1}{\mu}, \frac{T}{h\rho}\right\}.$$

Case 2: $0 < h\rho < 1$. In this case, we proceed in the same way obtaining the piecewise linear definition

$$y = \alpha(z,\rho) = \begin{pmatrix} z - \frac{1}{\mu} & \text{for} & z < z_0 = \frac{T - \frac{1}{\mu}}{h\rho - 1} \\ zh\rho - T & \text{for} & z > z_0 \end{pmatrix}$$

for the function α. Note, that the ranges for the linear pieces of α are now switched. Inverting α gives

$$z = \beta(y,\rho) = \begin{pmatrix} y + \frac{1}{\mu} & \text{for} & y < y_0 := z_0 - \frac{1}{\mu} \\ \frac{y+T}{h\rho} & \text{for} & y > y_0 \end{pmatrix},$$

and evaluating β at $y = 0$ gives

$$\beta(0,\rho) = \begin{pmatrix} \frac{1}{\mu} & \text{for} & 0 < y_0 = \frac{T - Dh\rho}{h\rho - 1} \\ \frac{T}{h\rho} & \text{for} & 0 > y_0 \end{pmatrix} = \begin{pmatrix} \frac{1}{\mu} & \text{for} & \frac{1}{\mu} > \frac{T}{h\rho} \\ \frac{T}{h\rho} & \text{for} & \frac{1}{\mu} < \frac{T}{h\rho} \end{pmatrix} = \max\left\{\frac{1}{\mu}, \frac{T}{h\rho}\right\}.$$

So, in both cases we obtain the same value for the inverse β evaluated at $y = 0$. Setting $\phi(\rho) = \frac{1}{\beta(0,\rho)}$ gives (14). ∎

Proof of Theorem 2. In order to avoid dealing with the boundary conditions, we extend the definition of the variables ρ, f, u, τ onto the whole real line $x \in \mathbb{R}$. We define

$$x_m = -mh_0 \text{ for } m \le -1, \quad x_m = X + (m-M)h_{M-1} \quad \text{for } m \ge M+1.$$

Next we extend the values of the arrival times τ for $m < 0$ and $m > M$. We set

$$\begin{aligned}
\tau^I(m) &= \tau^I(0) + mh_0, \quad m < 0,\\
\tau^I(m) &= \tau^I(M) + (m-M)h_{M-1}, \quad m > M,\\
\tau(m, n+1) &= \tau(m,n) + \varepsilon \Delta_n \tau^A(n), \quad m \le 0,\ n \ge 0,\\
\tau(m, n+1) &= \tau(m,n) + \varepsilon \Delta_n \tau(M,n), \quad m > M,\ n \ge 0.
\end{aligned}$$

With this definition the values of the flux function f_1 in (22) satisfy

$$f(x_m, \tau(m,n)) = \frac{1}{\Delta_n \tau(m,n)} = f^A(\tau(m,n)), \quad m < 0,\ n \ge 0,$$

and the corresponding interpolant f_2 in the continuous x-variable satisfies

$$\lim_{x \to 0-} f_2(x,t) = f^A(t). \tag{38}$$

Now we consider a compactly supported test function $\psi(x,t)$ and its discrete antiderivative Ψ given by

$$\Psi(x_m, t) = \sum_{m'=-\infty}^{m-1} h_{m'} \psi(x_{m'}, t), \quad \Psi(x_{m+1}, t) - \Psi(x_m, t) = h_m \psi(x_m, t). \tag{39}$$

Since ψ and Ψ are compactly supported in the time direction, we have for any fixed index n the trivial equality

$$\sum_{m=-\infty}^{\infty} [\Psi(x_{m+1}, \tau(m+1,n)) - \Psi(x_m, \tau(m,n))] = 0 \quad \forall n,$$

which we sum over the index n and multiply by ε, giving

$$0 = \varepsilon \sum_{n=0}^{\infty} \sum_{m=-\infty}^{\infty} [\Psi(x_{m+1}, \tau(m+1,n)) - \Psi(x_m, \tau(m,n))] = A - B,$$

and A and B are defined by

$$\begin{aligned}
\text{(a) } A &= \varepsilon \sum_{n=0}^{\infty} \sum_{m=-\infty}^{\infty} [\Psi(x_{m+1}, \tau(m+1,n)) - \Psi(x_m, \tau(m+1,n))]\\
&= \varepsilon \sum_{n=0}^{\infty} \sum_{m=-\infty}^{\infty} h_m \psi(x_m, \tau(m+1,n)),\\
\text{(b) } B &= \varepsilon \sum_{n=0}^{\infty} \sum_{m=-\infty}^{\infty} [\Psi(x_m, \tau(m,n)) - \Psi(x_m, \tau(m+1,n))].
\end{aligned} \tag{40}$$

We first estimate the spatial difference A. From the definition of the interpolant f_1 and the definition (17) of f, we have

$$\int_{\tau(m,n)}^{\tau(m,n+1)} f_1(x_m,t) = \varepsilon \Delta_n \tau(m,n) f(x_m, \tau(m,n)) = \varepsilon.$$

Inserting this into (40)(a) gives

$$\begin{aligned} A &= \sum_{n=0}^{\infty} \sum_{m=-\infty}^{\infty} h_m \psi(x_m, \tau(m+1,n)) \int_{\tau(m+1,n)}^{\tau(m+1,n+1)} f_1(x_{m+1},t)\ dt \\ &= \sum_{n=0}^{\infty} \sum_{m=-\infty}^{\infty} h_m \int_{\tau(m+1,n)}^{\tau(m+1,n+1)} \psi(x_m,t) f_1(x_{m+1},t)\ dt + O(\varepsilon) \\ &= \sum_{m=-\infty}^{\infty} h_m \int_{\tau^I(m+1)}^{\infty} \psi(x_m,t) f_1(x_{m+1},t)\ dt + O(\varepsilon), \end{aligned}$$

where we have committed an $O(\varepsilon)$ error by taking the test function ψ inside the integral. Because of the definition (24) of the interpolant f_2 in the spatial direction, the term $h_m f_1(x_{m+1},t)$ can be written as an integral with respect to x giving

$$A = \sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} \psi(x_m,t) \left[\int_{x_m}^{x_{m+1}} H(t-\tau_2^I(x)) f_2(x,t)\ dx \right]\ dt + O(\varepsilon),$$

$$A = \int\int H(t-\tau_2^I(x)) \psi(x,t) f_2(x,t)\ dx dt + O(\varepsilon), \tag{41}$$

where we have committed another $O(\varepsilon)$ error by taking the test function ψ inside the x-integral. Here H denotes the usual Heaviside function.

Now, we turn to the term B in (40)(b). We replace the difference in the time direction by a partial derivative, giving

$$B = -\varepsilon^2 \sum_{n=0}^{\infty} \sum_{m=-\infty}^{\infty} \partial_t \Psi(x_m, \tau(m+1,n)) \Delta_m \tau(m,n) + O(\varepsilon).$$

Again, we commit only an error of order $O(\varepsilon)$ in doing so, even if, according to the assumptions, a bounded number of the $\Delta_m \tau(m,n)$ is of order $O(\frac{1}{\varepsilon})$, since the test function Ψ will be bounded. We split the $n=0$ term in the sum and write

$$\begin{aligned} B = &-\varepsilon^2 \sum_{n=1}^{\infty} \sum_{m=-\infty}^{\infty} \partial_t \Psi(x_m, \tau(m+1,n)) \Delta_m \tau(m,n) \\ &-\varepsilon^2 \sum_{m=-\infty}^{\infty} \partial_t \Psi(x_m, \tau^I(m+1)) \Delta_m \tau^I(m) + O(\varepsilon). \end{aligned}$$

Clearly, the second term is of order $O(\varepsilon)$ again and can be neglected. Using the definition (17)(b) of ρ, we obtain

$$B = -\varepsilon \sum_{n=1}^{\infty} \sum_{m=-\infty}^{\infty} \frac{h_m}{X} \partial_t \Psi(x_m, \tau(m+1,n)) \rho(x_m, \tau(m+1,n-1)) \Delta_n \tau(m+1,n-1) + O(\varepsilon).$$

Now, we repeat essentially the same procedure used for the term A. From (17) and the definition of ρ_1 we obtain

$$\int_{\tau(m+1,n-1)}^{\tau(m+1,n)} \rho_1(x_m,t)\ dt = \varepsilon\Delta_n\tau(m+1,n-1)\rho(x_m,\tau(m+1,n-1))$$

$$\begin{aligned} B &= -\sum_{n=1}^{\infty}\sum_{m=-\infty}^{\infty} \frac{h_m}{X}\partial_t\Psi(x_m,\tau(m+1,n)) \int_{\tau(m+1,n-1)}^{\tau(m+1,n)} \rho_1(x_m,t)\ dt + O(\varepsilon) \\ &= -\sum_{m=-\infty}^{\infty} \frac{h_m}{X}\int_{\tau^I(m+1)}^{\infty} \partial_t\Psi(x_m,t)\rho_1(x_m,t)\ dt + O(\varepsilon). \end{aligned}$$

Using the definition of ρ_1 as the spatial difference of $-u_1$ gives

$$\sum_{m=-\infty}^{\infty}\int_{-\infty}^{\infty} H(t-\tau^I(m+1))\partial_t\Psi(x_m,t)[u_1(x_{m+1},t)-u_1(x_m,t)]\ dt + O(\varepsilon).$$

Regrouping the terms in the above expression yields

$$\begin{aligned} B &= -\sum_{m=-\infty}^{\infty}\int_{-\infty}^{\infty} H(t-\tau^I(m))\partial_t[\Psi(x_m,t)-\Psi(x_{m-1},t)]u_1(x_m,t)\ dt + O(\varepsilon) \\ &= -\sum_{m=-\infty}^{\infty}\int_{-\infty}^{\infty} H(t-\tau^I(m))h_{m-1}\partial_t\psi(x_{m-1},t)u_1(x_m,t)\ dt + O(\varepsilon) \\ &= -\sum_{m=-\infty}^{\infty}\int_{-\infty}^{\infty} \partial_t\psi(x_{m-1},t)\left[\int_{x_{m-1}}^{x_m} H(t-\tau_2^I(x))u_2(x,t)\ dx\right]\ dt + O(\varepsilon), \end{aligned}$$

giving altogether

$$B = -\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} \partial_t\psi(x,t)H(t-\tau_2^I(x))u_2(x,t)\ dx\right]\ dt + O(\varepsilon). \tag{42}$$

Combining (41) and (42) gives

$$\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} H(t-\tau_2^I(x))[\psi(x,t)f_2(x,t)+\partial_t\psi(x,t)u_2(x,t)]\ dx\right]\ dt = O(\varepsilon),$$

which, in the limit $\varepsilon\to 0$, is the weak formulation of

$$\partial_t u_2 = f_2,\ x\in\mathbb{R},\ t>\tau_2^I(x), \quad u_2(x,\tau^I(x)) = 0.$$

Because of the definition of $u_1(x_0,t)$ in (23) this is the solution of (25) on the interval $[0,X]$. □

REFERENCES

[1] E. J. Anderson, *A new continuous model for job shop scheduling,* Internat. J. Systems Sci., 12 (1981), pp. 1469–1475.

[2] D. Armbruster, D. Marthaler, C. Ringhofer, K. Kempf, and T.-C. Jo, *A continuum model for a re-entrant factory,* Oper. Res. 38, 2006, in print; preprint available online from http://math.la.asu.edu/∼chris.

[3] D. Armbruster, D. Marthaler, and C. Ringhofer, *A mesoscopic approach to the simulation of semiconductor supply chains*, in Proceedings of the International Conference on Modeling and Analysis of Semiconductor Manufacturing (MASM2002), G. Mackulak et al. eds., 2002, pp. 365–369.
[4] D. Armbruster, D. Marthaler, and C. Ringhofer, *Kinetic and fluid model hierarchies for supply chains*, Multiscale Model. Simul., 2 (2004), pp. 43-61.
[5] D. Armbruster and C. Ringhofer, *Thermalized kinetic and fluid models for reentrant supply chains*, Multiscale Model. Simul., 3 (2005), pp. 782–800.
[6] A. Aw and M. Rascle, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.
[7] J. Banks, J. Carson, II, and B. Nelson, *Discrete Event System Simulation*, Prentice-Hall, Englewood Cliffs, NJ, 1999.
[8] R. Billings and J. Hasenbein, *Applications of fluid models to semiconductor fab operations*, preprint, 2001.
[9] P. Brandt, A. Franken, and B. Lisek, *Stationary Stochastic Models,* Wiley, New York, 1990.
[10] C. Cercignani, *The Boltzmann Equation and Its Applications,* Appl. Math. Sci. 67, Springer-Verlag, Berlin, 1988.
[11] C. Daganzo, *Requiem for second order fluid approximations of traffic flow,* Transport. Res. B, 29 (1995), pp. 277–286.
[12] C. Daganzo, *A Theory of Supply Chains*, ITS Research report UCB-ITS-RR2001-7, 2001; also available online from http://www.ce.berkeley.edu/∼daganzo/.
[13] M. El-Taha and S. Stidham, *Sample Path Analysis of Queuing Systems,* Internat. Ser. Oper. Res. Management Sci. II, Kluwer Academic Publishers, Boston, 1999.
[14] D. Helbing, *Gas kinetic derivation of Navier Stokes like traffic equations,* Phys. Rev. E, 53 (1996), pp. 2366–2381.
[15] D. Helbing, *Traffic and related self-driven many particle systems,* Rev. Modern Phys., 73 (2001), pp. 1067–1141.
[16] M. Lighthill and J. Whitham, *On kinematic waves,* I: *Flow movement in long rivers,* II: *A theory of traffic flow on long crowded roads,* Proc. R. Soc. Lond. Ser. A Math. Phys. Engrg. Sci., 229 (1955), pp. 281–345.
[17] G. F. Newell, *A simplified theory of kinematic waves in highway traffic,* Transport. Res. B, 27 (1993), pp. 281–313.
[18] I. Prigogine and R. Herman, *Kinetic Theory of Vehicular Traffic,* Elsevier, New York, 1971.
[19] M. C. Pullan, *An algorithm for a class of continuous linear programs,* SIAM J. Control Optim., 31 (1993), pp. 1558–1577.

SIAM J. APPL. MATH.
Vol. 66, No. 3, pp. 921–944

STOCHASTIC MODELING AND SIMULATION OF TRAFFIC FLOW: ASYMMETRIC SINGLE EXCLUSION PROCESS WITH ARRHENIUS LOOK-AHEAD DYNAMICS*

ALEXANDROS SOPASAKIS† AND MARKOS A. KATSOULAKIS†

Abstract. A novel traffic flow model based on stochastic microscopic dynamics is introduced and analyzed. Vehicles advance based on the energy profile of their surrounding traffic implementing the "look-ahead" rule and following an underlying asymmetric exclusion process with Arrhenius spin-exchange dynamics. Monte Carlo simulations produce numerical solutions of the microscopic traffic model. Fluctuations play an important role in profiling observationally documented but, at the simulation level, elusive traffic phenomena. Furthermore, based on scaling and limit arguments we obtain a macroscopic description of this microscopic dynamics formulation which up to leading term of the expansions takes the form of integrodifferential Burgers or higher-order dispersive partial differential equations. We outline connections and comparisons of the hierarchical models presented here (microscopic, macroscopic) with other well-known traffic flow models.

Key words. traffic flow, look-ahead stochastic Arrhenius microscopic dynamics, Monte Carlo simulations

AMS subject classifications. 90B20, 60K30, 65C35, 65C05

DOI. 10.1137/040617790

1. Introduction and overview.

Building on the idea presented in [61], a new modeling approach of traffic flow is developed, which, based on Arrhenius microscopic stochastic dynamics [66], is adapted to vehicular traffic. We construct an asymmetric single exclusion process (ASEP) whose dynamics include interactions with other vehicles ahead ("look-ahead" rule). The model has remarkable attributes and similarities when compared to known observed traffic behavior and is able to predict most of the widely accepted traffic states at the microscopic level.

Traffic states that are commonly observed by researchers at the microscopic level are quite complex and at times even chaotic [22, 28, 29]. Some of the major traffic states that are commonly observed at the microscopic level are free flow, congested flow, synchronized traffic, and wide moving jams. Free flow is easily recognized by the ability of drivers to attain their desired speeds under very little, if any, interaction with other vehicles. Congested flow, in contrast, is characterized by heavy vehicular interactions and usually very low flows. In general in the regime above some critical density ($\approx$20% jam density), we usually observe the most interesting microscopic vehicular traffic phenomena. Among them, the phenomenon of synchronized traffic, as recently observed [29], may occur when above critical density and displays complex behavior. Synchronized traffic usually occurs at on ramps when vehicles are added to an already crowded highway. Although the corresponding flows are widely scattered, synchronized traffic is characterized by high vehicle flows and at the same time increasing vehicle densities [22, 27]. It has been observed [23, 28] that synchronized traffic breaks down (known as the "break-down" phenomenon) with a variety of "stop-

*Received by the editors October 27, 2004; accepted for publication (in revised form) June 28, 2005; published electronically March 3, 2006.
http://www.siam.org/journals/siap/66-3/61779.html
†Department of Mathematics and Statistics, Lederle Graduate Research Tower, University of Massachusetts, Amherst, MA 01003-9305 (sopas@math.umass.edu, markos@math.umass.edu). The research of the second author was partially supported by NSF-DMS-0413864 and NSF-ITR-0219211.

and-go" traffic waves and occurrence of even more chaotic phenomena [22, 28, 15]. Meta-stability and hysteresis [65, 9] effects (as vehicle densities increase) may also be associated with this behavior. Last, the so-called wide moving jams are localized structures moving upstream and are characterized by long (widthwise) waves whose "fronts are shorter than their width" [58]. Naturally vehicle speeds change sharply through the two fronts making up the wide moving jam. As those waves travel slowly upstream vehicles are forced to interact with a number of them. Wide moving jams appear also at densities above critical density [27, 29]. In general vehicles first transition from free flow to synchronized flow and then later, at a different location, we observe the transition to a wide moving jam [27, 58].

Attempts to model vehicular traffic date as far back as 1932 [19]. At first such attempts were mainly empirical and emphasized fitting parameters to the particular conditions at hand. However it is generally agreed that the "parameters of the model should be intuitively easy to calibrate and the corresponding values should be realistic" [22]. As a result, modeling approaches evolved and a variety of models and new approaches emerged, ranging from car-following to gas-dynamic to hydrodynamic and more recently cellular automaton to counter this problem.

Car-following (or follow-the-leader) models [17, 16, 54, 12, 69] appeared as a way of obtaining equations which can be used in a wider context than their (empirical) predecessors. As the name denotes, car-following models describe traffic behavior based on the vehicle leading closely up front. In some of these models [67] vehicles try to converge to their preferred (following) distance, relative to the vehicle in the front, thus creating an oscillatory behavior due to imperfect perception [29]. Clearly these models are not suitable, or designed, for large traffic streams.

The emergence of gas-dynamics-like (mesoscopic) [31, 25, 23, 64, 60, 59] and hydrodynamic-like (macroscopic) [52, 68] traffic models starting in the 1960s produced partial differential equations which attempt to describe traffic parameters of interest such as the density, velocity, and flow at larger time and spatial scales, thus encompassing larger traffic streams. Among them the fundamental model of Lighthill and Whitham [42], reflecting conservation of vehicles, is prominent. In its diffusive version [68], assuming a linear type of equilibrium velocity-density relationship [19], $V_{equil} = V_{max}(1 - c/c_{max})$ it takes the form of the Burgers equation,

$$\frac{\partial g}{\partial t} + g\frac{\partial g}{\partial r} = D\frac{\partial^2 g}{\partial r^2}, \tag{1.1}$$

where $D > 0$ is a diffusion constant, $g = V_{max}\,(1 - 2c/c_{max})$, and c denotes the density. We refer to [24] for further comments and references.

Optimal velocity models [47, 2, 64] emerged at almost the same time as car-following models. Newell [54] was the first to propose the following formulation for the equation of motion:

$$\frac{dx_j(t+\tau)}{dt} = V(\Delta x_j(t)),$$

where $x_j(t)$ is the position of vehicle j at time t and τ is a delay time. Here $\Delta x_j(t)$ is the headway of vehicle j at time t while V denotes a known optimal velocity rule. It has been shown [48] that some such optimal velocity traffic models can also be interpreted through the well-known mass transfer problem. Optimal velocity models have been used [47, 48, 34] to derive some of the well-known nonlinear wave equations such as Burgers, KdV, and modified KdV. Solutions of those nonlinear

wave equations are subsequently shown [48, 47] to describe different traffic regions: free flow, metastable region, unstable region. For an extended review of optimal velocity models see further comments and references in [48] and [6].

More recently, with the help of faster computers, cellular automaton (CA) microscopic traffic models have produced promising results when compared to traffic observations of similar spatial and time scales. Von Neumann introduced CA in 1950 in his abstract theory to study the logical conditions for self-reproducing machines [7]. It was, however, first due to Conway's game of life and later Wolfram [62, 70], which made CA well known in the dynamical systems community. The now famous rule CA 184, which has been studied in detail as a surface growth [36] (see also [14] for more comments), is widely implemented in CA traffic models. Since then the concept of CA has been applied and extended to model a wide variety of systems [39, 45].

To our knowledge [6] the first CA model for vehicular traffic was introduced by Cremer and Ludwig [8]. However, it was due to the contributions of Nagel and Schreckenberg [50] that CA models became widely known in traffic modeling. A number of improvements [32, 33, 49, 3] have been proposed since the original Nagel–Schreckenberg CA model, which also introduced random effects. These effects are deemed to better predict chaotic behavior of observed traffic at the microscopic scale. More recently a new CA, coded CA 184a, has been proposed by Nelson for traffic flow [53]. Similarly, development of discrete models proposed by Li [44], allows essential features of traffic to be captured and a better description of complex nonlinear phenomena to be studied.

In this work we extend the usual lattice-gas dynamics to vehicles. Based on an underlying stochastic model we reproduce some of the observed behavior of vehicular traffic when scrutinized under similar temporal and spatial scales. We assume a one-dimensional periodic (single-lane loop highway) lattice while we refer to [11] for generalizations (two-lane highway with entrances and exits). We start by obtaining suitable interpretations of many of the usual parameters of Ising systems and in many instances absorb as many of those parameters as possible to adapt our model to vehicular traffic observations. Special attention must be exercised to the application of the proper microscopic stochastic dynamics. Spin-exchange (diffusion) dynamics are therefore implemented to enforce conservation of vehicles. Overall we introduce the ASEP [13, 35, 40], guaranteeing that vehicles do not occupy the same site. An added novelty of this model is that vehicles are forced to move toward one direction since the dynamics, depending on spatial forward Arrhenius interactions, implement one-sided potentials and a look-ahead feature which can be considered to represent driver behavior. A variety of ASEP models without look-ahead Arrhenius interactions have been studied in [18]. Numerically we implement a kinetic Monte Carlo algorithm simulating aperiodic, consecutive (not parallel) microscopic stochastic dynamics. Among its many features the proposed stochastic model is able to predict spontaneous jam formation, "slow to start" (retarded acceleration) [63], and timely braking [58].

We also obtain kinetic mesoscopic PDEs derived in suitable asymptotic limits from the microscopic model which predict traffic observables for the appropriate validity ranges of these equations. Up to leading order we obtain Burgers and/or dispersive-type PDEs. We outline several connections between the hierarchical macroscopic models obtained through our expansions and other well-known traffic flow models [68, 54, 47, 34, 48, 28] and equations [26] of similar form. We examine these features in detail.

We start with an overview and derivation of the full model and present the details of the dynamics comprising our stochastic process in section 2. We subsequently

calibrate the free parameters of the stochastic model and produce the fundamental diagram (an important traffic engineering flow-density relationship) and other relevant solutions for our model in the numerical Monte Carlo simulations in section 3. In section 4 we obtain deterministic limit closures of our stochastic microscopic model, thus obtaining mesoscopic and macroscopic PDEs and systems of finite difference equations. Numerical comparisons of those models are presented in the same section. Final remarks and conclusions can be found in section 5. In the appendix we analyze briefly actual and theoretical data-gathering techniques which are in general of fundamental importance for the comparisons carried out in this work and elsewhere.

2. ASEP with Arrhenius look-ahead dynamics. We propose here an ASEP specifically equipped with conservative Arrhenius dynamics including a novel look-ahead parameter L. As a result we obtain a microscopic traffic flow model and its corresponding semigroup generator while introducing the notation to be used and general underlying assumptions.

We start by defining our physical space on a one-dimensional, periodic lattice, representing a one-lane loop highway. We partition our lattice into N cells, $\mathcal{L} = \{1, 2, \ldots, N\}$. On each of the lattice points $x \in \mathcal{L}$ we define an order parameter $\sigma(x)$ via

$$\sigma(x) = \begin{cases} 1 & \text{if a vehicle occupies site } x, \\ 0 & \text{if site at } x \text{ is empty (no vehicle).} \end{cases} \tag{2.1}$$

A spin configuration σ is an element of the configuration space $\Sigma = \{0,1\}^{\mathcal{L}}$ and we write $\sigma = \{\sigma(x) : x \in \mathcal{L}\}$ denoting by $\sigma(x)$ the spin at x.

Similarly to the Arrhenius dynamics for Ising systems [66] we let the interaction potential (which here will dictate the local behavior of vehicles) have the form

$$U(x, \sigma) = \sum_{\substack{y \in \mathcal{L} \\ y > x}} J(y - x)\sigma(y), \tag{2.2}$$

where J denotes an anisotropic short range intervehicle interaction potential,

$$J(x, y) = \gamma V\left(\gamma(y - x)\right), \quad x, y \in \mathcal{L}, \tag{2.3}$$

where $\gamma = 1/(2L+1)$ is a parameter prescribing the range of microscopic interactions and therefore L denotes the potential radius. Note that (2.2) in effect implements the interactions with vehicles up to range L ahead of the vehicle at position x, thus enforcing the look-ahead rule. Physically since a vehicle (or cell) length is taken to be 22 feet, then the look-ahead rule which is enforced by (2.2) allows drivers to see traffic at a physical distance which is $L \cdot 22$ feet ahead of their vehicle. Specifically we let $V : R \to R$ and set

$$V(r) = 0, \quad r \in R^{-} \quad \text{and} \quad V(r) = 0 \quad \text{for} \quad r \geq 1.$$

Note that the interaction potential $U(x)$ could be further enriched with the addition of an external potential h. This potential $h \equiv h(x, t)$ could vary in space but also time if so desired to account for temporal and spatial traffic situations (i.e., rush hour traffic).

There are several different choices for dynamics which we can employ: Arrhenius, Metropolis, Kawasaki, etc. We implement spin-exchange Arrhenius dynamics. Under this engine the simulation is driven based on the energy barrier a particle has

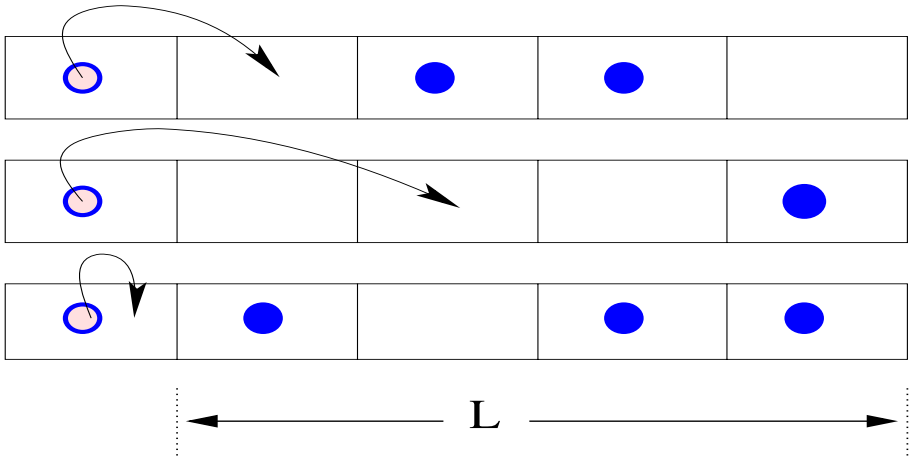


FIG. 1. *Simple schematic of the look-ahead rule (for $L = 4$ cells) pertaining to vehicle motion in the lattice for three different traffic examples. The process automatically simulates effects such as braking, acceleration, and simple exclusion rule through the interaction potential $U(x, \sigma)$ (2.2) for the provided range, L.*

to overcome in changing from one state to another. This energy barrier is found by calculating the potential energy of each vehicle based on (3.1) and performing a move only if that energy is higher than a given threshold. (Note that these dynamics are different from the usual Metropolis dynamics, where a move is encouraged whenever the energy difference between the current position and the new position is high enough.)

During such a spin-exchange between nearest neighbor sites x and y the system will actually allow the order parameter $\sigma(x)$ at location x to exchange value with the one at y. This is interpreted as advancing a vehicle from the site at x to the empty site at y. Note that based on the construction of our potential U it is not possible to move from an occupied site to another occupied cite; see Figure 1. In general, the rate at which a process will do this for spin-exchange Arrhenius dynamics is

$$c(x, y, \sigma) = \begin{cases} c_0 \exp[-U(x, \sigma)] & \text{if } y = x + 1, \text{ and } \sigma(x) = 1, \ \sigma(y) = 0, \\ & \text{if } \sigma(x) = 0 \ \ \sigma(y) = 1, \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

The parameters comprising the dynamics here are

$$c_0 = 1/\tau_0 \tag{2.5}$$

with τ_0 the characteristic or relaxation time for the process and $U(x, \sigma)$ as in (2.2).

Overall given (2.4) and the dynamics just described the probability of spin-exchange between x and y during time $[t, t + \Delta t]$ is

$$c(x, y, \sigma)\Delta t + O(\Delta t^2). \tag{2.6}$$

Clearly for one-lane traffic y corresponds to either $x - 1$ or $x + 1$ in (2.4). Note that the exchange, due to the specific construction of the interaction potential J in (2.3), can take effect if and only if the location at x is occupied while the location at y is not. A simple schematic of the lattice and some simple interactions are provided in Figure 1. At the same time, vehicles are restricted (exclusion rule) with performing

an exchange move backward—an unrealistic move for vehicular traffic—by definition of the dynamics (2.4). Note we do not allow backward moves at all since they are not part of the dynamics.

In summary this model is determined by the characteristic time, interaction strength, and look-ahead

$$\tau_0, J_0 \text{ and } L,$$

respectively, and driven by our Arrhenius dynamics stochastic process $\{\sigma_t\}_{t\geq 0}$ with rate (2.4).

The parameter c_0 will allow us to set the maximum speed of vehicles (which we do in the calibration section (3.1)) while the other parameters L and J_0 will be chosen appropriately based on known traffic behavior or other physical constraints which we present in our numerical investigations.

More rigorously we can define the generator for this stochastic process $\{\sigma_t\}_{t\geq 0}$. Let f be an arbitrary test function defined on $L^\infty(\Sigma)$, where Σ is the configuration space $\Sigma = \{0,1\}^{\mathcal{L}}$. Then the generator, M, of the process is defined,

$$Mf(\sigma) = \sum_{x\in\mathcal{L}} c_0\sigma(x)(1-\sigma(x+1))\exp(-U(x,\sigma))[f(\sigma^{x,x+1}) - f(\sigma)] \tag{2.7}$$

with c_0 from (2.5) and

$$\sigma^{x,x+1} = \begin{cases} \sigma(y), & y \neq x, x+1, \\ \sigma(x+1), & y = x, \\ \sigma(x), & y = x+1, \end{cases}$$

denotes an exchange of the spins between locations x and $x+1$. We refer to [40, 41] for background on generators and exclusion processes.

3. Monte Carlo simulations and benchmarks. We now present numerical implementations of the microscopic stochastic traffic flow model derived in section 2. We start by physically interpreting the nondimensional model variables and subsequently calibrating our model for the free parameters τ_0, J_0 and L with respect to well-known quantities from real traffic data. Last, as a basic benchmark, we also obtain the fundamental diagram corresponding to the flow-density relationship dictated by our model.

In the simulations we choose a simple constant potential of the form

$$V(r) = \begin{cases} J_0 & \text{if } 0 < r < 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{for} \quad J_0 > 0, \tag{3.1}$$

where J_0 is a parameter which based on its sign describes attractive repulsive or no-interactions. In our simulations we implement $J_0 > 0$, which implies that vehicles are attracted by the empty space in front of them. Note that (3.1) imposes only forward interactions and the dynamics (2.4) do not allow for backward moves. Thus vehicles move toward the empty space in the highway with an exclusion rule since we allow only one vehicle per site; see Figure 1. We have therefore created an ASEP [35] with interactions in one direction. In fact, the proposed dynamics reduce to an ASEP-type evolution if a particle/vehicle does not have any other vehicles in front of it beyond the interaction radius L in (2.3). In our numerical simulations we implement piecewise constant ($J_0 = 1$), short-range local interactions. However, we can also impose a V, in the form of a one-sided Maxwellian, which could be more realistic.

We let the actual physical length of each cell to be 22 feet. This allows for the average vehicle length plus safe distance. Therefore for a vehicle which has an average speed of 65 miles per hour we obtain a natural estimate of time to cross a cell,

$$\Delta t_{cell} = \frac{22 \text{ feet}}{65 \text{ miles/hour}} \approx \frac{1}{4} \text{ sec.} \tag{3.2}$$

In this work we start by modeling an infinite length road (ring road) with no entrances or exits so as to observe certain known traffic behavior and compare with other similar works [50].

Given the (one-sided) structure of the potential (2.2) we allow vehicles (or drivers) to be able to perceive traffic forward of their position up to a possible distance (look-ahead) of four cells or up to 88 feet—vehicle lengths plus safe distance. This is automatically implemented in the calculation of the interaction potential (2.2) and therefore plays a decisive role in the decision for making a move to a new location ahead through (2.4) and subsequently (2.6). However, we have also run simulations (not presented here) with a look-ahead of three and two vehicle lengths for comparison purposes. Further remarks regarding the influence of the look-ahead parameter can be found in section 4.4 and especially Figure 9.

We implement a kinetic Monte Carlo (KMC) simulation based on [4]. We refer to [61] for details on how the spin-exchange Arrhenius dynamics algorithm is applied. As expected, a KMC algorithm produces no null steps and therefore every iteration is a success. In that respect our KMC algorithm continues to choose and move vehicles at every step by skipping the idle waiting which occurs in usual Monte Carlo simulations and simply adjusting the simulation time by the appropriate amount, as if it had waited for that long. This is quite useful for the cases of high densities of vehicles or even more generally when a process reaches equilibration.

3.1. Calibration and validity. We start by calibrating our code through the free parameters J_0 and τ_0 by simulating a free-flow regime where we expect all vehicles to drive at their desired speed. We set such a speed to be 65 miles per hour. This is accomplished by the characteristic time τ_0, which allows us to calibrate the maximum velocity at which vehicles would like to drive assuming no other vehicles up front. Naturally, due to the stochasticity inherent in our simulation some vehicles will drive faster while some slower than the set limit of 65 miles per hour. As pointed out earlier the free parameter J_0 indirectly influences how drivers react to conditions in front of them and subsequently allows us to set the velocity of an upstream front (which researchers estimate it to be approximately -10 miles per hour [22, 58]). In Figure 2 we record all the parameters used and the average velocities of a stream of traffic which is initially randomly distributed under free-flow conditions. Note that for the chosen parameters $\tau_0 = .23$ and $J_0 = 6$ we obtain the desired velocity of 65 miles per hour and velocity out of a jam of ≈ -10 miles per hour. Also note that other pairs of τ_0 and J_0 are possible which easily adjust the traffic model for different standards set in other countries or regions.

3.2. Monte Carlo simulations. Using the calibrated parameters for J_0 and τ_0 from section 3.1 we now obtain the fundamental diagram (see the Appendix for details), the density—flow relationship in Figure 3, and the flow-velocity relationship in Figure 4. In general we can implement a number of different types of initial conditions which we make specific for each example considered. For these figures we use a random initial vehicle distribution and observe the behavior of the traffic stream as density increases incrementally.

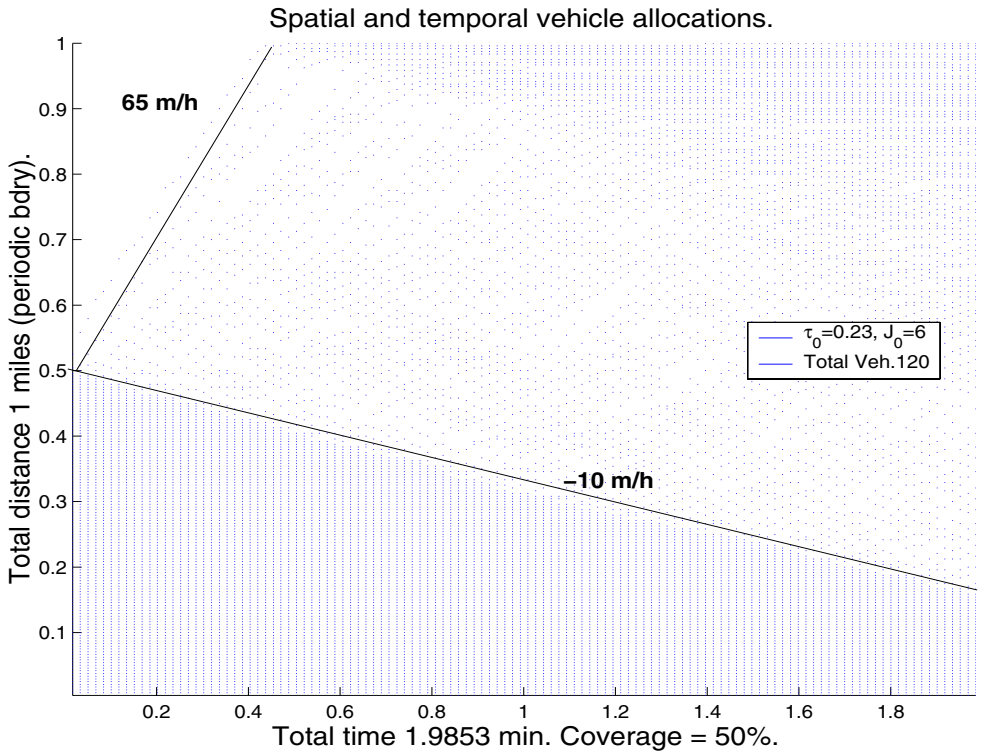


FIG. 2. *Calibration of parameters permitting a desired vehicle speed of* 65 *miles per hour and upstream front velocity of* ≈ -10 *miles per hour. We take a look-ahead of* $L = 4$. *Initial density corresponds to a traffic release problem (i.e., bumper-to-bumper vehicles up to* 0.5 *miles and no vehicles after that).*

It is of particular interest to analyze the simulated results for the well-established behavior of actual traffic at high densities, which includes stop-and-go waves, congested traffic, and other such interesting phenomena, all of which could be identified at the microscopic level. The phenomena observed at small time scales disappear in long averaged runs of the microscopic models. Cassidy [5] and Munoz and Daganzo [56] suggest how to eliminate two-regime flow by appropriately filtering the data. It is clear that these traffic states could not possibly be predicted, at their full complexity, as solutions of deterministic systems of differential equations [10] since they almost disappear as we aggregate observables. This is a very important point which should be emphasized. It is in fact possible that a given PDE traffic model could be just an extension (in the correct asymptotic limit) of a given microscopic model. In fact, we show how this can be done for our stochastic traffic flow model in the following section. Note, however, that caution is in order here since long-time averages may be unreliable due to the possible presence of phase transitions. In this model, since metastability is expected, for some concentration regimes (near c_{crit}), the simulations should be analyzed under small-time averages (see the appendix). Metastability, phase transitions, and hysteresis effects in general are quite difficult to conclusively detect and as such we propose to further study them in a forthcoming publication.

Several very interesting observations can be made from Figures 3 and 4. We compare our results with those of Nagel and Schreckenberg [50] but also with observations

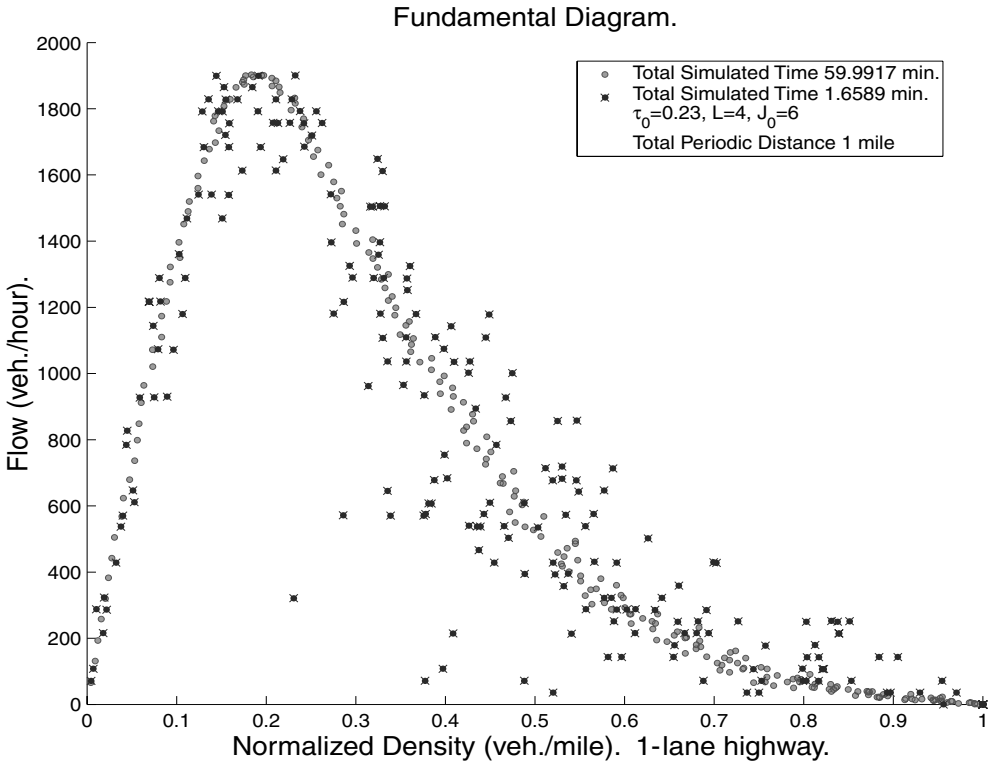


FIG. 3. *The flow density relationship for one-lane highway. Spatial periodic length of* 1 *mile, relaxation time* $\tau_0 = .23$*, interaction strength* $J_0 = 6$*, and look-ahead* $L = 4$ *cells. We superimpose two plots in this figure. In the first plot each point of the fundamental diagram is averaged over the usual* 1.65 *minutes aggregation time of data while in the second we display the long averaged flow and density for an aggregation time of a total of* 1 *hour. The aggregation time of* 1.65 *minutes was selected so that we can compare with observed data in* [51].

from Wiedemann [69] and observe qualitative agreement. Specifically, the region of free flow is clearly displayed up to approximately 50 vehicles per mile. Note here that the value of $c_{crit} = 50$ vehicles per mile is not forced on our simulation but instead is naturally created by the process dynamics through the calibration of the two parameters J_0 and τ_0. Similarly we observe a maximum vehicle flow of approximately 2000 vehicles per hour, which also agrees with observations [69], [51], and Figure 1(b) from [22]. (The aggregation time of 1.65 minutes was selected so that we can compare with observed data in [51].) An interesting question is whether this long averaged flux coincides with the equivalent mesoscopic partial differential equation of the stochastic limit. We answer this in section 4. The fluctuations in vehicle flows shown in Figures 3 and 4 are sizable for densities above c_{crit} and display a smeta-stable stop-and-go region. We should point out here that we did not fit parameters to obtain c_{crit} or q_{max} yet these parameters scale (for one-lane traffic) in agreement with observations of $c_{crit} \approx 50$ vehicles per miles and $q_{max} \approx 2000$ vehicles per hour [69].

In Figure 5 we display a close-up of vehicle trajectories for a congested traffic setup (50% jam density). In that figure we also observe that jams randomly form and sometimes disappear in time. In fact it can be easily calculated that these jams move backward in traffic with a speed of ≈ -10 miles per hour, which closely agrees with the reported speed of -15 ± 5 km per hour from [29, 58, 22].

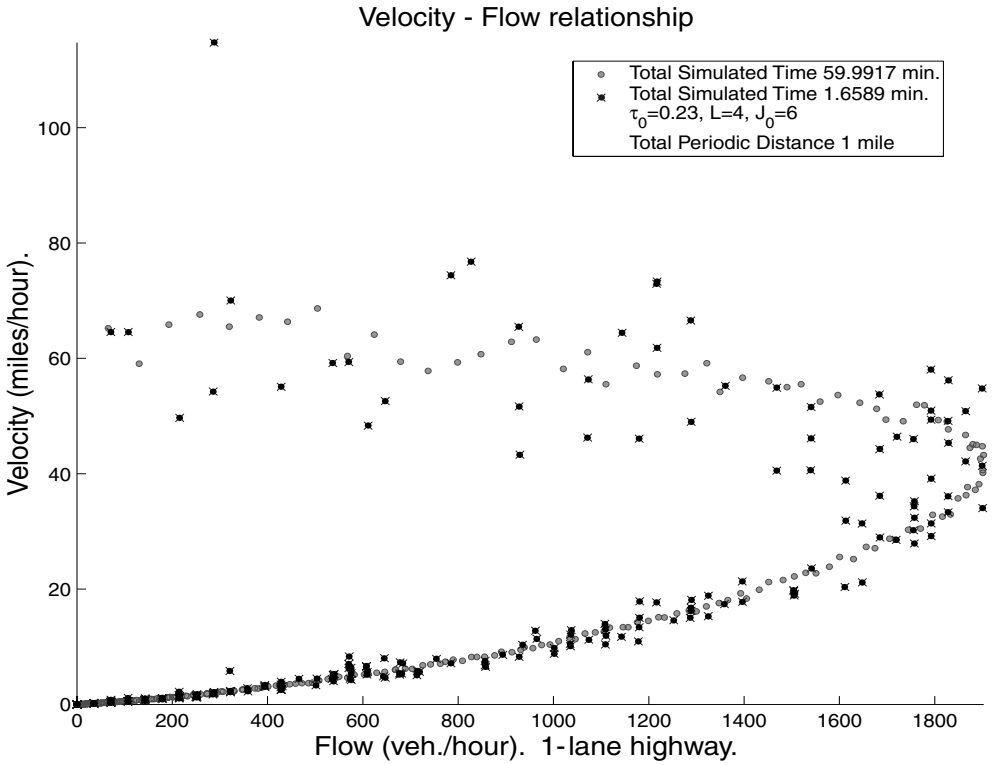


FIG. 4. *Velocity versus flow relationship for one-lane highway. Spatial periodic length of* 1 *mile, relaxation time* $\tau_0 = .23$*, interaction strength* $J_0 = 6$*, and look-ahead* $L = 4$ *cells. Once again we superimpose the* 1.65*-minutes aggregation time over the* 1*-hour aggregation time. This figure also compares favorably with observed data in* [51].

A consequence of the model versatility and ease of calibration presented is that calculations can be performed in real time for prediction purposes. The size of the traffic stream simulated in the numerical simulations here is small (< 3000 vehicles) and therefore the algorithm presented can easily produce real-time predictions. However, an improvement applicable for large traffic streams (> 3000 vehicles) is under development in [11], which uses a coarse graining idea and also produces real-time predictions. This has obvious important consequences since, as an example, we could possibly obtain the traffic input at a section of a highway in real time and immediately predict whether a traffic jam for the given highway capacity is imminent downstream, thus diverting traffic before the problem occurs.

4. Average behavior and deterministic closures. Although the emphasis of this paper is on new stochastic models we would like to establish connections with known CA and PDE models in various asymptotic regimes where mean field theory applies. Therefore we formally derive here a kinetic formulation of the stochastic model which is subsequently used to obtain an approximating finite difference (FD) scheme from (4.2) in the case of weak long-range interactions. Further, we also derive a PDE by a rescaling argument.

From our definition of a generator (2.7) we have

$$\frac{d}{dt}Ef(\sigma) = EMf(\sigma) \tag{4.1}$$

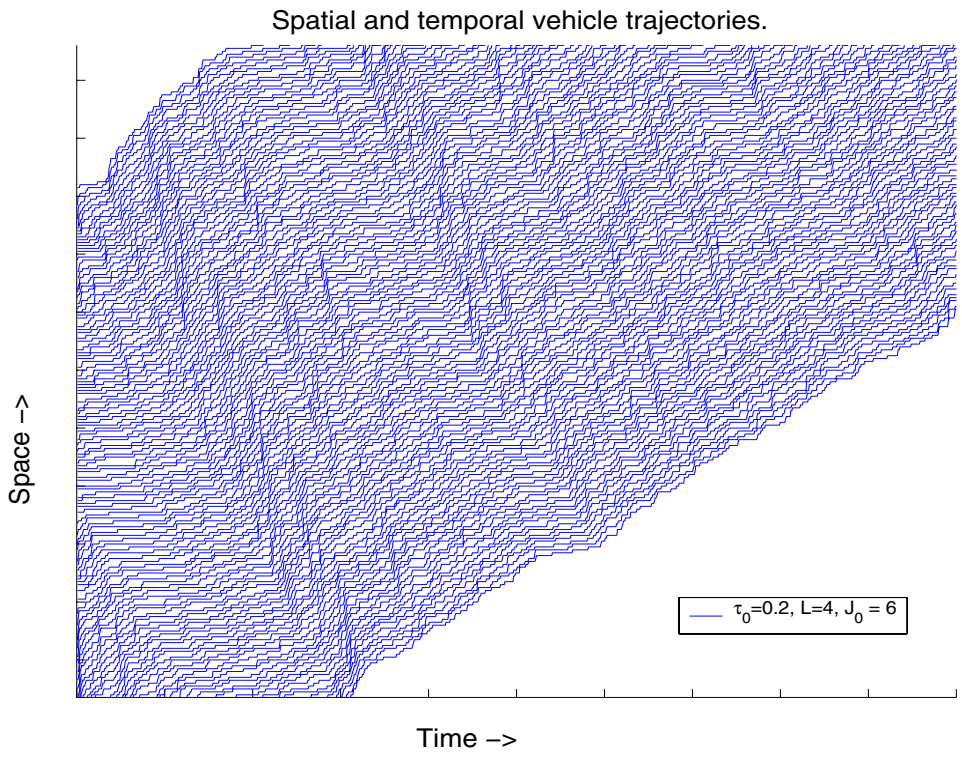


FIG. 5. *Vehicle stream lines in time and space. Spatial periodic length of* 1 *mile (corresponding to a total of* 240 *cells). We allow* 50% *initial capacity, relaxation time* $\tau_0 = .23$*, interaction strength* $J_0 = 6$*, and look-ahead* $L = 4$ *cells. Note the traffic waves in the stream displaying fronts and rarefactions.*

for any test function f, where $\sigma = \{\sigma_t : t \geq 0\}$ is the process corresponding to (2.7) and E denotes expected value. In particular we pick $f(\sigma) = \sigma(z)$ for z fixed in $\mathcal{L}$. Therefore

$$f(\sigma^{x,x+1}) = \sigma^{x,x+1}(z) = \begin{cases} \sigma(z), & x \neq z, x \neq z-1, \\ \sigma(z+1), & x = z, \\ \sigma(z-1), & x = z-1, \end{cases}$$

and based on (2.7) we have

$$f(\sigma^{x,x+1}) - f(\sigma) = \begin{cases} 0, & z \neq x, x+1, \\ \sigma(z+1) - \sigma(z), & x = z, \\ \sigma(z-1) - \sigma(z), & x = z-1. \end{cases}$$

By (2.7) and (4.1) we have the relation

(4.2)
$$\frac{d}{dt} E\sigma_t(z) = -Ec_0\sigma(z)(1-\sigma(z+1))e^{-U(z,\sigma)} + Ec_0\sigma(z-1)(1-\sigma(z))e^{-U(z-1,\sigma)}.$$

Note that relation (4.2) is exact and can be used to evaluate the closures discussed below. However, it is not yet a closed equation for $E\sigma_t(x) = \text{Prob}(\sigma_t(x) = 1)$.

4.1. Finite difference scheme. Suppose now that J (for J_0 fixed) has fairly long and weak interaction. We may assume that the stochastic process in (2.7) is a perturbation of the simple exclusion process considered in [40]. This process has a Bernoulli product invariant measure; thus at local equilibrium the probability measure is expected to be approximately a product measure. As in [57] we assume "propagation of chaos" for the microscopic system, in which case the fluctuations of the spins $\{\sigma(x), x \in \mathcal{L}\}$ about their mean values are independent and the law of large numbers formally applies. Thus the fluctuations of $\sum_{y\neq x} J(y-x)\sigma(y)$ about their mean will be small such that in the long-range interaction limit we have

$$Ee^{-U(x,\sigma)} = Ee^{-\sum_{y\neq x} J(y-x)\sigma(y)} \overset{N,L\to\infty}{\approx} e^{-\sum_{y\neq x} J(y-x)E\sigma(y)} + o_N(1). \tag{4.3}$$

Using the product property in (4.3) we formally obtain that the right-hand side of (4.2) becomes

$$-E\sigma(z)E(1-\sigma(z+1))c_0e^{-J\circ E\sigma(z)} + E\sigma(z-1)E(1-\sigma(z))c_0e^{-J\circ E\sigma(z-1)} + o_N(1), \tag{4.4}$$

where for an arbitrary function $v(z)$ we define

$$J\circ v(z) := \sum_{\substack{y\in\mathcal{L}\\ y>z}} J(y-z)v(y).$$

Next we drop $o_N(1)$ in (4.4) and define the density $u(z,t) = E\sigma_t(z)$. ($u(z,t)$ is the probability that site z is occupied at time t.) Then (4.2) and (4.4) give the following approximate semidiscrete FD scheme:

$$\begin{aligned}\frac{d}{dt}u(z,t) =& -c_0u(z,t)(1-u(z+1,t))\exp(-J\circ u(z,t))\\ &+c_0u(z-1,t)(1-u(z,t))\exp(-J\circ u(z-1,t)) \quad \text{for all } z\in\mathcal{L}.\end{aligned} \tag{4.5}$$

Note that our semidiscrete FD scheme (4.5) is conservative. Simply define

$$F(z,t) = c_0u(z-1,t)(1-u(z,t))\exp(-J\circ u(z-1,t)) \tag{4.6}$$

and without loss of generality, assuming a periodic lattice of N nodes, we sum the right-hand side of (4.5) and obtain

$$\begin{aligned}\sum_{z=0}^{N-1} c_0[-F(z+1,t)+F(z,t)] &= c_0[F(0,t)-F(N-1,t)]\\ &= c_0[u(-1,t)(1-u(0,t))\exp(-J\circ u(-1,t))\\ &\quad -u(N-1,t)(1-u(N,t))\exp(-J\circ u(N-1,t))] = 0\end{aligned}$$

since $u(-1,t) = u(N-1,t)$ and $u(0,t) = u(N,t)$ due to spatial periodicity.

Note that based on (4.6) we can rewrite (4.5) as

$$\frac{du(z,t)}{dt} + F(z+1,t) - F(z,t) = 0.$$

4.2. PDE limit. We now formally obtain the resulting PDE from (4.5) and make connections with other known traffic flow models. We start by expanding the spatial variables in Taylor series. We set $h = \Delta x$ and use

$$u(z \pm 1, t) = u(z,t) \pm h u_z(z,t) + \frac{h^2}{2} u_{zz}(z,t) + \cdots .$$

Substituting into (4.5) we obtain

$$\frac{d}{dt}u + h\left[u(1-u)c_0 \exp(-J \circ u)\right]_z = O(h^2). \tag{4.7}$$

Rescaling time in (4.5),

$$t \to th^{-1},$$

to absorb h and by omitting the $O(h^2)$ terms we have

$$\frac{d}{dt}u + [u(1-u)c_0 \exp(-J \circ u)]_z = 0 \tag{4.8}$$

$$\text{for } z \in R \quad \text{and} \quad J \circ u(z) = \int_z^\infty J(y-z)u(y)\, dy.$$

The transport equation obtained is

$$u_t + F(u)_z = 0, \tag{4.9}$$

where

$$F(u) = u(1-u)c_0 \exp(-J \circ u). \tag{4.10}$$

It is interesting to point out here that (4.10) under the simplest case of no interactions ($J_0 = 0$) corresponds to the well-known [68, 42] Lighthill–Whitham flux (1.1), thus producing a traffic stream formulation equivalent in form to the Burgers equation.

Note that in fact (4.5), when fully discretized, provides a natural finite difference scheme for (4.9),

$$\begin{aligned} u(z, t^{n+1}) = u(z,t^n) + \frac{\Delta t}{h}[&u(z-1,t^n)(1-u(z,t^n))c_0 \exp(-J \circ u(z-1,t^n)) \\ &- u(z,t^n)(1-u(z+1,t^n))c_0 \exp(-J \circ u(z,t^n))]. \end{aligned} \tag{4.11}$$

4.3. Stochastic versus semidiscrete approximation. We numerically compare the solutions of the stochastic model (4.2) against the semidiscrete scheme (4.5). In that respect we implement the same initial normalized density of vehicles (a traffic light release type) for both schemes with the following form:

$$u(z,0) = \begin{cases} 1 & \text{for cells in } 20 < z < 40, \\ 0 & \text{otherwise.} \end{cases} \tag{4.12}$$

The comparisons are performed under the assumption in (4.3) (i.e., $L = 240$ cells) for the semidiscrete scheme while the stochastic is averaged over several realizations to obtain an approximation of $E(\sigma_t(z))$. The solutions are shown in Figures 6 and 7 for the semidiscrete and stochastic models, respectively. The final profiles of these

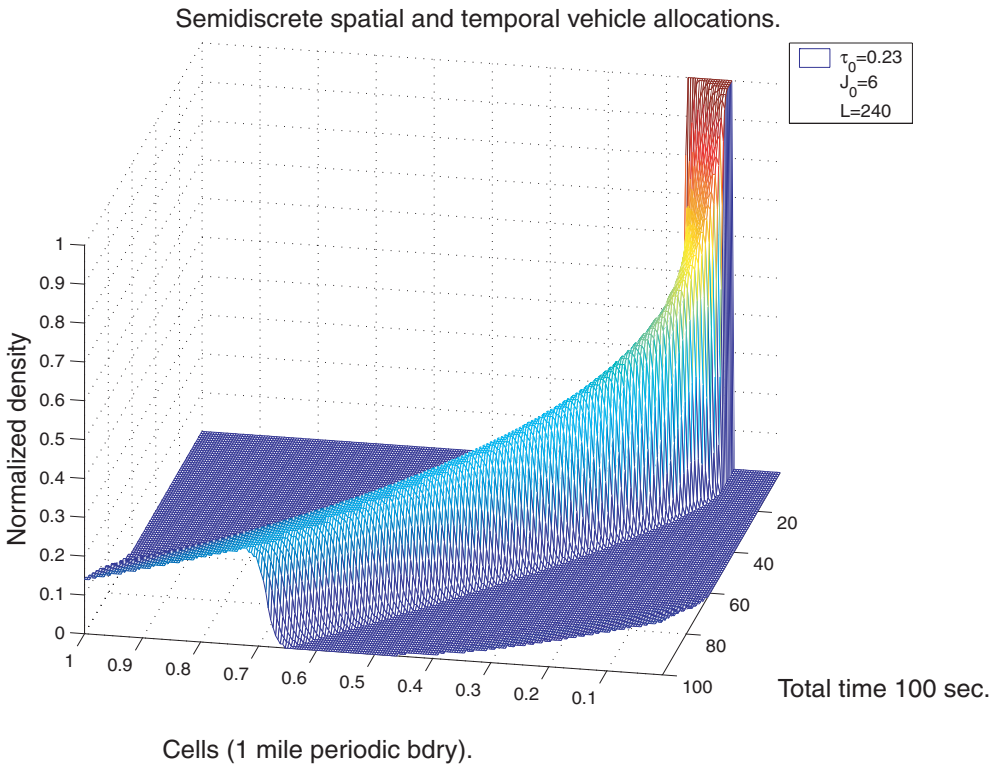


FIG. 6. *Solution of the semidiscrete scheme* (4.5) *in time and space for a traffic light type of initial condition* (4.12). *We set* $L = 240$ *based on the assumption in* (4.3). *Parameters:* $\tau_0 = .23$ *and* $J_0 = 6$.

solutions are compared in Figure 7. We clearly observe the expected rarefaction wave on one side and shock wave on the other.

In Table 4.1 we display the l_1 relative error estimates of the solutions of each model for different sizes of the interaction potential L at a specific time. The relative l_1 error is calculated from the final (in time) solutions of the semidiscrete and stochastic densities u_{sd} and u_{stoch}, respectively,

$$\frac{|u_{sd} - u_{stoch}|_{l_1}}{|u_{stoch}|_{l_1}} = \frac{\sum_z |u_{sd}(z, t_{final}) - u_{stoch}(z, t_{final})|}{\sum_z |u_{stoch}(z, t_{final})|}.$$

We observe the smallest relative errors in Table 4.1 for the case of $L = 240$. This is expected based on assumption (4.3). We compare the resulting stochastic microscopic (4.1) and FD PDE (4.11) models against each other and give possible connections with other well-known models in the following section.

4.4. Connections and comparisons between models and parameters. The task in this section is twofold. First we display possible connections between our derived models (4.1) and (4.11) with other well-known traffic flow models. Second, to further understand how parameters influence the fundamental diagram (more precisely the flux) of our derived mesoscopic formulations we also compare these formulations (4.10, 6.1) against each other to understand their possible limitations and range of predictions.

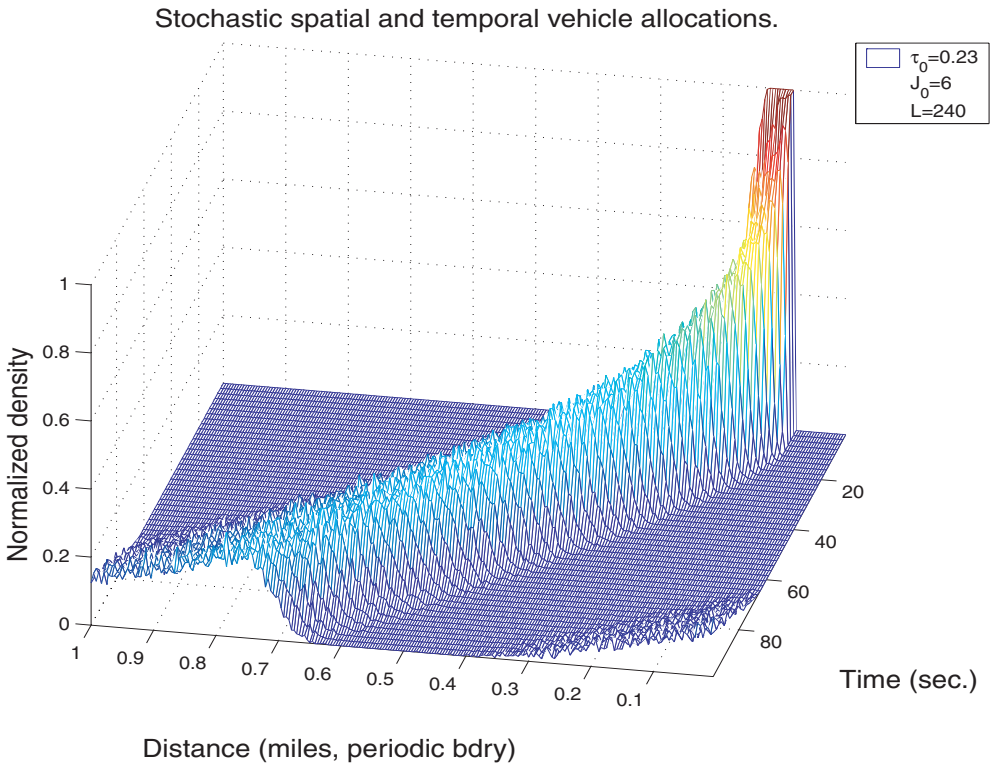


FIG. 7. *Solution of the stochastic model* (4.2) *in time and space for a traffic light type of initial condition* (4.12). *We average* 500 *realizations to produce the averaged density presented. Parameters:* $\tau_0 = .23, J_0 = 6$, *and* $L = 240$.

TABLE 4.1

Relative error of final solutions, similar to that presented in Figure 8, *comparing the semi-discrete FD scheme* (4.5) *against the stochastic model* (4.2) *solution for different sizes of the potential radius* L. *The stochastic solution has been averaged over* 500 *realizations. Other parameters:* $\tau_0 = .23$ *and* $J_0 = 6$.

Potential Radius L	240	100	50	10	4	1
l_1 Rel. Error	.0013	.0029	.0051	.0066	.0126	.02

4.4.1. Theoretical connections. We obtain here hierarchical connections with other well-known traffic flow models based on expansions of our underlying macroscopic equation (4.9) with (4.10),

$$u_t + [u(1-u)c_0 \exp(-J \circ u)]_z = 0. \tag{4.13}$$

We start by expanding the convolution term $J \circ u$,

$$J \circ u = \int_z^\infty V(y-z)u(y)\,dy \overset{x=y-z}{=} \int_0^\infty V(x)u(x+z)\,dx = J_0 u + J_1 u_z + J_2 u_{zz} + \cdots, \tag{4.14}$$

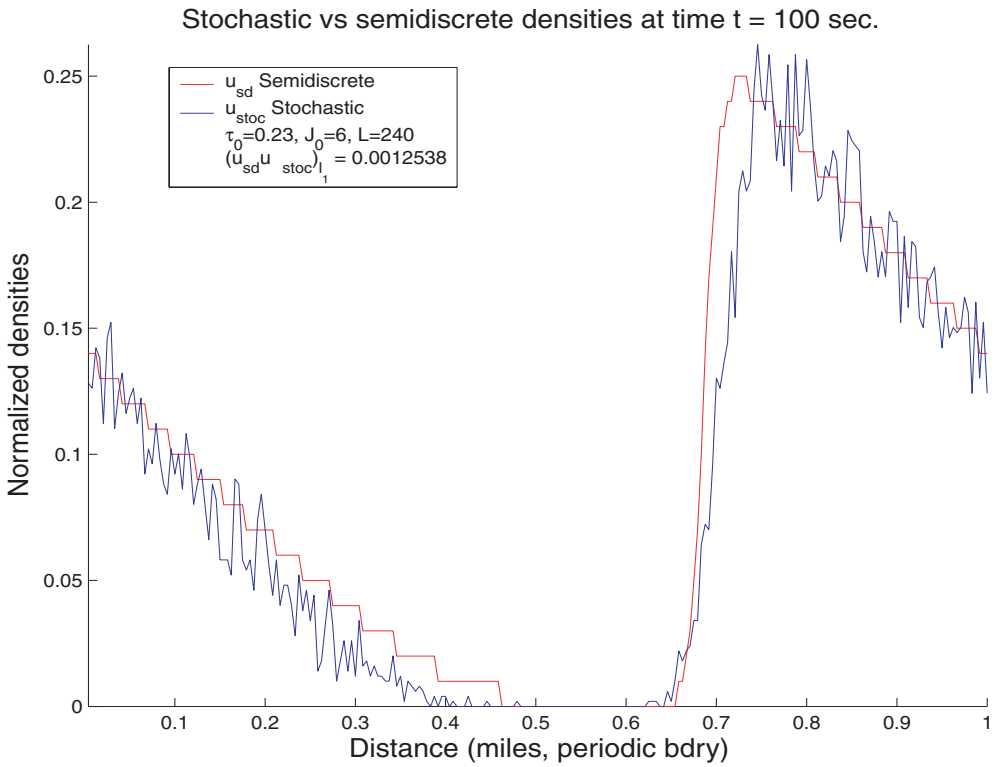


FIG. 8. *Density comparisons at final time $t = 100$ seconds. The error at that time in l_1 is also displayed. Once again, the stochastic is averaged over* 500 *realizations. Note the agreement between the solutions in both quantity and form. Parameters: $\tau_0 = .23$ and $J_0 = 6$ for $L = 240$.*

where assuming $V > 0$ (recall that the positive uniform potential we use here (3.1) implies that vehicles are attracted to the empty space in front of them) gives

(4.15)
$$J_0 = \int_0^\infty V(x)\,dx > 0, \quad J_1 = \int_0^\infty xV(x)\,dx > 0 \quad J_2 = \int_0^\infty \frac{x^2}{2}V(x)\,dx > 0.$$

We can therefore approximate the exponential as

$$e^{-J\circ u} \approx e^{-[J_0 u + J_1 u_z + J_2 u_{zz}]} = e^{-J_0 u} e^{-J_1 u_z - J_2 u_{zz}} \approx e^{-J_0 u}[1 - J_1 u_z - J_2 u_{zz}],$$

which based on (4.13) gives the higher-order traffic flow model

$$u_t + c_0[u(1-u)e^{-J_0 u}]_z = c_0[J_1 u(1-u)e^{-J_0 u} u_z]_z + c_0[J_2 u(1-u)e^{-J_0 u} u_{zz}]_z \tag{4.16}$$

with J_0, J_1, and J_2 from (4.15). Note that (4.16) is a third-order dispersive PDE with diffusion which is similar in form to the PDEs derived from optimal velocity models and usually referred to as modified KdV in [47, 48, 54].

We make general remarks below about the behavior of (4.16) as well as the more general (4.13) under different scales and/or parameters:

- Assuming first that there are no interactions $J = 0$ in the potential (4.10) of (4.13) we obtain $F(u) = c_0 u(1-u)$, which gives the well-known diffusive Lighthill–Whitham or Burgers equation flux.

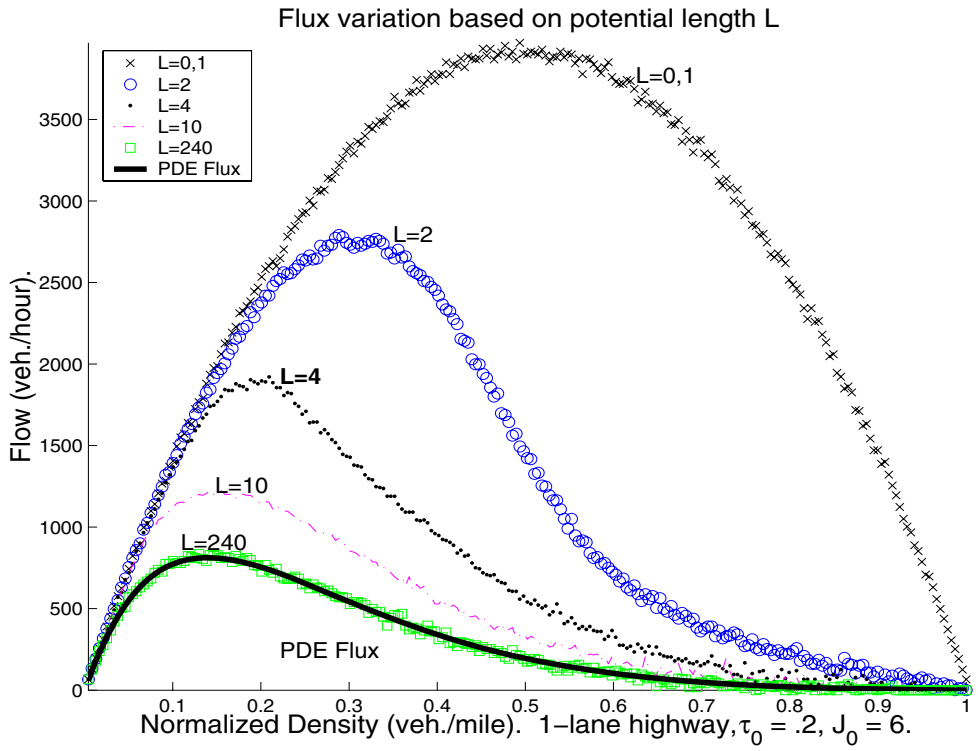


FIG. 9. *Long time averages. Comparing how the flux* (6.1) *changes with respect to increasing* L. *We set* $J_0 = 6, \tau_0 = .23$ *and run all microscopic simulations for the same total time and under the same initial conditions before plotting the flow per concentration. Note that for long-range interactions we observe that the PDE flux* (4.10) *coincides with the long-range interaction* ($L = 240$) *microscopic model flux* (4.17) *which fluctuates around it.*

- In the opposite case, however, of long-range interactions between vehicles, $L = N$, we obtain the following nonlocal flux from (4.10):

$$F(u) = c_0 u(1-u)\exp(-J_0\bar{u}). \tag{4.17}$$

  As we will see below (see Figure 9), under this long-range interaction case the flux of the stochastic model and that of the PDE (4.9, 4.10) agree.
- Note further that the hyperbolic equation obtained by including terms up to J_0 in the convolution (4.14) (disregarding J_1, etc.),

$$u_t + c_0[u(1-u)\exp(-J_0 u)]_z = 0, \tag{4.18}$$

  has a nonconvex flux. Indeed note in Figure 10 that if $J_0 \geq 3$ the flux is neither convex nor concave.
- If on the other hand we include terms up to order J_1 in (4.14), then (4.16) takes the form of a nonlinear diffusive Lighthill–Whitham type equation [68, 54, 48].
- Returning to the higher-order dispersive PDE (4.16) we note the similarities with other usual higher-order traffic flow models found in [34, 47, 48, 28, 38], although the coefficients obtained here include nonlinearities. Coherent structures can emerge as solutions of (4.16) which are similar to Figures 7 and 6 and

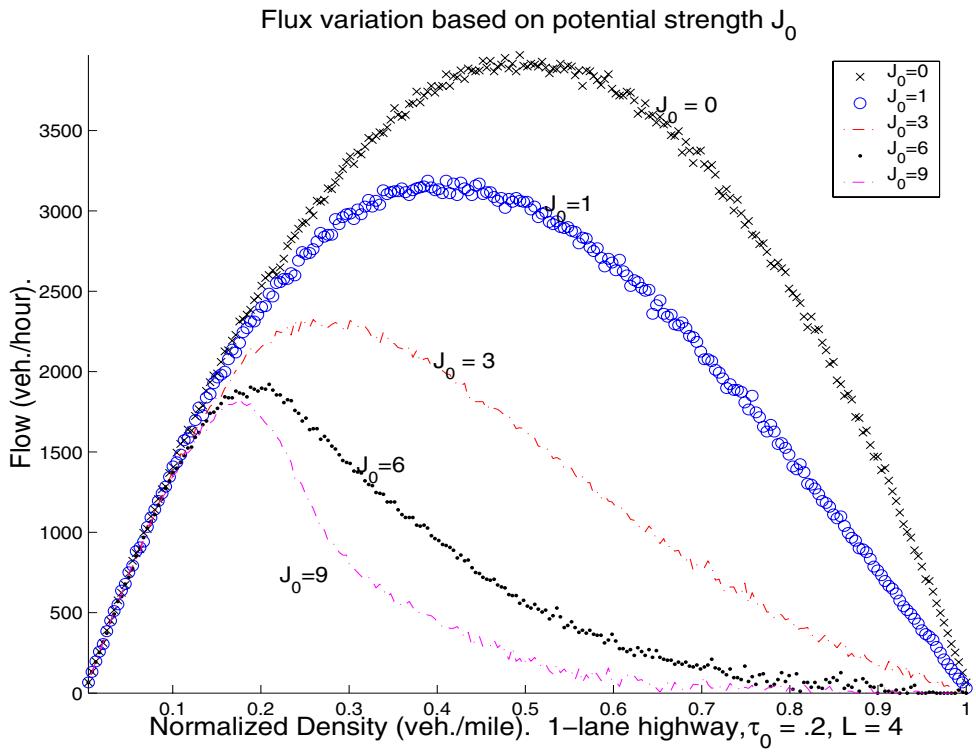


FIG. 10. *Long time averages. Comparing the influence of potential strength J_0 in the stochastic flux* (6.1). *In these comparisons we keep $L = 4$ and run all microscopic simulations for the same total time before plotting the flow per concentration.*

especially the profile density solution presented in Figure 8. It is known that traveling wave solutions of the Payne–Whitham model [37], [28] with nonconcave fundamental diagrams, which resemble the form of our higher-order PDEs, are asymptotically stable under small perturbations for a subcharacteristic type of condition [43]. It would be interesting to further examine traveling wave solutions of (4.13) as well as the higher-order approximation (4.16) and compare them with observed soliton, kink-antikink, or mixtures of other density solutions as have been noted in [34, 47, 28].

- Further, we remark that equations similar in form to (4.16) have also been studied in [26]. Diffusive and dispersive KdV type equations are emerging from Chapman-Enskog expansions of hyperbolic models with coexisting diffusion and relaxation contributions. The structure of the wave solutions which are shown to emerge for the dispersive, KdV type, equation presented there consists of enhanced solitary waves.

4.4.2. Numerical comparisons of fluxes. Clearly the microscopic model contains the maximum amount of information while the PDE is a rough averaging of the same system. It is also important, in terms of applicability of each model, to identify the range of parameters for which the PDE corresponds to our microscopic description. In that respect we expect certain parameters such as the interaction potential radius, the interaction size L, and the potential strength J_0 to be of significance. We

therefore display, in Figure 9, numerical simulations of the microscopic model predictions for different potential lengths L while keeping all other parameters the same. In that same figure we also plot the flux as obtained from our PDE formulation for a uniform potential strength value of $J_0 = 6$ and $\tau_0 = .23$ (to match the value implemented in our stochastic simulations). Note that, remarkably, the PDE flux and the microscopic flux corresponding to $L = 240$ (the maximum number of potential length) coincide (excluding the small fluctuations)! Respectively we also note that the PDE is a reasonable approximation for any microscopic model with $L > 10$ look-ahead. This is unfortunate, however, since physically we do not expect that drivers would (or even could) have a perception of traffic up front for more than possibly five vehicles. Notably the discrepancy between the microscopic and PDE fluxes is substantial for $L \approx 100$ cells or less. Such differences between long and short values for L have also been recorded in [66] for catalytic surface diffusion models.

We similarly examine how the potential strength J_0 influences the corresponding flux for either formulation (4.1) and (4.9) in Figure 10. We plot the microscopic flux for potential length of $L = 4$, which is more appropriate for actual traffic conditions. The case of $J_0 = 6$ connects Figure 9 to Figure 10 and allows for comparisons between the two. It is interesting to note how the expected concavity of traffic flow flux changes in Figure 10 as potential strength increases. In fact for values of $J_0 \geq 3$ the flux is neither concave not convex. Observe this loss of convexity also for $L \geq 2$ in Figure 9, thus producing a richer behavior than the typical Lighthill–Whitham (or Burgers) type traffic model predicts. On the other hand, in both Figures 9 and 10 we obtain one of the most basic traffic flow fluxes, the diffusive Lighthill–Whitham [68] (or Burgers) flux for the trivial case of either $J_0 = 0$ or $L = 0, 1$.

We also note the similarities of the simulated flux depicted in Figure 10 obtained here for $J_0 = 0$ and $J_0 = 6$ with Figures 2 and 3 from [50], respectively, as well as Figure 2 from [44]. Similar such agreement is also observed between Figure 10 (case of $J_0 = 6$) here and the observed data in Figure 1(b) and simulations in Figure 8 of [22].

5. Conclusions. In this work we have presented a stochastic microscopic model for traffic flow. The modeling and subsequent Monte Carlo simulations relied on simple calibration procedures of just three parameters: the potential interaction length of J_0, the relaxation time τ_0, and the look ahead L. Note that J_0 indirectly influences the desired speed of vehicles while τ_0 and L set the speed as vehicles are emerging out of a jam. Based on these settings we were able to produce realistic fundamental diagrams, for one-lane traffic, with qualitatively meaningful flows ($q_{max} \approx$2000 vehicles per hour and $c_{crit} \approx$50 vehicles per mile) which display many of the observed traffic states including phenomena such as stop-and-go traffic, spontaneous jam generation, retarded acceleration [63], and timely braking [58]. Overall the ease of calibration and richness of solution behavior make this stochastic model valuable in terms of describing, even at this simple one-lane setting, complex traffic flows.

Furthermore we derived (in the weak, long-range interactions limit) the corresponding macroscopic traffic model from our microscopic stochastic description. Subsequently we compared the solutions of each model for a variety of parameters making connections, for those regimes, with other well-known CA and PDE models for traffic flow. We obtain, up to leading order terms, hierarchical macroscopic diffusive and/or dispersive PDEs which resemble other well-known traffic flow models [68, 54, 28] of the same order. It would be interesting to examine our hierarchical macroscopic models (4.18), (4.16), and (4.13) for coherent structure solutions exhibiting behavior similar to that observed in [26, 48, 47, 28] (solitons, kink-antikink, etc.).

Last we point out that the simulations presented here can be performed in real time for the current one-lane traffic model proposed but also for multilane traffic which will be presented in a forthcoming work [11]. Extending the current one-dimensional stochastic microscopic model to multilane traffic will further allow us to directly compare our simulated results with available observational data [51, 69]. To further improve the model we also include entrances and exits by implementing two different, nonconservative (in contrast to the conservative spin exchange studied here) dynamics mechanisms: adsorption/desorption and surface diffusion.

6. Appendix: Data gathering and analysis techniques. The underlying method of recording, comparing, and analyzing traffic data for both the simulations and the observations is of paramount importance and must be clearly explained. Studies have been curried out by Athol [1] and more recently others [20, 46, 53] regarding proper data collecting procedures and underlying assumptions put forth in theory and in practice.

Observational traffic data can be obtained via a variety of methods: visual observations (video cameras), single or double loop detectors, magnetic overhead detectors, satellites, etc. In the majority, traffic data are gathered via detector loop techniques. In the case of double detector loops, for instance, pneumatic tubes (or, more recently, point detectors) are placed in close proximity on the highway and each is connected to a detector which records occurrences as vehicles pass on top of each tube. Quantities of interest that need to be determined from the collected data are usually the flow and velocity.

In general (regardless of detection method) the flow is measured as the number of vehicles $n(\tau)$ passing a detector at a given time interval τ via

$$q = \frac{1}{\tau} n(\tau) \tag{6.1}$$

(therefore flow cannot be found based on a single snapshot of vehicles over a length of interval). Usually flow is reported in number of vehicles per hour although the actual time length of recorded observation is much smaller (1/2 to 2 minutes). As a result some concerns have been raised regarding sustainability of such high volumes when data correspond to measurements over time intervals which are less than 15 minutes [21].

Average velocity on the other hand requires observations over both time and space. There are two ways to calculate velocity: time mean speed and space mean speed. As the name suggests, we calculate the time mean speed and space mean speed respectively through

$$\bar{u}_\tau = \frac{1}{N} \sum_{i=1}^{N} u_i \quad \text{and} \quad \bar{u}_s = \frac{s}{\frac{1}{N} \sum_{i=1}^{N} t_i}.$$

Here N denotes the number of time observations, s is the total distance covered in that time, and t_i is the corresponding time per observation. Note that in terms of (6.1) the following holds: $\tau = \sum_i t_i$. It is important to note here that the two definitions for mean speed provided above differ, as shown by Wardrop [67], by the ratio of the variance to the mean of the space mean speed,

$$\bar{u}_\tau = \bar{u}_s + \frac{\sigma_s^2}{\bar{u}_s}.$$

Therefore although they are similar for free-flow conditions they actually differ, especially near the highest flow regimes [20], for the key regions of stop-and-go and traffic breakdown. In practice double loop detectors are well suited to collect correctly space mean speeds. In this work (for comparative and other reasons which become clear below) we report our findings by computing space mean speed instead of time mean speed.

Density on the other hand is a quantity which is quite hard to measure empirically [46] and can be measured only along a length [20]. Having flow and space mean speeds, collected as described above, it is not uncommon to estimate the density from the well known macroscopic formula ("fundamental identity") as originally developed by Wardrop [67],

$$q = c\bar{v}. \tag{6.2}$$

A lot has been written regarding (6.2) but most importantly, as pointed out by Hall [20, p. 10], "its use has often exceeded the underlying assumptions." Clearly by applying (6.2) researchers introduce an assumption in terms of scales, ranges, and even continuity of observed variables [55] for which such an equation is valid. The validity of (6.2) is therefore sometimes questionable and its application varies among researchers, based on concentration, [30] (away from jam concentrations), [58]. The underlying problem of applying (6.2) to obtain the concentration is that (6.2) holds under "some very restrictive conditions" [20] which among other things imply that both space and time measurement intervals approach zero.

If point measurements are taken the best way [46] to estimate density is to instead calculate the lane occupancy, OC, of the traffic stream [20, 46]. Occupancy is the fraction of time that vehicles are over the detector,

$$OC = \frac{\text{total time detector is occupied by vehicle}}{\text{total study time}} = \frac{(L+D)/v}{\text{headway}} = c(L+D),$$

where L denotes the average vehicle length, D the detector length, and v the speed. (The headway is understood here as a time headway.) Therefore vehicle density c is found as

$$c = \frac{OC}{L+D}. \tag{6.3}$$

It is important to point out a common discrepancy: "almost all the theoretical work done prior to 1985 either ignores occupancy ... or else uses it ... as a surrogate for, density. On the other hand much of the freeway traffic management work ... (practical as opposed to theoretical work) relied on occupancy" [20].

In the simulations presented in our work we use virtual detectors in an effort of reproducing real traffic data collection procedures. As a result we place our virtual detector in a specific cell of our lattice and collect data on observables at that cell. Flow is computed based on (6.1) while density is found through calculation of occupancy as explained above and in the same fashion as done in [50] and similar other works. The data are averaged over small intervals of time (which we make precise in each of the presented simulations) as in real traffic data collection practices. In this work we simulate a closed round road without entrances or exits (race track) which we initialize with a specified total density (which naturally remains constant through the simulation) of randomly distributed vehicles. The fundamental diagrams are constructed by collecting data with a given density and then increasing that density and repeatedly restarting the complete simulation over again.

Acknowledgments. The first author would like to thank Chalmers Institute of Technology and Goteborg University, where part of this work was carried out during the summer of 2004. The first author would also like to thank Professor Paul Nelson for insightful discussions and comments during the preparation of this work.

REFERENCES

[1] P. Athol, *Interdependence of certain operational characteristics within a moving traffic stream*, Highway Research Record, 72 (1972), pp. 58–97.
[2] M. Bando, K. Hasebe, A. Hakayama, A. Shibata, and Y. Sugiyama, *Dynamical model of traffic congestion and numerical simulation*, Phys. Rev. E, 51 (1995), pp. 1035–1042.
[3] R. Barlovic, L. Santen, A. Schadschneider, and M. Schreckenberg, *Metastable states in cellular automata for traffic flow*, Eur. J. Phys. B, 5 (1998), pp. 793–800.
[4] A. B. Bortz, M. H. Kalos, and J. L. Lebowitz, *A new algorithm for Monte Carlo simulations of ising spin systems*, J. Comput. Phys., 17 (1975), pp. 10–18.
[5] M. J. Cassidy, *Bivariate relations in nearly stationary highway traffic*, Transportation Research B, 32 (1998), pp. 49–59.
[6] C. Chowdhury, L. Santen, and A. Schadschneider, *Statistical physics of vehicular traffic and some related systems*, Phys. Rep., 329 (2000), p. 199.
[7] E. F. Codd, *Cellular Automata*, Academic Press, New York, 1968.
[8] M. Cremer and J. Ludwig, *A fast simulation model for traffic flow on the basis of Boolean operation*, Math. Comput. Simulation, 28 (1986), p. 297.
[9] C. F. Daganzo, M. J. Cassidy, and R. L. Bertini, *Some traffic features at freeway bottlenecks*, Transportation Research B, 33 (1999), pp. 25–42.
[10] C. F. Daganzo, *Requiem for second-order fluid approximations of traffic flow*, Transportation Research B, 29 (1995), pp. 277–286.
[11] N. Dundon and A. Sopasakis, *Stochastic modeling and simulation of multi-lane traffic*, in progress.
[12] L. C. Edie, *Following and steady-state theory for non-congested traffic*, Oper. Res., 9 (1961), pp. 66–76.
[13] M. R. Evans, N. Rajewsky, and E. R. Speer, *Exact Solution of a Cellular Automaton for Traffic*, arXiv:cond-mat/9810306.
[14] H. Fukś, *Exact Results for Deterministic Cellular Automata Traffic Models*, arXiv: comp-gar/9902001.
[15] N. H. Gardner and N. H. Wilson, eds., *Freeway Speed Distribution and Acceleration Noise-Calculations from a Stochastic Continuum Theory and Comparison with Measurements*, Elsevier, New York, 1987.
[16] D. C. Gazis, R. Hermann, and R. W. Rothery, *Nonlinear follow-the-leader models of traffic flow*, Oper. Res., 9 (1961), pp. 545–567.
[17] D. L. Gerlough and M. J. Huber, *Traffic Flow Theory*, Technical Report 165, Transportation Research Board, Washington, DC, 1975.
[18] L. Gray and D. Griffeath, *The ergodic theory of traffic jams*, J. Statist. Phys., 105 (2001), pp. 413–452.
[19] B. N. Greenshields, *A study of traffic capacity*, in Proceedings of the 14th Annual Meeting of the Highway Research Board, 1934, pp. 448–474.
[20] F. L. Hall, *Traffic Flow Theory*, Chapter 2, Federal Highway Administration, Washington, DC, 1996, pp. 2–34.
[21] *Highway Capacity Manual*, Transportation Research Board, Washington, DC, 1985.
[22] D. Helbing, A. Hennecke, V. Shvetsov, and M. Treiber, *Micro- and macrosimulation of freeway traffic*, Math. Comput. Modelling, 35 (2002), pp. 517–547.
[23] D. Helbing and M. Treiber, *Gas-kinetic-based traffic model explaining observed hysteretic phase transition*, Phys. Rev. Lett., 81, 1998, pp. 3042–3045.
[24] D. Helbing, *Gas-kinetic derivation of Navier-Stokes-like traffic equations*, Phys. Rev. E, 53 (1995), pp. 2366–2381.
[25] R. Illner, A. Klar, and T. Materne, *Vlasov-Fokker-Planck models for multilane traffic flow*, Commun. Math. Sci., 1 (2003), pp. 1–12.
[26] S. Jin and J. G. Liu, *Relaxation and diffusion enhanced dispersive waves*, Proc. Roy. Soc. London A, 446 (1994), pp. 555–563.
[27] B. S. Kerner, S. L. Klenov, and D. E. Wolf, *Cellular automata approach to three-phase traffic theory*, J. Phys. A, 35 (2002), pp. 9971–10031.

[28] B. S. Kerner and P. Konhäuser, *Structure and parameters of clusters in traffic flow*, Phys. Rev. E, 50 (1994), pp. 54–83.
[29] B. S. Kerner and H. Rehborn, *Experimental properties of phase transitions in traffic flow*, Phys. Rev. Lett., 79 (1997), pp. 4030–4033.
[30] B. S. Kerner, *Dependence of Empirical Fundamental Diagram on Spatial-Temporal Traffic Patterns Features*, arXiv: cond-mat/0309018.
[31] A. Klar and R. Wegener, *Kinetic derivation of macroscopic anticipation models for vehicular traffic*, SIAM J. Appl. Math., 60 (2000), pp. 1749–1766.
[32] W. Knospe, L. Santen, A. Schadschneider, and M. Schreckenberg, *Towards a realistic microscopic description of highway traffic*, J. Phys. A., 33 (2000), pp. L477–L485.
[33] W. Knospe, L. Santen, A. Schadschneider, and M. Schreckenberg, *Single-vehicle data of highway traffic: Microscopic description of traffic phases*, Phys. Rev. E, 65 (2002), 056133.
[34] T. Komatsu and S. Sasa, *Kink soliton characterizing traffic congestion*, Phys. Rev. E, 52 (1995), pp. 5574–5582.
[35] J. Krug and P. A. Ferrari, *Phase transitions in driven diffusive systems with random rates*, J. Phys. A, 29 (1996), pp. L465–L471.
[36] J. Krug and H. Spohn, *Universality classes for deterministic surface growth*, Phys. Rev. A, 38 (1988), pp. 4271–4283.
[37] R. D. Kühne, *Macroscopic freeway model for dense traffic-stop-start wave and incident detection*, in Ninth International Symposium on Transportation and Traffic Theory, 1984, pp. 21–42.
[38] D. A. Kurtze and D. S. Hong, *Traffic jams, granular flow, and soliton selection*, Phys. Rev. E, 52 (1995), pp. 218–221.
[39] M. Duff and K. Preston, *Modern Cellular Automata: Theory and Applications*, Plenum, New York, 1984.
[40] C. Landim and C. Kipnis, *Scaling Limits of Interacting Particle Systems*, Springer, Berlin, 1999.
[41] T. M. Liggett, *Interacting Particle Systems*, Springer, Berlin, 1985.
[42] M. J. Lighthill and G. B. Whitham, *On kinematic waves* ii: *A theory of traffic flow on long crowde roads*, Proc. Roy. Soc. London Ser. A, (1955), pp. 317–345.
[43] T. Li and H. Liu, *Stability of a traffic flow model with non-convex relaxation*, Commun. Math. Sci., 3 (2005), pp. 101–118.
[44] T. Li, *Nonlinear dynamics of traffic jams*, Phys. D., 207 (2005), pp. 41–51.
[45] J. Marro and R. Dickman, *Nonequilibrium Phase Transitions in Lattice Models*, Cambridge University Press, Cambridge, UK, 1999.
[46] W. R. McShane and R. P. Roess, *Traffic Engineering*, Prentice-Hall, Englewood Cliffs, NJ, 1990.
[47] M. Muramatsu and T. Nagatani, *Soliton and kink jams in traffic flow with open bondaries*, Phys. Rev. E, 60 (1999), pp. 180–187.
[48] T. Nagatani, *The physics of traffic jams*, Rep. Prog. Phys., 65 (2002), p. 1331–1386.
[49] K. Nagel and M. Paczuski, *Emergent traffic jams*, Phys. Rev. E, 51 (1995), pp. 2909–2918.
[50] K. Nagel and M. Schreckenberg, *A cellular automaton model for freeway traffic*, J. Phys. I, 2 (1992), pp. 2221–2229.
[51] K. Nagel, D. E. Wolf, P. Wagner, and P. Simon, *Two-lane traffic rules for cellular automata: A systematic approach*, Phys. Rev. E, 58 (1998), pp. 1425–1437.
[52] P. Nelson, *Synchronized traffic flow from a modified Lighthill–Whitham model*, Phys. Rev. E, 61 (2000), p. R6052–R6055.
[53] P. Nelson, *On two-regime flow, fundamental diagrams and kinematic-wave theory*, in progress.
[54] G. F. Newell, *Nonlinear effects in theory of car following*, Oper. Res., 9 (1961), pp. 209–229.
[55] G. F. Newell, *Applications of Queueing Theory*, Chapman and Hall, London, 1982.
[56] J. C. Muñoz and C. F. Daganzo, *Structure of the transition zone behind freeway queues*, Transportation Sci., 37 (2003), p. 312–329.
[57] O. Penrose, *A mean-field equation of motion for the dynamic ising model*, J. Statist. Phys., 63 (1991), pp. 975–986.
[58] A. Schadschneider, *Traffic flow: A statistical physics point of view*, Phys. A, 312 (2002), pp. 153–187.
[59] A. Sopasakis, *Unstable flow theory and modeling*, Math. Comput. Modell., 35 (2002), pp. 623–641.
[60] A. Sopasakis, *Formal asymptotic models of vehicular traffic model closures*, SIAM J. Appl. Math., 63 (2003), pp. 1561–1584.
[61] A. Sopasakis, *Stochastic noise approach to traffic flow modeling*, Phys. A, 342 (2004), pp. 741–754.

[62] D. Stauffer, *Computer simulations of cellular automata*, J. Phys. A, 24 (1991), pp. 909–927.
[63] M. Takayasu and H. Takayasu, *Phase transition and $1/f$ noise in one dimensional asymmetric particle dynamics*, Fractals, 1 (1993) pp. 860–866.
[64] M. Treiber, A. Hennecke, and D. Helbing, *Derivation, properties and simulation of a gas-kinetic based non-local traffic model*, Phys. Rev. E, 59 (1999), pp. 239–253.
[65] J. Treiterer and J. A. Myers, *The hysteresis phenomenon in traffic flow*, in Proceedings of the 6th ISTT, Sydney, Australia, D. J. Buckley, ed., A. W. Reed, London, 1974, pp. 13–38.
[66] D. G. Vlachos and M. A. Katsoulakis, *Derivation and validation of mesoscopic theories for diffusion of interacting molecules*, Phys. Rev. Lett., 85 (2000), pp. 3898–3901.
[67] J. G. Wardrop, *Some theoretical aspects of road traffic research*, in Proceedings of the Institut. of Civil Engineers, Part II, 1952, pp. 325–378.
[68] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley, New York, 1974.
[69] R. Wiedemann, *Simulation des straßenverkehrsflusses*, Schriftenreihe des Instituts für Verkenhrswesen der Universität Karlsruhe, Germany, 1974.
[70] S. Wolfram, *Cellular Automata and Complexity*, Addison-Wesley, Reading, MA, 1994.

SIAM J. APPL. MATH.
Vol. 66, No. 3, pp. 945–976

TRAVELING WAVE SOLUTIONS TO A COUPLED SYSTEM OF SPATIALLY DISCRETE NAGUMO EQUATIONS*

MICHAEL D. BATEMAN† AND ERIK S. VAN VLECK†

Abstract. We consider a coupled system of discrete Nagumo equations and derive traveling wave solutions to this system using McKean's caricature of the cubic. A certain form of this system is used to model ephaptic coupling between pairs of nerve axons. We study the difference $g(c) = a_1 - a_2$ between the detuning parameters a_i that is required to make both waves move at the same speed c. Of particular interest is the effect of a coupling parameter α and an "alignment" parameter A on the function g. Numerical investigation indicates that for fixed A, there exists a time delay value β that results in $g = 0$, and for large enough wave speeds, multiple such β values exist. Also, numerical results indicate that the perturbation of α away from zero will yield additional solutions with positive wave speed when $A = \frac{1}{2}$. We employ both analytical and numerical results to demonstrate our claims.

Key words. discrete Nagumo equations, ephaptic coupling, traveling waves

AMS subject classifications. 35K57, 74N99

DOI. 10.1137/050624352

1. Introduction.

In myelinated nerve axons, transmembrane ion flow occurs only at the spatially periodic nodes of Ranvier, and the activity at these nodes may affect the activity at nodes of neighboring fibers. This so-called ephaptic coupling is an electrical effect that causes neighboring fibers to interact and possibly synchronize with each other. Accompanying the problem of ephaptic coupling is the issue of the relative positioning of the nodes of Ranvier on the different fibers. That is, given two parallel nerve fibers, the nodes on one fiber may or may not align perfectly with the nodes on the other fiber.

Our contribution in this paper is to derive a solution to a system that models these phenomena and use this model to show that coupling decreases the size of the range of propagation failure when the nodes of Ranvier are staggered, but that coupling increases the size of this range when the nodes are perfectly aligned. To do this, we consider a system of two myelinated nerve axons coupled ephaptically. In particular, our goal is to study the effect of this coupling and the effect of nonalignment on the propagation of action potentials. Different types of coupling between fibers are possible. In [2], Binczak, Eilbeck, and Scott model "saltatory" conduction present in these myelinated neurons with equations used to govern the behavior of electrical circuits and introduce the effect of ephaptic coupling between two myelinated neurons. In [1] a different type of coupling, called "ohmetric" coupling, is considered and it is shown that the introduction of this kind of coupling causes waves on two adjacent myelinated axons to match speeds with each other. Earlier work of Keener [13] and Bose and Jones [4, 5] addressed the issue of ohmetric coupling.

The strength of the ephaptic coupling α depends on the electrical resistance R_{int} inside the axons (assumed to be the same for both axons) and the resistance R_o of the medium between the axons. We must also consider the positioning of the axons

*Received by the editors February 14, 2005; accepted for publication (in revised form) October 3, 2005; published electronically March 3, 2006. This work was supported in part by NSF grants DMS-0139824 and DMS-0513438.

http://www.siam.org/journals/siap/66-3/62435.html

†Department of Mathematics, University of Kansas, Lawrence, KS 66045 (mbat@ku.edu, evanvleck@math.ku.edu).

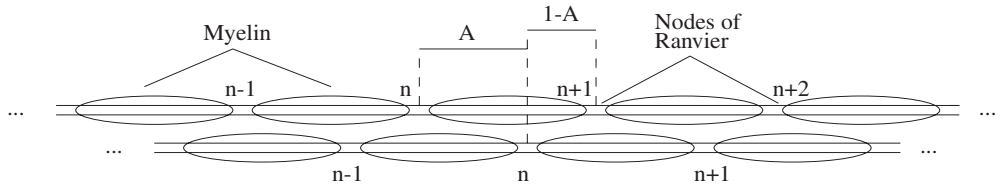


FIG. 1. *Diagram of two myelinated nerve axons.*

relative to each other: here we assume that they are parallel, that the nodes on a given fiber are evenly spaced, and that the distance between nodes is the same on both fibers, but we allow the nodes of Ranvier to line up or not line up, as shown in Figure 1. That is, we will introduce an alignment parameter A that will reflect the positioning of the nodes on one fiber relative to those on the other. We will also introduce a parameter β that will act as a time delay on one of the fibers. Specifically, we consider a system of differential equations on two one-dimensional lattices coupled together. Study of the uncoupled case [6, 11, 10, 14] provides an idea of the kind of behavior to expect from the coupled system and also serves as a precedent to which our results may be compared.

We consider the system

$$\begin{aligned}
\dot{V}_n^{(1)} + f_1(V_n^{(1)}) &= \frac{1}{1-\alpha^2}[D_1(\bar{L}V^{(1)})_n - \alpha D_2(\bar{L}V^{(2)})_n - \alpha A(\bar{N}V^{(2)})_n \\
&\qquad - \alpha^2 A(\bar{N}V^{(1)})_{n+1}], \\
\dot{V}_n^{(2)} + f_2(V_n^{(2)}) &= \frac{1}{1-\alpha^2}[-\alpha D_1(\bar{L}V^{(1)})_n + D_2(\bar{L}V^{(2)})_n + \alpha^2 A(\bar{N}V^{(2)})_n \\
&\qquad + \alpha A(\bar{N}V^{(1)})_{n+1}],
\end{aligned} \tag{1.1}$$

where $(\bar{L}V)_n = V_{n+1} - 2V_n + V_{n-1}$, $(\bar{N}V^{(i)})_n = (\dot{V^{(i)}}_{n-1} - \dot{V^{(i)}}_n) + (f_i(V_{n-1}^{(i)}) - f_i(V_n^{(i)}))$, $D_1, D_2 > 0$ are diffusion coefficients, $A \in [0,1)$ is the alignment parameter, and $\alpha \in [0,1)$ is the coupling coefficient. When $\alpha = 0$, the ephaptic coupling is completely turned off, and increasing α increases the strength of the coupling. When $A = 0$, the nodes are in perfect alignment; setting $A = \frac{1}{2}$, for example, staggers the nodes so that the nth node on one fiber is equidistant from the nth and $(n+1)$th nodes on the other fiber. The nonlinearities $f_i : \mathbb{R} \to \mathbb{R}$ are often cubics of the form $f_i(u) = u(u - a_i)(u - 1)$, where the $a_i \in (0,1)$ are "detuning" parameters, and the quantities $V_n^{(i)}$ represent the ionic flow through the membrane at the nth node of the respective fibers. However, to facilitate the construction of a solution, we consider an idealized, piecewise linear f_i known as McKean's caricature of the cubic (see [15, 17, 12, 22, 23, 6, 11, 8, 9]). When considering one fiber, it is only necessary to think of the detuning parameter $a \in (0, \frac{1}{2})$ because of symmetry in the relationship between the detuning parameter and the wave speed. A priori, we do not know if the same is true when considering two fibers. However, we will see in Theorem 4.4 and Corollary 4.5 that certain symmetries do hold in the coupled problem.

In general, the wave speed c depends on all the other parameters in the problem. Here, however, we approach the issue from a different angle: we fix all parameters except the detuning parameters a_i and assume that the wave speed c is the same for

both fibers. This way we "solve," in a sense, for the detuning parameters. Instead of fixing the detuning parameters, and studying how they affect the wave speeds, we demand that the wave speeds be the same and determine which pairs of detuning parameters give us this effect.

Although our particular model is motivated by this neurological application, similar models may also be used to study action potentials in cardiac cells, among other things (see [3, 16, 20, 21]). The authors in [1] also mention the possible application of coupled two-dimensional lattices to the study of image processing. For this reason, we consider a more general system and allow the parameters to range over values that may not be physically reasonable for our particular application.

This paper is organized as follows: we start by using a piecewise linear nonlinearity f to derive candidate solutions using a Fourier transform method. To make this process easier, we make a key assumption about the shape of our solutions—this assumption will be verified after we have obtained the candidate solutions. The details of the construction have been left to an appendix. After proving the existence of traveling backs, we look at some properties held by these solutions. We investigate the relationship between the wave speed c and the detuning parameters a_i, as well as the effects of the coupling coefficient α and the alignment parameter A on this system. We have interspersed some numerical computations throughout the paper to illustrate the analytical results and provide insight into the problem.

1.1. Derivation of the system. The system considered here is a version of the model used in [18] (see pp. 177–183) and [2]. The variables used here are as follows: I_n is the mesh current, a term used to analyze circuits using Kirchhoff's equations; $I_{ion,n}$ is a cubic modeling the ionic current (taking into consideration both sodium and potassium); $V_n^{(i)}$ is the voltage across a node of Ranvier, and V_b is a constant representing the Nernst potential for sodium ions; $R_{i,j}$ is the resistance inside fiber j, R_o is the resistance outside the axons, and R_f is a constant representing the internodal resistance, and is on the same order as $R_{i,j}$; and C is the capacitance of a node plus the capacitance of the adjacent internodal myelin sheath.

Using the circuit diagrams presented in [2] and [18], we sum the voltages around the mesh and set the result to zero. Note that here we will use A to represent the amount that fiber 1 leads fiber 2. This gives us

$$\begin{aligned} v_n^{(1)} - v_{n+1}^{(1)} &= (R_{i,1} + R_o)I_n^{(1)} + R_o(AI_{n-1}^{(2)} + (1-A)I_n^{(2)}), \\ v_n^{(2)} - v_{n+1}^{(2)} &= (R_{i,2} + R_o)I_n^{(2)} + R_o(AI_{n+1}^{(1)} + (1-A)I_n^{(1)}). \end{aligned} \tag{1.2}$$

In addition, we have the following relationship between the mesh currents and the voltages:

$$I_{n-1}^{(j)} - I_n^{(j)} = C\frac{dv_n^{(j)}}{dt} + I_{ion,n}^{(j)}, \quad j = 1, 2. \tag{1.3}$$

Now we multiply (1.2) by $\frac{R_f}{V_b(R_{i,j}+R_o)}$ and (1.3) by $\frac{R_f}{V_b}$. Using the notation

$$V_n^{(j)} = \frac{v_n^{(j)}}{V_b}, \quad i_n^{(j)} = \frac{R_f I_n^{(j)}}{V_b}, \quad D_j = \frac{R_f}{R_{i,j}+R_o}, \text{ and } \alpha_j = \frac{R_o}{R_{i,j}+R_o}, \tag{1.4}$$

we arrive at

$$\begin{aligned} D_1(V_n^{(1)} - V_{n+1}^{(1)}) &= i_n^{(1)} + \alpha(Ai_{n-1}^{(2)} + (1-A)i_n^{(2)}), \\ D_2(V_n^{(2)} - V_{n+1}^{(2)}) &= i_n^{(2)} + \alpha(Ai_{n+1}^{(1)} + (1-A)i_n^{(1)}), \end{aligned} \tag{1.5}$$

$$
\begin{aligned}
i_{n-1}^{(1)} - i_n^{(1)} &= R_f C \dot{V}_n^{(1)} + f_1(V_n^{(1)}),\\
i_{n-1}^{(2)} - i_n^{(2)} &= R_f C \dot{V}_n^{(2)} + f_2(V_n^{(2)}).
\end{aligned} \tag{1.6}
$$

As noted in [2], experimental results indicate that $R_o \ll R_{i,j} \approx R_f$, so we take $\alpha_1 = \alpha_2$ but allow $D_1 \neq D_2$.

Using (1.5) and (1.6), we have

$$
\begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix} \begin{bmatrix} i_n^{(1)} \\ i_n^{(2)} \end{bmatrix} = \begin{bmatrix} X_n^{(1)} \\ X_n^{(2)} \end{bmatrix}, \tag{1.7}
$$

where

$$
\begin{aligned}
X_n^{(1)} &= D_1(V_n^{(1)} - V_{n+1}^{(1)}) - \alpha A(R_f C \dot{V}_n^{(2)} + f(V_n^{(2)})),\\
X_n^{(2)} &= D_2(V_n^{(2)} - V_{n+1}^{(2)}) + \alpha A(R_f C \dot{V}_{n+1}^{(1)} + f(V_{n+1}^{(1)})),
\end{aligned} \tag{1.8}
$$

which gives us

$$
i_n^{(1)} = \frac{1}{1-\alpha^2}(X_n^{(1)} - \alpha X_n^{(2)}), \quad i_n^{(2)} = \frac{1}{1-\alpha^2}(X_n^{(2)} - \alpha X_n^{(1)}). \tag{1.9}
$$

Substituting into (1.6), we have

$$
\begin{aligned}
R_f C \dot{V}_n^{(1)} + f_1(V_n^{(1)}) &= \frac{1}{1-\alpha^2}[(X_{n-1}^{(1)} - X_n^{(1)}) - \alpha(X_{n-1}^{(2)} - X_n^{(2)})],\\
R_f C \dot{V}_n^{(2)} + f_2(V_n^{(2)}) &= \frac{1}{1-\alpha^2}[(X_{n-1}^{(2)} - X_n^{(2)}) - \alpha(X_{n-1}^{(1)} - X_n^{(1)})].
\end{aligned} \tag{1.10}
$$

Computing $X_{n-1}^{(j)} - X_n^{(j)}$ for $j = 1, 2$ and inserting the resulting expressions into (1.10) brings us to system (1.1). Note that we will take $R_f C = 1$ since changing these parameters will only amount to a rescaling of the wave speed c after we impose a traveling wave ansatz.

2. Construction of a solution. In the construction of our solution, the speed c is assumed to be nonzero unless otherwise noted, and the nonlinearity f_i will be the idealized cubic-like function

$$
f_i(u) = u - h(u - a_i), \tag{2.1}
$$

where h denotes the Heaviside step function

$$
h(u) = \begin{cases} 0 & \text{if } u < 0,\\ [0,1] & \text{if } u = 0,\\ 1 & \text{if } u > 0. \end{cases} \tag{2.2}
$$

Note that $h(u)$ is a set-valued function, evaluating to the interval $[0, 1]$ when $u = 0$ and evaluating to a singleton everywhere else. This results in the $f_i(u)$ being set-valued functions as well. Thus, (1.1) with f_i given in (2.1) and (2.2) should be interpreted as a differential inclusion when $V_n^{(i)} = a_i$.

After imposing the traveling wave ansatz $V_n^{(i)}(t) = \varphi_i(n - ct)$, and letting $\xi = n - ct$, we have

$$
\begin{aligned}
(2.3)\quad -c\varphi_1'(\xi) + f_1(\varphi_1(\xi)) &= \frac{1}{1-\alpha^2}\left[D_1(L\varphi_1)(\xi) - \alpha D_2(L\varphi_2)(\xi)\right. \\
&\qquad \left. - \alpha A(N\varphi_2)(\xi) - \alpha^2 A(N\varphi_1)(\xi+1)\right], \\
-c\varphi_2'(\xi) + f_2(\varphi_2(\xi)) &= \frac{1}{1-\alpha^2}\left[-\alpha D_1(L\varphi_1)(\xi) + D_2(L\varphi_2)(\xi)\right. \\
&\qquad \left. + \alpha^2 A(N\varphi_2)(\xi) + \alpha A(N\varphi_1)(\xi+1)\right],
\end{aligned}
$$

where

$$
\begin{aligned}
(L\varphi)(\xi) &= \varphi(\xi+1) - 2\varphi(\xi) + \varphi(\xi-1), \\
(N\varphi)(\xi) &= -c[\varphi'(\xi-1) - \varphi'(\xi)] + [f(\varphi(\xi-1)) - f(\varphi(\xi))]. \qquad (2.4)
\end{aligned}
$$

It is natural to require the boundary conditions

$$
\varphi_i(-\infty) = 0, \qquad \varphi_i(+\infty) = 1 \qquad (2.5)
$$

for $i = 1, 2$. To help construct solutions to this system, we will initially assume that each φ_i satisfies $\varphi_i(\beta_i) = a_i$ for only one value, and we may assume one of these values to be zero by simply translating the argument. That is, we assume

$$
\varphi_i(\xi) \begin{cases} < a_i & \text{for} \quad \xi < \beta_i, \\ = a_i & \text{for} \quad \xi = \beta_i, \\ > a_i & \text{for} \quad \xi > \beta_i, \end{cases} \qquad (2.6)
$$

where we will take $\beta_1 = 0$ and $\beta_2 = \beta$. After we construct our candidate solutions, we will verify that they satisfy these assumptions.

With (2.6), we have that $h(\varphi_i(\xi) - a_i) = h(\xi - \beta_i)$, which gives us $f(\varphi_i(\xi)) = \varphi_i(\xi) - h(\xi - \beta_i)$. Now the system (2.3) becomes

$$
\begin{aligned}
(2.7)\quad -c\varphi_1'(\xi) &= \frac{1}{1-\alpha^2}\left[D_1(L\varphi_1)(\xi) - \alpha D_2(L\varphi_2)(\xi)\right. \\
&\qquad \left. - \alpha A(N\varphi_2)(\xi) - \alpha^2 A(N\varphi_1)(\xi+1)\right] - \varphi_1(\xi) + h(\xi), \\
-c\varphi_2'(\xi) &= \frac{1}{1-\alpha^2}\left[-\alpha D_1(L\varphi_1)(\xi) + D_2(L\varphi_2)(\xi)\right. \\
&\qquad \left. + \alpha^2 A(N\varphi_2)(\xi) + \alpha A(N\varphi_1)(\xi+1)\right] - \varphi_2(\xi) + h(\xi-\beta).
\end{aligned}
$$

Notice the presence of several Heaviside functions in each equation of (2.7). Each instance of the Heaviside function results in a discontinuity of the first derivative of the functions φ_i. For example, we expect to have discontinuities in the first derivative of φ_1 at $\xi = 0$, $\xi = 1$, $\xi = \beta$, and $\xi = \beta + 1$. Further, an appearance of $\varphi'(\xi + 1)$ in the first equation leads us to expect a discontinuity in the first derivative of φ_1 at $\xi = -1$. These discontinuities will be reflected in the solution we compute.

As outlined in Appendix A, candidate solutions to this system can be expressed as

$$
\varphi_i(\xi) = \frac{1}{2} + \frac{1}{2\pi}\int_0^\infty \sum_{j \in M_i} F_{i,j}(s)\cos(s(\xi - j)) + G_{i,j}(s)\frac{1}{s}\sin(s(\xi - j))ds, \qquad (2.8)
$$

where $M_i = \{\beta_i, \beta_i + 1, \beta_i - 1, \beta_{i^*}, \beta_{i^*} + 1\}$, $1^* = 2$ and $2^* = 1$, and

(2.9)
$$\begin{aligned}
F_{i,\beta_i}(s) &= [W(s)(1+2Q) - R_{i^*}(s)Y(s)], \quad & G_{i,\beta_i}(s) &= [X(s)(1+2Q) - R_{i^*}(s)Z(s)],\\
F_{i,\beta_i+1}(s) &= -QW(s), & G_{i,\beta_i+1}(s) &= -QX(s),\\
F_{i,\beta_i-1}(s) &= -QW(s), & G_{i,\beta_i-1}(s) &= -QX(s),\\
F_{i,\beta_{i^*}}(s) &= -\alpha(1-A)R_{i^*}(s)Y(s), & G_{i,\beta_{i^*}}(s) &= -\alpha(1-A)R_{i^*}(s)Z(s),\\
F_{i,\beta_{i^*}+1}(s) &= -\alpha A R_{i^*}(s)Y(s), & G_{i,\beta_{i^*}+1}(s) &= -\alpha A R_{i^*}(s)Z(s),
\end{aligned}$$

where

$$\begin{aligned}
W(s) &= \frac{B(s)}{\det M(s)} - \frac{B(-s)}{\det M(-s)} = \frac{2c[b_2c^2s^2 + b_2 - b_0]}{|\det M(s)|^2},\\
X(s) &= \frac{B(s)}{\det M(s)} + \frac{B(-s)}{\det M(-s)} = \frac{2[(b_2+b_1)c^2s^2 + b_2 + b_1 + b_0]}{|\det M(s)|^2},\\
Y(s) &= \frac{1}{\det M(s)} - \frac{1}{\det M(-s)} = \frac{2c[2b_2 + b_1]}{|\det M(s)|^2},\\
Z(s) &= \frac{1}{\det M(s)} + \frac{1}{\det M(-s)} = \frac{2[-b_2c^2s^2 + b_2 + b_1 + b_0]}{|\det M(s)|^2},
\end{aligned} \tag{2.10}$$

and

(2.11)
$$\begin{aligned}
Q &= k\alpha^2 A(1-A), \quad & \det M(s) &= b_2(s)B(s)^2 + b_1(s)B(s) + b_0(s) = a(s) + ib(s),\\
R_i(s) &= 2kD_iC(s), & a(s) &= (1 - c^2s^2)b_2(s) + b_1(s) + b_0(s),\\
k &= \frac{1}{1-\alpha^2}, & b(s) &= -cs(2b_2(s) + b_1(s)),\\
B(s) &= 1 - ics, & b_0(s) &= 4kD_1D_2C^2(s),\\
C(s) &= \cos(s) - 1, & b_1(s) &= -2kC(s)(D_1 + D_2),\\
E(s) &= 1 - e^{is}, & b_2(s) &= 1 - 2k\alpha^2 A(1-A)C(s).
\end{aligned}$$

Notice that the set M_i consists of all values at which the first derivative of φ_i is discontinuous. This agrees with our prediction based on inspection of the equations (2.7).

For convenience, we will suppress the s-dependence of the functions a, b, b_0, b_1, and b_2. Our next proposition establishes some basic properties of the φ_i, but to do this, we need to bound the integrands in the solutions (2.8). This is accomplished by the following lemma.

LEMMA 2.1. *Let $\alpha \in [0,1)$, $A \in [0,1)$, $\beta \in \mathbb{R}$, and $c \neq 0$. Then*

$$|\det M(s)|^2 \geq 1 \tag{2.12}$$

for $s \in \mathbb{R}$.

Proof. We leave the proof of this lemma to the second appendix. ∎

PROPOSITION 2.2. *Use the definition of φ_i as given in* (2.8)*, and let $A \in [0,1), \alpha \in [0,1)$, $\beta \in \mathbb{R}$, $\xi \in \mathbb{R}$, and $c \neq 0$. Then φ_i is continuous in A, α, β, ξ, and c.*

Proof. The previous lemma gives us a lower bound on the denominator of the integrands in (2.8), except when $s = 0$. However, the s appears in the denominator

only when there is a sine in the numerator. The numerators of the integrands are also bounded, and from the definitions (2.8), we see that the integrands are continuous in the variables listed above. Also note from the definitions in (2.10) and (2.11) that the integrands are $O(s^{-2})$ as $s \to \infty$, making the integral in (2.8) absolutely convergent. The continuity claims follow from this fact. ∎

3. Existence.

3.1. Proof of existence. Recall that the candidate solutions were derived using the assumptions given in (2.6).

THEOREM 3.1. *Let $A \in [0,1)$, $\beta \in \mathbb{R}$, and $c \neq 0$, and consider our solutions as functions of α and ξ. That is, let $\varphi_i = \varphi_i(\alpha, \xi)$. Also recall $\beta_1 = 0$ and $\beta_2 = \beta$. Suppose the following conditions are met:*

C1. *The φ_i are continuous in α for all $\xi \in \mathbb{R}$ and for $\alpha \in [0,1)$.*

C2. *There exists $\delta = \delta(A, \beta, D_1, D_2, c) > 0$ such that $\varphi_i'(0,\xi) > 0$ for $\xi \neq \beta_i$ in the interval $(\beta_i - \delta, \beta_i + \delta)$ (where $'$ indicates differentiation with respect to ξ).*

Then there exists a range of α for which $\varphi_i(\alpha,\xi) > a_i$ when $\xi > \beta_i$ and $\varphi_i(\alpha,\xi) < a_i$ when $\xi < \beta_i$,

Proof. Consider the functions

$$h_i(\alpha,\xi) = \varphi_i(\alpha,\xi) - \varphi_i(\alpha,\beta_i). \tag{3.1}$$

From the work done in [6], we know that $h_i(0,\xi) > 0$ for $\xi > \beta_i$ and $h_i(0,\xi) < 0$ for $\xi < \beta_i$ for $i = 1,2$. We also know that the h_i are continuous in α since, by Proposition 2.2, the φ_i are continuous in α. These two facts, together with the boundary conditions, guarantee that there is an α^* such that for $\alpha \in [0,\alpha^*)$, $h_i(\alpha,\xi) > 0$ for $\xi \geq \beta_i + \delta$ and $h_i(\alpha,\xi) < 0$ for $\xi \leq \beta_i - \delta$. So let $\alpha \in [0,\alpha^*)$. To ensure that $h_i(\alpha,\xi) > 0$ for $\xi \in (\beta_i, \beta_i + \delta)$ and $h_i(\alpha,\xi) < 0$ for $\xi \in (\beta_i - \delta, \beta_i)$, we must also require that $\varphi_i'(0,\xi) > 0$ for $\xi \in (\beta_i - \delta, \beta_i + \delta)\backslash\{\beta_i\}$ (where, once again, $'$ denotes differentiation with respect to ξ). So, condition C2 gives us that $\varphi_i(0,\xi) > 0$ for $\xi \in (\beta_i - \delta, \beta_i + \delta)\backslash\{\beta_i\}$, which satisfies the requirement given above. ∎

It remains to verify the hypotheses of Theorem 3.1. The strict monotonicity result from the $\alpha = 0$ case done in [6] is enough to satisfy condition C2 (this is true because when $\alpha = 0$, the solution is the same, up to a translation, for all β) and Proposition 2.2 satisfies condition C1. We now verify that the φ_i satisfy our boundary conditions.

PROPOSITION 3.2. *For $i = 1,2$,*

$$\lim_{\xi\to-\infty} \varphi_i(\xi) = 0 \qquad \text{and} \qquad \lim_{\xi\to+\infty} \varphi_i(\xi) = 1. \tag{3.2}$$

Proof. Recall the definition of φ_i given in (2.8), and assume $\xi > 0$. Now use the change of variables $s \to \frac{s}{\xi}$ and take the limit as $\xi \to +\infty$, using the Lebesgue dominated convergence theorem. We have

$$\begin{aligned}
\lim_{\xi\to+\infty} \varphi_i(\xi) &= \frac{1}{2} + \frac{1}{2\pi} \lim_{\xi\to+\infty} \int_0^\infty \sum_{j\in M_i} \left[\frac{1}{\xi} F_{i,j}\left(\frac{s}{\xi}\right) \cos\left(s - \frac{sj}{\xi}\right) \right. \\
&\qquad \left. + G_{i,j}\left(\frac{s}{\xi}\right) \frac{1}{s} \sin\left(s - \frac{sj}{\xi}\right) \right] ds \\
&= \frac{1}{2} + \frac{1}{2\pi} \int_0^\infty \sum_{j\in M_i} \left[0 \cdot F_{i,j}(0) \cos(s) + G_{i,j}(0) \frac{1}{s} \sin(s) \right] ds \\
&= \frac{1}{2} + \frac{1}{2\pi} \int_0^\infty \frac{\sin(s)}{s} X(0) ds = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \frac{\sin(s)}{s} ds = 1.
\end{aligned}$$

Proof for the limit as $\xi \to -\infty$ is the same, except we assume $\xi < 0$. But note that when $\xi < 0$, the change of variables $s \to \frac{s}{\xi}$ results in the limits of integration becoming zero to $-\infty$, and everything else remains unchanged. The last expression then becomes

$$\frac{1}{2} + \frac{1}{\pi}\int_0^{-\infty} \frac{\sin(s)}{s} ds = \frac{1}{2} - \frac{1}{\pi}\int_{-\infty}^{0} \frac{\sin(s)}{s} = 0, \tag{3.3}$$

giving us the result. □

To summarize, we know there is an $\alpha^* > 0$ such that solutions exist for all $\alpha \in [0, \alpha^*)$. This next proposition gives a condition that implies our candidate solutions are not solutions of (2.7).

PROPOSITION 3.3. *Fix $A \in [0,1), \alpha \in [0,1)$, $\beta \in \mathbb{R}$, and $c \neq 0$. If either $\varphi_i(\beta_i) > 1$ or $\varphi_i(\beta_i) < 0$, then the candidate solutions given in* (2.8) *are not solutions of* (2.7).

Proof. By Proposition 3.2, $\varphi_i(\xi) \to 1$ as $\xi \to \infty$. Hence if $\varphi_i(\beta_i) > 1$, there will be a $\xi > \beta_i$ for which $\varphi_i(\xi) < \varphi_i(\beta_i)$, violating our original assumptions (2.6). Recall also that $\varphi_i(\xi) \to 0$ as $\xi \to -\infty$, so if $\varphi_i(\beta_i) < 0$, we have violated the assumption that $\varphi_i(\xi) < \varphi_i(\beta_i)$ for all $\xi < \beta_i$. □

Having proved the existence of our solutions, we turn our attention to uncovering some basic properties of these solutions and to plots of several solution curves.

3.2. Plots of waveforms. We wish to show the form of the traveling wave solutions, but first, a few comments about our numerical studies are in order. Because we have an explicit formula for our candidate solutions, we are able to compute a broad range of numerical results. That said, we must also note that the large number of parameters in this problem makes it unfeasible to completely canvass the parameter space. Instead, we have focused on parameter values we expect will be representative of a larger range of values or on those that illustrate interesting phenomena. We approximate the integrals using the adaptive Gaussian quadrature code `adapt` of [19] and to find zeros we use the combined secant/bisection code `zero` of [19].

In Theorem 3.1, we proved the existence of solutions for some range of α. The size of this range is unknown, and it is important to keep in mind that the existence of our candidate solutions may not be guaranteed for all combinations of parameter values explored in this section. We have, however, checked the necessary condition that $a_i \in (0,1)$, as mentioned in Proposition 3.3 for most of the parameter values in this section. In some instances, we have taken the additional step of numerically verifying (an admittedly finite range of) the candidate solutions themselves to verify that they satisfy our original assumptions. In particular, we have numerically determined waveforms with $D_1 = D_2 = 1$, $A = \beta = 0$, and c values of $10^{-1}, 10^0$, and 10^1, for α between 0 and .95, with a step size of .05.

Our intent in providing numerical results is twofold: to illustrate some of the results from the theoretical sections and to use numerical evidence to extend our knowledge of the problem and to possibly provide ideas for further study. (In particular, we wish to answer, at least tentatively, the questions of the effects of ephaptic coupling and nonalignment on the propagation of action potentials.) Figure 2 shows plots of wave forms for a small wave speed. Notice the large jumps; these correspond to discontinuities in the first derivative of the functions φ_i. A larger wave speed results in waves with much less pronounced jumps. An example of wave forms with larger c is given later in Figure 6, but we withhold these plots for now since we will use them to illustrate a different phenomenon.

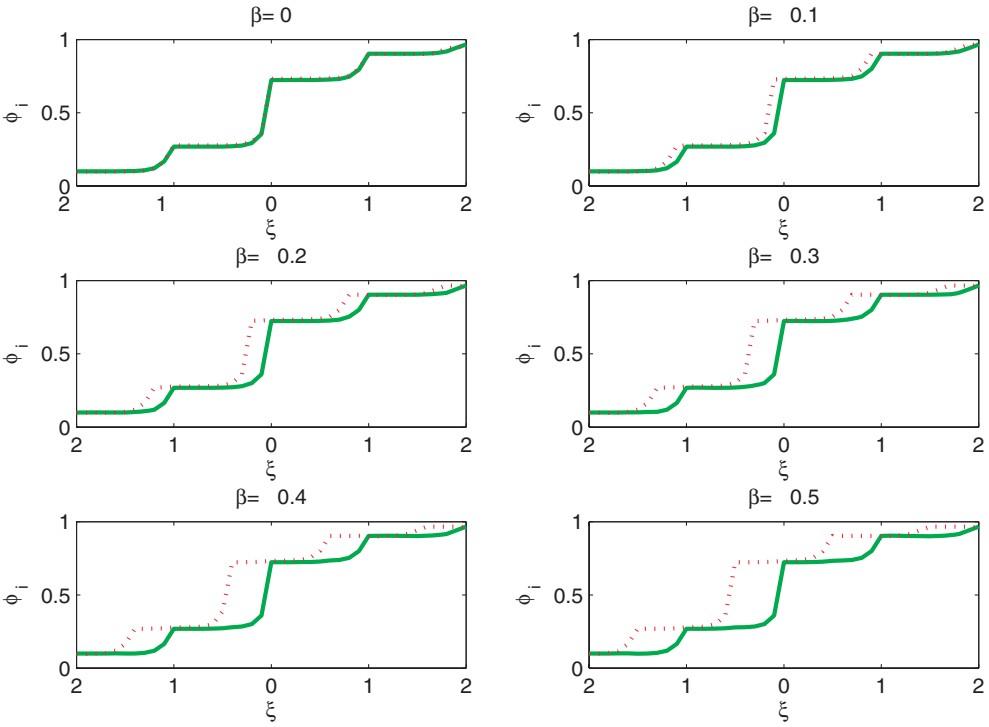


FIG. 2. *Wave profiles for* $c = .1$, $A = .5$, $\alpha = .1$, *and* $D_1 = D_2 = 1$. *The steep jumps are due to the small wave speed.*

4. Some properties of the solution. Recall that, from the outset, we have required both φ_1 and φ_2 to have the same wave speed c, because we are interested in solutions that move together. Also recall that we are declaring c and then allowing this choice of c to determine the values for a_i. The idea is this: given a wave speed c, we want to know how much the detuning parameters will have to change under different circumstances in order for the waves to stay together. We start by writing the a_i as functions of c and then by investigating the behavior of these functions. In particular, to find the range of propagation failure, we will compute the value of a_i as c approaches zero; this will tell us which values of a_i can support a nonzero wave speed. We also give an expression for the rate of change in the range of propagation failure when α moves away from zero, and we evaluate this expression explicitly in a few special cases.

4.1. Plots of $a_i(c)$ curves. Recall that from our original assumptions (2.6), we have that

$$a_i(c) = \varphi_i(\beta_i). \tag{4.1}$$

In Figures 3 and 4 we present $a_i(c)$ curves, which relate a given wave speed to the detuning parameters a_i. Take special note of the distance between $a_i(c)$ and $\frac{1}{2}$ at $c = 0$: this is the range of propagation failure mentioned several times already. Also note that smaller values of the diffusion constants D_i result in larger ranges of propagation failure. For $D_1 = D_2$, increasing the value of the D_i results in a smaller value of the detuning parameter required to achieve a certain wave speed.

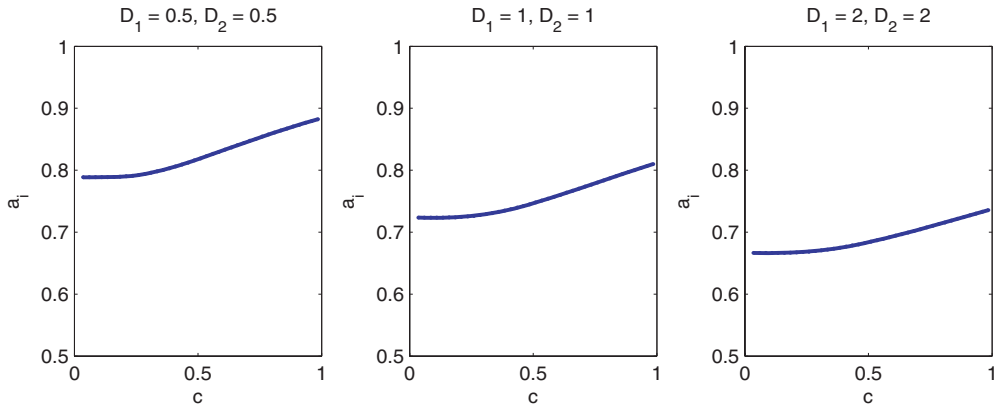


FIG. 3. *$a_i(c)$ curves for $A = .5, \beta = -.5$, and $\alpha = 0.1$. In this case, $a_1 = a_2$. Notice that $lim_{c\to 0}\ a_i(c)$ is significantly greater than $\frac{1}{2}$, implying a nontrivial range of propagation failure. Also notice that, for a given c, $a_i(c)$ decreases as $D_1 = D_2$ increases. Similar plots result in the case $A = 0 = \beta$.*

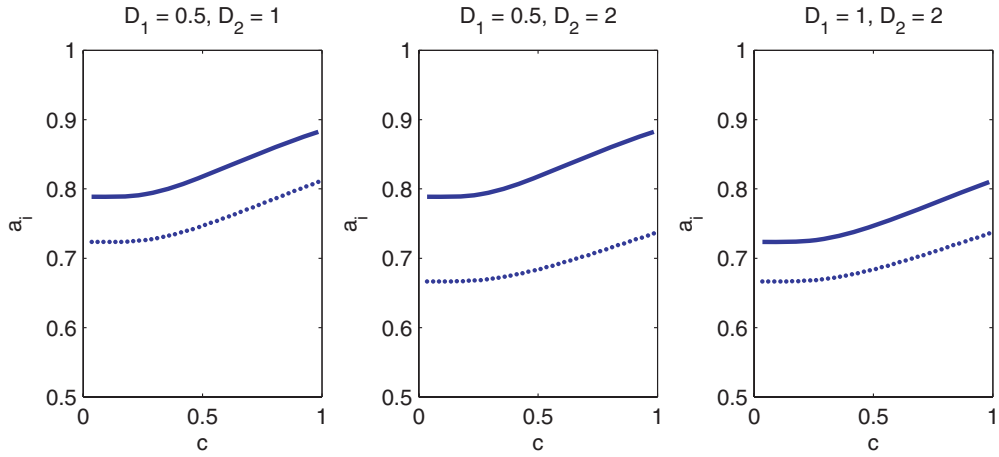


FIG. 4. *$a_i(c)$ curves for $A = .5, \beta = -.5$, and $\alpha = 0.1$, when $D_1 \neq D_2$. The solid line represents a_1, and the dotted line represents a_2. Similar plots result in the case $A = 0 = \beta$.*

4.2. Analytical results related to the $a_i(c)$ relationship. We comment on notation: some of the functions introduced in this section depend on several parameters; since there are so many parameters, we consider all parameters to be fixed except the ones listed explicitly in the argument of the function. Now consider the quantity

$$\Gamma_i(c) = a_i(c) - \frac{1}{2} = \frac{1}{2\pi}\int_0^\infty \sum_{j\in M_i} F_{i,j}(s)\cos(s(\beta_i - j)) + G_{i,j}(s)\frac{1}{s}\sin(s(\beta_i - j))ds. \tag{4.2}$$

Also define

$$g(c) = a_1(c) - a_2(c) = \Gamma_1(c) - \Gamma_2(c), \tag{4.3}$$

and note that this quantity represents the difference in the detuning parameters a_i required to make both waves travel at speed c. This definition allows us to state several important questions: 1. Given a noncoupled system with certain parameter

values, what happens to the quantity g when the coupling coefficient α moves away from zero? 2. Given a coupled, aligned system with certain parameter values, what happens to the quantity g when the alignment coefficient A moves away from zero? In this section we begin to answer these questions.

The following results about the oddness of the Γ_i illustrate the symmetry held by these functions under certain circumstances.

PROPOSITION 4.1. *Let $\Gamma_i(c)$ be defined as in* (4.2)*, and let $c \neq 0$. Then Γ_i is odd in c if and only if*

$$\int_0^\infty \sum_{j \in M_i} G_{i,j}(s) \frac{1}{s} \sin(s(\beta_i - j)) ds = 0 \tag{4.4}$$

for all $c \neq 0$. Further,

$$\lim_{c \to \pm\infty} \Gamma_i(c) = \frac{1}{2}. \tag{4.5}$$

Proof. Note that $X(s)$ and $Z(s)$, as given in (2.10), are even with respect to c, $W(s)$ and $Y(s)$ are odd with respect to c, and all other terms in Γ_i are independent of c. If the condition (4.4) holds, then we see from the definitions of $F_{i,j}$ and $G_{i,j}$ that all appearances of $X(s)$ and $Z(s)$ vanish, leaving us with Γ_i odd. Conversely, if the condition does not hold, then the dependence of Γ_i on c is not odd.

For the second statement, we use the change of variables $s \to \frac{s}{c}$, interchange the limiting process with the integration, and then evaluate the result:

$$\begin{aligned} \lim_{c \to +\infty} \Gamma_i(c) &= \frac{1}{2\pi} \int_0^\infty \lim_{c \to \infty} \sum_{j \in M_i} \cos\left(\frac{-sj}{c}\right) F_{i,j}\left(\frac{s}{c}\right) ds \qquad (4.6) \\ &= \frac{1}{2\pi} \int_0^\infty \lim_{c \to \infty} W\left(\frac{s}{c}\right) ds \\ &= \frac{1}{\pi} \int_0^\infty \frac{s^2+1}{s^4+2s^2+1} ds = \frac{1}{2}. \qquad (4.7) \end{aligned}$$

Once again, the limit as $c \to -\infty$ is evaluated in the same way, but the limits of integration change, much like in the proof of Proposition 3.2, giving us the result stated above. □

The second claim in this last theorem implies that increasing the wave speed to arbitrarily large values requires the detuning parameters to be very close to zero or one.

We also have the following corollary.

COROLLARY 4.2. *In particular, Γ_i is odd in the following cases:*

1. $\alpha = 0$,
2. $A = 0$ *and* $\beta = 0$,
3. $A = \frac{1}{2}$ *and* $\beta = -\frac{1}{2}$.

Proof. Conditions 1, 2, and 3 all imply (4.4) for all $c \neq 0$. □

Note that conditions 1, 2, and 3 are enough to imply the oddness of Γ_i, regardless of the values of the other parameters. It is not clear, however, whether there exist other solutions to the equation in (4.4), much less whether there exist other solutions that are independent of the other parameters.

Now consider the functions

$$P_Y(c) = \frac{1}{2\pi}\int_0^\infty \sum_{j\in M_1} F_{1,j}(s)\cos(s(\beta_1 - j)) - \sum_{j\in M_2} F_{2,j}(s)\cos(s(\beta_2 - j))ds$$

$$= \frac{k}{\pi}\int_0^\infty Y(s)C(s)[D_1 - D_2][1 + \alpha(1-A)\cos(s\beta) + \alpha A\cos(s(\beta+1))]ds, \tag{4.8}$$

$$P_Z^*(c) = \frac{1}{2\pi}\int_0^\infty \sum_{j\in M_1} G_{1,j}(s)\sin(s(\beta_1 - j)) - \sum_{j\in M_2} G_{2,j}(s)\sin(s(\beta_2 - j))ds$$

$$= \frac{k}{\pi}\int_0^\infty \frac{1}{s}Z(s)C(s)[D_1 + D_2][(1-A)\sin(s\beta) + A\sin(s(\beta+1))]ds. \tag{4.9}$$

In fact an α factors out of P_Z^*, so we will denote

$$P_Z(c) = \frac{P_Z^*(c)}{\alpha}, \tag{4.10}$$

which allows us to write

$$g(c) = P_Y(c) + \alpha P_Z(c), \tag{4.11}$$

where $g(c)$ is the difference between the detuning parameters a_1 and a_2, as given in (4.3). This representation of g will be helpful in proving the following results.

THEOREM 4.3. *Let g be defined as above. If either $\alpha = 0$ or $P_Z(c) = 0$, then*

$$c > 0 \Rightarrow g(c)\begin{cases} < 0 & \text{for } D_1 > D_2, \\ = 0 & \text{for } D_1 = D_2, \\ > 0 & \text{for } D_1 < D_2, \end{cases} \qquad c < 0 \Rightarrow g(c)\begin{cases} > 0 & \text{for } D_1 > D_2, \\ = 0 & \text{for } D_1 = D_2, \\ < 0 & \text{for } D_1 < D_2. \end{cases}$$

Proof. Clearly, if either assumption is true, then $g(c) = P_y(c)$. The only remaining task is to determine the sign of $P_y(c)$. To do this, note that $Y(s,c) > 0$ when $c > 0$ and $Y(s,c) < 0$ when $c < 0$ and that $C(s) \le 0$. Further, we have that

$$1 + \alpha(1-A)\cos(s\beta) + \alpha A\cos(s(\beta+1)) \tag{4.12}$$

$$> 1 - \alpha(1-A) - \alpha A = 1 - \alpha > 0, \tag{4.13}$$

and these estimates give the result above. ∎

The previous theorem determines the sign of g for a number of cases, since if either condition 2 or condition 3 from Corollary 4.2 is met, then $P_Z(c) \equiv 0$. The following results provide more examples of the symmetries of the functions Γ_i and g.

THEOREM 4.4. *Fix $\alpha \in [0,1)$. Then for $A \in (0,1)$,*

$$\Gamma_i(D_1, D_2, c, A, \beta) = -\Gamma_{i^*}(D_2, D_1, -c, A, \beta) = \Gamma_{i^*}(D_2, D_1, c, 1-A, -\beta - 1), \tag{4.14}$$

and for $A = 0$,

$$\Gamma_i(D_1, D_2, c, 0, \beta) = -\Gamma_{i^*}(D_2, D_1, -c, 0, \beta) = \Gamma_{i^*}(D_2, D_1, c, 0, -\beta). \tag{4.15}$$

Proof. Follows from the definitions of the functions $F_{i,j}$ and $G_{i,j}$ given in (2.9) and the definition of Γ_i given in (4.2). ∎

COROLLARY 4.5. *Fix $\alpha \in [0,1)$. Then for $A \in (0,1)$,*

$$\Gamma_i(D_1, D_2, c, A, \beta) = -\Gamma_i(D_1, D_2, -c, 1-A, -\beta - 1) \quad \text{and} \tag{4.16}$$

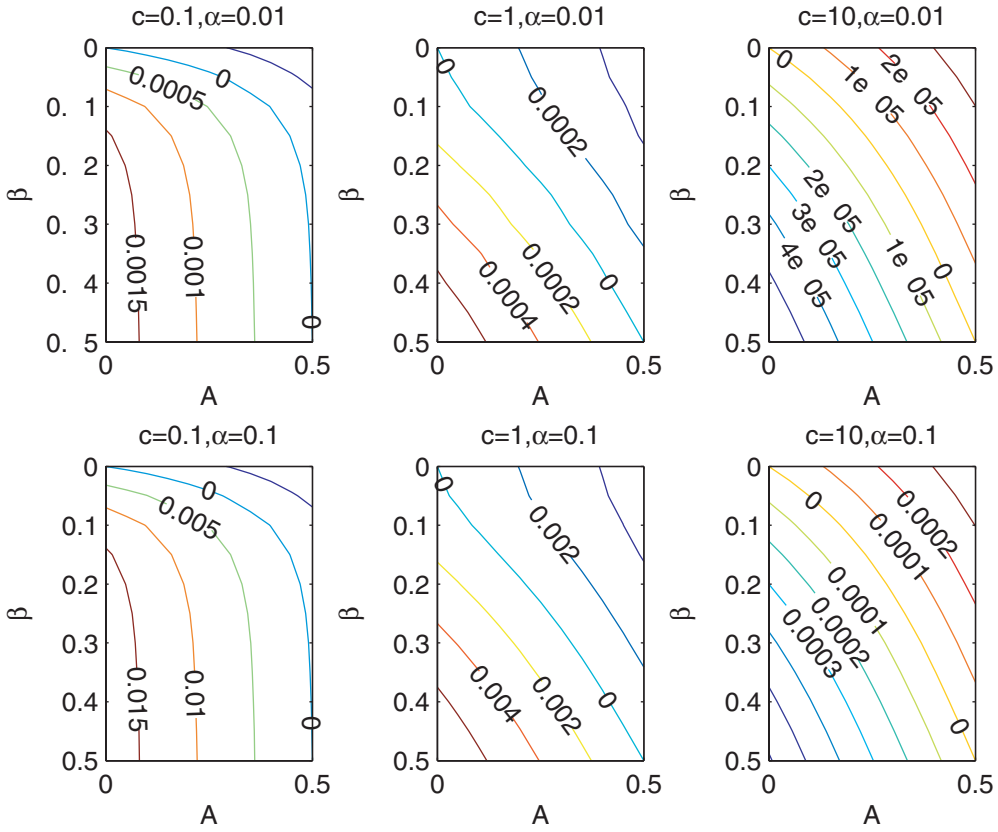


FIG. 5. *Contours of $g(A,\beta)$ for $D_1 = D_2 = 1$. Notice that for all $A \in [0, \frac{1}{2}]$, there is a β such that $g(A,\beta) = 0$.*

$$g(D_1, D_2, c, A, \beta) = -g(D_2, D_1, c, 1 - A, -\beta - 1) \tag{4.17}$$

$$= -g(D_1, D_2, -c, 1 - A, -\beta - 1) = g(D_2, D_1, -c, A, \beta), \tag{4.18}$$

and for $A = 0$,

$$\Gamma_i(D_1, D_2, c, 0, \beta) = -\Gamma_i(D_1, D_2, -c, 0, -\beta) \quad \textit{and} \tag{4.19}$$

$$g(D_1, D_2, c, 0, \beta) = -g(D_2, D_1, c, 0, -\beta) \tag{4.20}$$

$$= -g(D_1, D_2, -c, 0, -\beta) = g(D_2, D_1, -c, 0, \beta). \tag{4.21}$$

Proof. The proof follows from Theorem 4.4 and the definitions of Γ_i and g given in (4.2) and (4.3). □

Note that (4.14), (4.16), and (4.17) are also true for $A = 0$, but the statements given in (4.15), (4.19), and (4.20) are better results. We turn our attention now to the issue of nonalignment.

4.3. Numerical results related to nonalignment. We mentioned earlier that it is important to note which values of the parameters cause $a_1 = a_2$. This is equivalent, of course, to finding the zeros of our function g. The contour plots in Figure 5

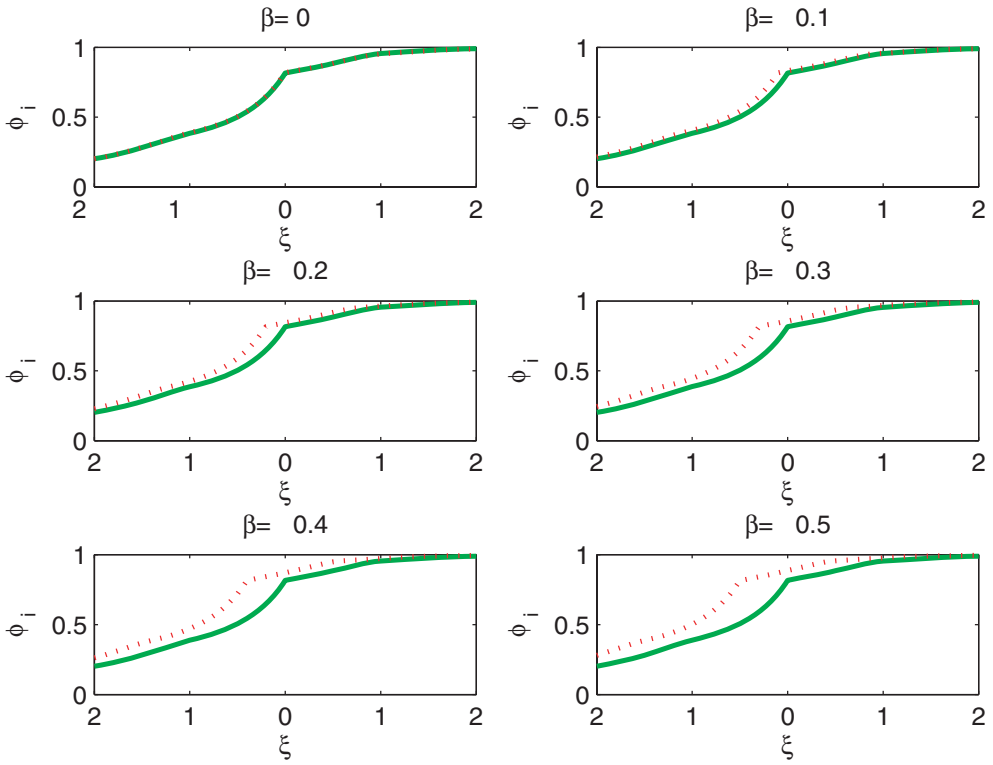


FIG. 6. *Wave profiles for* $D_1 = D_2 = 1$, $\alpha = .1$, $c = 1$, *and* $A = .3$. *The cusps in the graphs at* $a_i(\beta_i)$ *are the result of the Heaviside functions switching from* 0 *to* 1. *The solid line represents* φ_1, *and the dotted line represents* φ_2.

of $g = 0$ show that, given conditions with $A = 0$ that yield $g = 0$, we may move A away from zero, and, as long as we adjust β appropriately, g will remain zero.

For example, consider the bottom middle plot, in which $c = 1$ and $\alpha = .1$. When $A = 0 = \beta$, we see that $g(A, \beta) = 0$. In addition, for any $A \in [0, \frac{1}{2}]$, there is a $\beta \in [-\frac{1}{2}, 0]$ such that $g(A, \beta) = 0$. we see that this last statement is true of all plots in Figure 5, although we have not proved that such a β exists in $[-\frac{1}{2}, 0]$ in all circumstances.

Here we will fix α and examine the effects of moving the alignment parameter A away from zero. Recall our earlier mentioning of the parameter β as a time-delay. It is our contention that, loosely speaking, problems caused by moving A away from zero can be remedied by moving β a corresponding (but not necessarily equal) distance from zero. Consider, for instance, the case $\beta = 0$. Recalling our traveling wave ansatz $\xi = n - ct$, we see that at time $t = 0$, for instance, $\xi = 0$ will correspond with $n = 0$ for both fibers. However, if $A > 0$, then wave 2 will be "ahead" of wave 1. To make this notion a little more precise, we will speak of the "location" of a wave as being at β_i. Moving β below zero corrects this problem by having wave 2 reach a given node at a time β later than wave 1 reaches the same node. The upper left plot in Figure 6 gives an example of two waves that are lined up with respect to ξ when $\beta = 0$. Since $A \neq 0$, however, this lining up actually corresponds to one wave leading the other. It is only when β is moved below zero that the waves travel together, as indicated by the other plots in Figure 6.

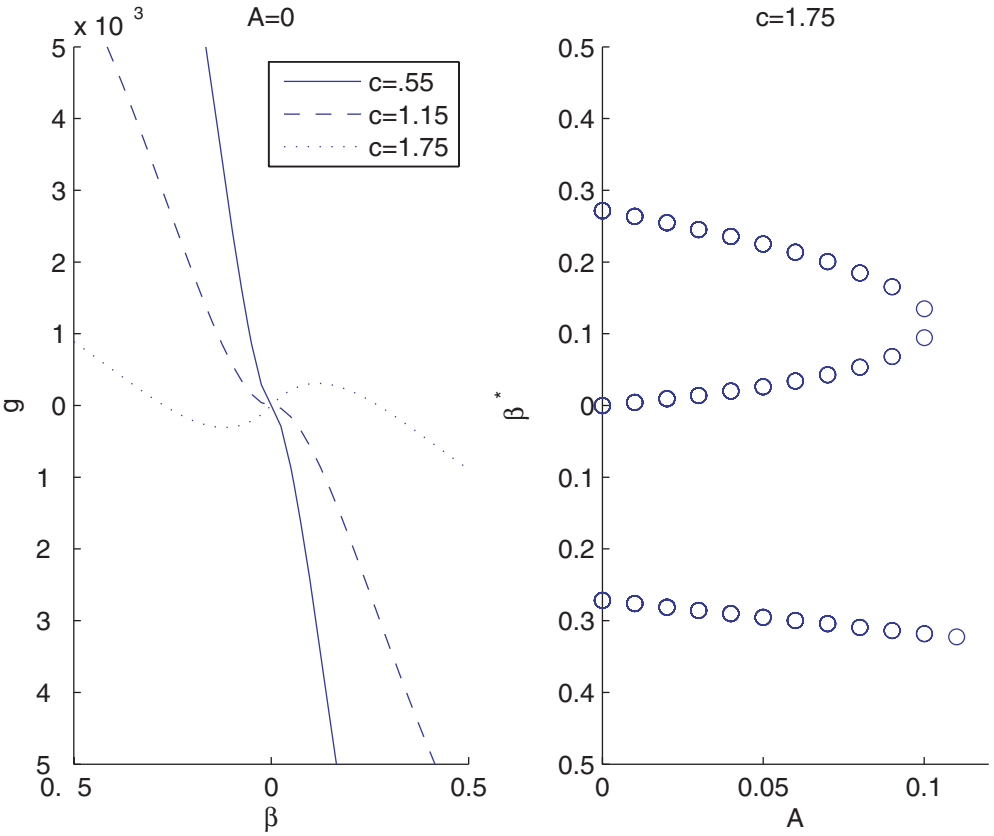


FIG. 7. *Left: plots of $g(\beta)$ for $A = 0$. When $c = 1.75$, we see that $g(\beta) = 0$ for three different values of β. Right: zeros of $g(\beta)$ for $c = 1.75$. This plot shows how the three zeros shown in the left plot for $c = 1.75$ change as A changes. In both plots, $D_1 = D_2 = 1$ and $\alpha = .1$.*

The relationship between A and β has been relatively unexplored analytically. We want to know if A and β are independent of each other, or if fixing a value of A determines a value of β. Consider an uncoupled system in which $D_1 = D_2$, and note that this is equivalent to the condition $a_1 = a_2$. We want to know, given fixed values of A and $\alpha > 0$, whether there exists a critical value of β that still results in $a_1 = a_2$. This section provides evidence that such a β does exist for a variety of combinations of A, α and c (see Tables 4.1 and 4.2).

Further, these numerical results show that multiple such β values exist for certain combinations of these parameters, particularly for higher wave speeds. We find evidence for the existence of a pitchfork bifurcation in the parameter c with respect to these critical β values. That is, fixing all parameters but c, there exists a c^* (dependent on the other parameters) such that for $c \leq c^*$ there is only one critical β value, but that for $c > c^*$, there are three critical β values. The left plot in Figure 7 shows the cubic-like shape of $g(\beta)$ and shows the transition between $g(\beta)$ having one zero to $g(\beta)$ having three zeros in the range shown. The plot on the right shows implicitly the dependence of this value c^* on the parameter A in particular. For smaller values of A, a wave speed of $c = 1.75$ results in three critical β values, whereas for large enough values of A, there is only one critical β value for which $g(\beta, c) = 0$. The possible implications of a pitchfork bifurcation are interesting. For instance, if the

TABLE 4.1
Values of β that yield $a_1 = a_2$ for $\alpha = .1$, $D_1 = D_2 = 1$, and different pairs of the wave speed c and the alignment parameter A.

c\ A	0	.1	.2	.3	.4	.5
.25	0	−.0433	−.0792	−.1328	−.2381	−.5
.75	0	−.0938	−.1651	−.2506	−.3623	−.5
1.25	0	−.1409	−.2200	−.3035	−.3974	−.5
1.75	0	−.3183	−.3634	−.4085	−.4540	−.5

TABLE 4.2
Values of $a_1 = a_2$ that result from the β value listed in Table 4.1 for different pairs of the wave speed c and the alignment parameter A.

c\ A	0	.1	.2	.3	.4	.5
.25	.7346	.7328	.7311	.7294	.7277	.7267
.75	.7891	.7878	.7864	.7852	.7844	.7840
1.25	.8521	.8510	.8500	.8492	.8487	.8485
1.75	.8975	.8959	.8955	.8952	.8950	.8949

single traveling wave before the bifurcation is stable, which is reasonable to expect for small values of the coupling parameter α (see [7]), then we might expect an exchange of stability at the bifurcation point.

The existence of multiple β values that result in $g = 0$ suggests that perhaps the two waves can travel at the same speed even if one wave is lagging behind the other. This allows for the possibility of two waves traveling together even if their respective nerve fibers were not activated at exactly the same time. Other numerical results indicate that the value of c^* increases as A increases. This suggests that in the case of staggered nodes, higher wave speeds are required in order to have multiple critical β values. The implications of this would be very interesting if there is in fact an exchange of stability at the bifurcation point. If the middle branch becomes unstable after the bifurcation, then this would imply that the middle branch stays stable for a larger range of the wave speed when the nodes are staggered than when they are aligned.

5. Propagation failure.

5.1. Analytical results on propagation failure. As mentioned earlier, propagation failure is a well-documented phenomenon in lattice differential equations (see [6] and [11], among others), and it is certainly important to compute the range of a_i that result in a zero wave speed. We address this issue here. This section involves a few lengthy calculations, so we highlight the most important results of this section with a few theorems and then proceed with the calculations that lead to the results. To determine the range of propagation failure, we need to compute $\lim_{c\to 0}\Gamma_i(c)$. In light of the symmetries of Γ_i presented previously, we will focus on the computation of $\lim_{c\to 0+}\Gamma_i(c)$. Also, we will start by considering only the case where $\Gamma_i(c)$ is odd, and the consideration of the other case will follow. These steps are carried out in the following proposition.

PROPOSITION 5.1. *Let $\alpha \in [0,1)$, $A \in [0,1)$, $\beta \in \mathbb{R}$, and $\gamma_i = \lim_{c\to 0+}\Gamma_i(c)$. Then*

$$\gamma_i = \lim_{c\to 0+}\Gamma_i(c) = \sum_{j\in M_i}\frac{1}{2\pi p}\int_0^\infty\int_0^p F^*_{i,j}\left(\frac{t}{c},s\right)\cos(s(\beta_i-j))dsdt + \int_0^\infty [G_{i,j}(s)]_{c=0}\frac{1}{s}\sin(s(\xi-j))ds. \tag{5.1}$$

Proof. We will compute $\lim_{c\to 0}\int_0^\infty F_{i,j}(s)\cos(s(\beta_i - j))ds$ individually for a given (i,j) pair. First, let's assume $\beta_i - j$ is rational, and then define

$$p_{i,j} := p^*_{i,j}2\pi := \operatorname{lcm}\left(2\pi, \frac{2\pi}{\beta_i - j}\right), \tag{5.2}$$

the common period of the periodic functions that define $F_{i,j}$, and

$$p = \operatorname{lcm}(p_{i,j}), \tag{5.3}$$

the least common multiple over all i and j of the $p_{i,j}$. For notational purposes, we consider W, X, Y, and Z as functions of two variables, e.g., $W(s) = W(u,v)$: the first argument will go in place of all nonperiodic instances of s, and the second argument will go in place of all periodic instances of s. As an example, we would write the function $q(s) = as^4 + bs^2 + sin(s)$ as $q(u,v) = au^4 + bu^2 + \sin(v)$.

For the $F_{i,j}$ and $G_{i,j}$ we will use the same convention. So, for instance,

$$F_{1,0}(u,v) = [W(u,v)(1+2Q) - R_2(v)Y(u,v)], \tag{5.4}$$

and $F_{i,j}(s,s) = F_{i,j}(s)$. Then

(5.5)

$$\begin{aligned}\frac{1}{2\pi}\int_0^\infty F_{i,j}(s)\cos(s(\beta_i - j))ds &= \frac{1}{2\pi}\sum_{n=0}^{\infty}\int_{np}^{(n+1)p} F_{i,j}(s)\cos(s(\beta_i - j))ds \\ &= \frac{1}{2\pi}\sum_{n=0}^{\infty}\left(\int_{np}^{(n+1)p} F_{i,j}(np,s)\cos(s(\beta_i - j))ds + E_n\right) \\ &:= \Phi + \frac{1}{2\pi}\sum_{n=0}^{\infty} E_n,\end{aligned}$$

where

$$E_n = \int_{np}^{(n+1)p} F_{i,j}(s)\cos(s(\beta_i - \xi))ds - \int_{np}^{(n+1)p} F_{i,j}(np,s)\cos(s(\beta_i - j))ds \tag{5.6}$$

and

$$\Phi = \frac{1}{2\pi}\sum_{n=0}^{\infty}\int_{np}^{(n+1)p} F_{i,j}(np,s)\cos(s(\beta_i - j))ds. \tag{5.7}$$

We will show in an appendix that

$$\sum_{n=0}^{\infty} E_n \to 0 \quad \text{as} \quad c \to 0, \tag{5.8}$$

but for now, we will assume this is true and complete the calculation of Φ. From the definitions of W and Y given in (2.10), notice that c factors out of the $F_{i,j}$. This allows us to write

$$F_{i,j}(s) = cF^*_{i,j}(s), \tag{5.9}$$

and we will denote

$$P_{i,j}(t) = \frac{1}{p}\int_0^p F^*_{i,j}(\frac{t}{c}, s)\cos(s(\beta_i - j))ds. \tag{5.10}$$

Then

$$\begin{aligned}\Phi &= \frac{c}{2\pi}\sum_{n=0}^{\infty}\int_{np}^{(n+1)p} F^*_{i,j}(np, s)\cos(s(\beta_i - j))ds \\ &= \frac{pc}{2\pi}\sum_{n=0}^{\infty}\frac{1}{p}\int_0^p F^*_{i,j}(np, s)\cos(s(\beta_i - j))ds \\ &= \frac{pc}{2\pi}\sum_{n=0}^{\infty} P_{i,j}(npc).\end{aligned} \tag{5.11}$$

Viewing this final quantity as a Riemann sum, we arrive at

$$\Phi = \frac{1}{2\pi}\int_0^{\infty} P_{i,j}(t)dt = \frac{1}{2\pi p}\int_0^{\infty}\int_0^p F^*_{i,j}\left(\frac{t}{c}, s\right)\cos(s(\beta_i - j))dsdt. \tag{5.12}$$

This brings us to the expression, for Γ_i odd,

$$\gamma_i = \lim_{c\to 0}\Gamma_i(c) = \sum_j \frac{1}{2\pi p}\int_0^{\infty}\int_0^p F^*_{i,j}\left(\frac{t}{c}, s\right)\cos(s(\beta_i - j))dsdt. \tag{5.13}$$

If Γ_i is not odd, then we need to consider the effect of the $G_{i,j}$. However, this is much easier to compute since

$$\lim_{c\to 0}\int_0^{\infty} G_{i,j}(s)\frac{1}{s}\sin(s(\xi - j))ds = \int_0^{\infty}\left[G_{i,j}(s)\right]_{c=0}\frac{1}{s}\sin(s(\xi - j))ds, \tag{5.14}$$

and so we have the general formula

$$\begin{aligned}\gamma_i = \lim_{c\to 0}\Gamma_i(c) = &\sum_j \frac{1}{2\pi p}\int_0^{\infty}\int_0^p F^*_{i,j}\left(\frac{t}{c}, s\right)\cos(s(\beta_i - j))dsdt \\ &+ \int_0^{\infty}\left[G_{i,j}(s)\right]_{c=0}\frac{1}{s}\sin(s(\xi - j))ds. \qquad \square\end{aligned} \tag{5.15}$$

One of our goals here is to investigate the changes in this quantity with respect to a change in α and, in particular, the difference in the range of propagation failure between when $\alpha = 0$ and when $\alpha > 0$. To this end, we let $\gamma_i = \gamma_i(\alpha)$, calculate $\gamma_i'(\alpha)$, and evaluate this expression at $\alpha = 0$. In a few special cases, we have an explicit value for $\gamma_i'(\alpha)$ in terms of D_1 and D_2. The following theorem gives this result.

THEOREM 5.1. *Let $D_1 = D_2 := D$, and consider the function $\gamma_i = \gamma_i(\alpha)$.*

1. *If $A = 0 = \beta$, then*

$$\gamma_i'(0) = \frac{D}{(4D+1)^{\frac{3}{2}}} > 0. \tag{5.16}$$

2. *If $A = \frac{1}{2}$ and $\beta = -\frac{1}{2}$, then*

$$\gamma_i'(0) = 0. \tag{5.17}$$

Recall that $\gamma_i = \lim_{c\to 0+} a_i(c) - \frac{1}{2}$. If $\gamma_i > 0$, then there is a positive range of propagation failure. The results above quantify how the size of this range changes as α moves away from zero. Before we proceed with the proof of this theorem, let us interpret the results in terms of the ephaptic coupling model. This theorem states that turning the coupling on results in an increase in the range of propagation failure when the nodes are aligned, but that turning the coupling on results in a decrease in the range of propagation failure when the nodes are staggered. Intuitively, it seems reasonable that staggering the nodes would result in a decrease in this range. The two-fiber problem with staggering seems very much like the one-fiber problem with twice as many nodes squeezed into the same space. A shorter internodal distance is reflected by an increase in the diffusion coefficient, which has the effect of decreasing the range of propagation failure. As the coupling increases, the fibers interact more, increasing the effect of the shorter internodal distance.

The result for the aligned case is more easily understood by thinking about the system (2.7). When $A = 0 = \beta$, we are left with

$$\begin{aligned} -c\varphi_1'(\xi) &= \frac{1}{1-\alpha^2}\left[D_1(L\varphi_1)(\xi) - \alpha D_2(L\varphi_2)(\xi)\right] - \varphi_1(\xi) + h(\xi), \\ -c\varphi_2'(\xi) &= \frac{1}{1-\alpha^2}\left[-\alpha D_1(L\varphi_1)(\xi) + D_2(L\varphi_2)(\xi)\right] - \varphi_2(\xi) + h(\xi), \end{aligned} \tag{5.18}$$

and when $D_1 = D_2$, we expect to have $\varphi_1 = \varphi_2$. In this situation, the system would become two copies of the equation

$$\begin{aligned} -c\varphi'(\xi) &= \frac{1}{1-\alpha^2}\left[D(1-\alpha)(L\varphi)(\xi)\right] - \varphi(\xi) + h(\xi) \\ &= \frac{D}{1+\alpha}(L\varphi)(\xi) - \varphi(\xi) + h(\xi), \end{aligned} \tag{5.19}$$

which is just the one-fiber problem with diffusion coefficient $\frac{D}{1+\alpha}$. This makes it easy to see that increasing α decreases the diffusion coefficient in this related one-fiber problem, which has the effect of increasing the range of propagation failure. Using the value of γ computed for the one-fiber problem in [6], and viewing γ as a function of α, we have

$$\gamma(\alpha) = \frac{1}{2(1+\frac{4D}{1+\alpha})^{\frac{1}{2}}} \tag{5.20}$$

and

$$\gamma'(\alpha) = \frac{D}{(1+\alpha)^2(1+\frac{4D}{1+\alpha})^{\frac{3}{2}}}. \tag{5.21}$$

In particular,

$$\gamma'(0) = \frac{D}{(1+4D)^{\frac{3}{2}}}, \tag{5.22}$$

which is in agreement with the result obtained in the theorem. This formula also sheds light on how the range of propagation failure might decrease for all values of α. Unfortunately, such an argument cannot be made when $A \neq 0$. However, the reasoning used here works for any nonlinearity f for which the wave speed increases

with the diffusion coefficient. In particular, we expect a similar claim holds for a cubic nonlinearity, although we are not able to determine the size of the range quantitatively in that case.

Note that here we have an expression for $\gamma'(0)$ when $A = 0 = \beta$ and when $A = \frac{1}{2}, \beta = -\frac{1}{2}$, but not for values of A in $(0, \frac{1}{2})$. We are interested in the quantity $\gamma'(0)$ primarily when $a_1 = a_2$. That is, if we consider $\gamma = \gamma(\alpha, A, \beta)$, we want to compute $\left.\frac{\partial \gamma}{\partial \alpha}\right|_{(\alpha,A,\beta^*)=(0,A,\beta^*)}$ for $A \in (0, \frac{1}{2})$, where β^* is such that $a_1 = a_2$. The difficulty in extending the results of Theorem 5.1 to $A \in (0, \frac{1}{2})$ lies in determining analytically the relationship between A and β^*.

We now turn our attention to the proof of Theorem 5.1.

Proof. Let $\gamma_i = \gamma_i(\alpha)$ and let $W, X, Y, Z, F_{i,j}$, and $G_{i,j}$ be functions of α as well. We would like to calculate $\gamma_i'(0)$. To do this note that

$$W'(0) = X'(0) = Y'(0) = Z'(0) = Q'(0) = 0, \tag{5.23}$$

and from this we obtain

$$F'_{i,\beta_i}(0) = F'_{i,\beta_i+1}(0) = F'_{i,\beta_i-1}(0) = G'_{i,\beta_i}(0) = G'_{i,\beta_i+1}(0) = G'_{i,\beta_i-1}(0) = 0 \tag{5.24}$$

and

$$\begin{aligned} F'_{i,\beta^*}(0) &= -(1-A)R_{i^*}Y, \\ F'_{i,\beta^*+1}(0) &= -AR_{i^*}Y, \\ G'_{i,\beta^*}(0) &= -(1-A)R_{i^*}Z, \\ G'_{i,\beta^*+1}(0) &= -AR_{i^*}Z. \end{aligned} \tag{5.25}$$

Since $D_1 = D_2 := D$, we have $R_1 = R_2 := R$, and

$$\gamma_i'(0) = \frac{2}{\pi p}\int_0^p \int_0^\infty \frac{T(s)}{(t^2 + T^2(s))^2} P(s)dtds - \frac{1}{\pi}\int_0^\infty \frac{1}{sT^2(s)}Q(s)ds, \tag{5.26}$$

where

$$\begin{aligned} T(s) &= 2DC(s) - 1, \\ P(s) &= R(s)\left[(1-A)\cos\left(s(\beta_i - \beta_{i^*})\right) + A\cos\left(s(\beta_i - (\beta_{i^*}+1))\right)\right], \\ Q(s) &= R(s)\left[(1-A)\sin\left(s(\beta_i - \beta_{i^*})\right) + A\sin\left(s(\beta_i - (\beta_{i^*}+1))\right)\right]. \end{aligned} \tag{5.27}$$

Some calculation yields that the complex equation $T(z) = 0$ has roots in

$$\left\{ z = a + ib \;\middle|\; a = 2m\pi \text{ for } m \in \mathbb{Z}, b = \cosh^{-1}\left(1 + \frac{1}{2D}\right)\right\}. \tag{5.28}$$

We focus on the calculation of the inner integral in the first term of (5.26). The complex roots of $z^2 + T^2(s))^2 = 0$ are $z = \pm iT(s)$. Note that since $T(s) < 0$ for real s, $P_1 = -iT(s)$ lies in the upper half-plane. Denoting

$$f(z) = \frac{1}{(z^2 + T^2(s))^2} = \frac{1}{(z - iT(s))^2(z + iT(s))^2}, \tag{5.29}$$

we have that

$$\operatorname{res}_f(P_1) = \frac{i}{4T^3(s)}. \tag{5.30}$$

Then

$$\int_0^\infty \frac{T(s)}{(t^2+T^2(s))^2}P(s)dt = \pi i \mathrm{res}_f(P_1) = \frac{-\pi}{4T^3(s)} \tag{5.31}$$

and

$$\frac{2}{\pi p}\int_0^p \int_0^\infty \frac{T(s)}{(t^2+T^2(s))^2}P(s)dtds = -\frac{1}{2p}\int_0^p \frac{1}{T^2(s)}P(s)dt := \Psi. \tag{5.32}$$

Let

$$p^* = \mathrm{lcm}(p^*_{i,j}), \text{ and } q_{i,j} = \frac{p^*}{p^*_{i,j}}. \tag{5.33}$$

To compute the integral Ψ, we use the change of variables $z = e^{\frac{is}{p^*}}$, $dz = \frac{i}{p^*}e^{\frac{is}{p^*}}ds = \frac{iz}{p^*}ds$, which results in a contour integral around the unit circle. Now we may write

(5.34)

$$T(z) = 2D\left(\frac{z^{p^*}+z^{-p^*}}{2}-1\right)-1,$$

$$P(z) = 2D\left(\frac{z^{p^*}+z^{-p^*}}{2}-1\right)\left[(1-A)\left(\frac{z^{q_{i,\beta_{i^*}}}+z^{-q_{i,\beta_{i^*}}}}{2}\right)+A\left(\frac{z^{q_{i,\beta_{i^*+1}}}+z^{-q_{i,\beta_{i^*+1}}}}{2}\right)\right].$$

Multiplying the top and bottom of the integrand by z^{2p^*-1} leaves the denominator as the square of a quadratic in z^{p^*}, with zeros given by

$$z_\pm^{p^*} = 1+\frac{1}{2D}\pm\frac{\sqrt{4D+1}}{2D}. \tag{5.35}$$

We have $|z_-^{p^*}|<1<|z_+^{p^*}|$, so the integrand has p^* distinct poles inside the unit circle, each of order 2. Labeling these poles $r_1,\dots,r_{p^*}$ we have

$$\Psi = -\frac{q}{i}\frac{1}{2p}2\pi i \mathrm{res}\sum_{m=1}^{q}\left(\frac{P(z)}{T^2(z)},r_i\right) = -\frac{1}{2}\mathrm{res}\sum_{m=1}^{q}\left(\frac{P(z)}{T^2(z)},r_i\right). \tag{5.36}$$

This quantity can be computed explicitly in several special cases. We will start with $A=0$, $\beta=0$. In this case, $P(s)=R(s)$, $p^*=1$, and $Q(s)=0$. Then we have

$$r_1 = 1+\frac{1}{2D}-\frac{\sqrt{4D+1}}{2D} \tag{5.37}$$

and

$$\mathrm{res}\left(\frac{P(z)}{T^2(z)},r_1\right) = \frac{-2D}{(4D+1)^{\frac{3}{2}}}, \tag{5.38}$$

which gives us

$$\gamma_i'(0) = \frac{D}{(4D+1)^{\frac{3}{2}}} > 0. \tag{5.39}$$

For $A = \frac{1}{2}$, $\beta = -\frac{1}{2}$, we have $P(s) = R(s)cos(\frac{s}{2})$, $p^* = 2$, and $Q(s) = 0$. We use the roots

$$r_1, r_2 = \pm \frac{\sqrt{2D(2D + 1 - \sqrt{4D + 1})}}{2D} \tag{5.40}$$

and

$$\operatorname{res}\left(\frac{P(z)}{T^2(z)}, r_1\right) + \operatorname{res}\left(\frac{P(z)}{T^2(z)}, r_2\right) = 0, \tag{5.41}$$

implying $\gamma_i'(0) = 0$ in this case. □

Unfortunately, explicit results like those given above are very difficult to find for more general parameter values. The problem lies in the computation of the second integral in equation (5.26). In an attempt to carry out this calculation, we see that the integral

$$\Omega = \frac{1}{\pi} \int_0^\infty \frac{Q(z)}{zT^2(z)} dz \tag{5.42}$$

can be evaluated using the calculus of residues. The poles of this integrand are given by

$$r_m = 2m\pi + i\cosh^{-1}\left(1 + \frac{1}{2D}\right), \tag{5.43}$$

which are of order 2. Then

$$\Omega = \pi i \sum_{m=-\infty}^{\infty} \operatorname{res}\left(\frac{Q(z)}{zT^2(z)}, r_m\right). \tag{5.44}$$

Now using $'$ to denote differentiation with respect to the complex variable z, we write

$$\operatorname{res}\left(\frac{Q(z)}{zT^2(z)}, r_m\right) = -\frac{Q'(r_m)}{r_m(4D + 1)} + \frac{Y(r_m)}{r_m D(4D + 1)}\left[\frac{1}{r_m} - \frac{i(2D + 1)\sqrt{4D + 1}}{4D + 1}\right]. \tag{5.45}$$

Since $\beta_i - \beta_{i^*}$ may take on any rational value, it is very difficult to simplify, in general, the expressions for $Q'(r_m)$ and $Y(r_m)$.

5.2. Numerical results on propagation failure. Figures 8 and 9 illustrate the effect of moving the coupling parameter α away from zero. Notice that when $A = 0$, $\beta = 0$, an increase in the coupling parameter α results in an increase in the range of propagation failure. However, when $A = .5$, $\beta = -.5$, this range decreases when α increases, suggesting that in the case of myelinated axons, nonalignment of the nodes of Ranvier allows for propagation in a larger range of values of the detuning parameter.

The plots in Figure 10 show the effects of coupling for positive wave speeds away from zero. Note that although our analytical expressions describe a_i as a function of c, these plots allow us to see what happens to the wave speed if we fix a value for $a_1 = a_2$ and then begin coupling. The results are especially interesting for small wave speeds: these plots indicate that coupling increases the wave speed for small wave

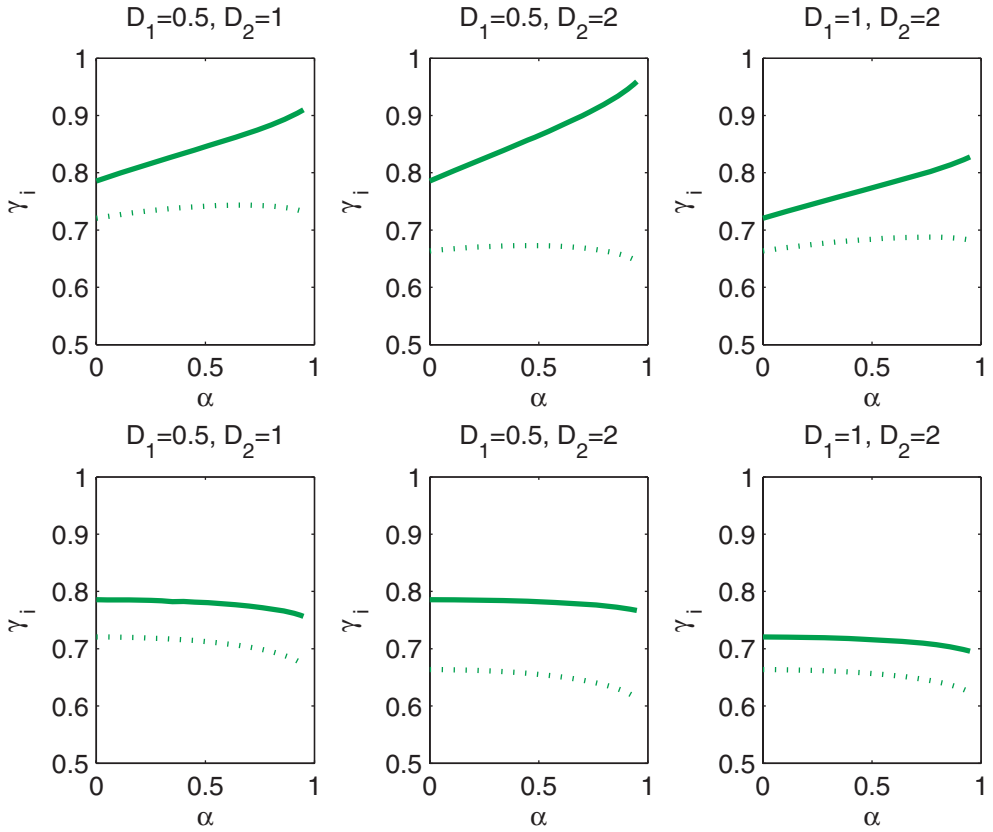


FIG. 8. *In the top plots, $A = 0 = \beta$; in the bottom plots, $A = .5$, $\beta = -.5$. The plots show the range of propagation failure when $D_1 \neq D_2$. The solid line represents a_1 and the dotted line a_2. (Note: values for failure were computed at $c = 10^{-2}$.)*

speeds when $A = .5$ and $\beta = -.5$ but that the same is not true when $A = 0 = \beta$. This also agrees with the results shown in Figure 9 when $D_1 = D_2$, and the results obtained for $\gamma_i'(\alpha)|_{\alpha=0}$ in previous sections.

6. Conclusion. We have shown that traveling back solutions to our system exist (at least) for small values of the coupling coefficient α, although further study is required to know the size of the range of α for which solutions exist. Using these solutions, we find that if the nodes are perfectly lined up or evenly staggered, and the waves travel together, then the sign of $g(c) = a_1 - a_2$ remains unchanged for all $\alpha \in [0, 1)$.

If we limit ourselves to the case $D_1 = D_2$, we can comment on the questions we set out to answer: namely, what are the effects of nonalignment and ephaptic coupling on the propagation of action potentials? We find that (a) if the nodes are perfectly lined up and the waves travel together, then the introduction of ephaptic coupling increases the size of the range of propagation failure, and (b) if the nodes are evenly staggered and the waves travel together, then the introduction of ephaptic coupling decreases the the size of the range of propagation failure.

The first result, in particular, agrees in spirit with results obtained in [2] with the cubic nonlinearity. There the authors use numerical methods to show that larger

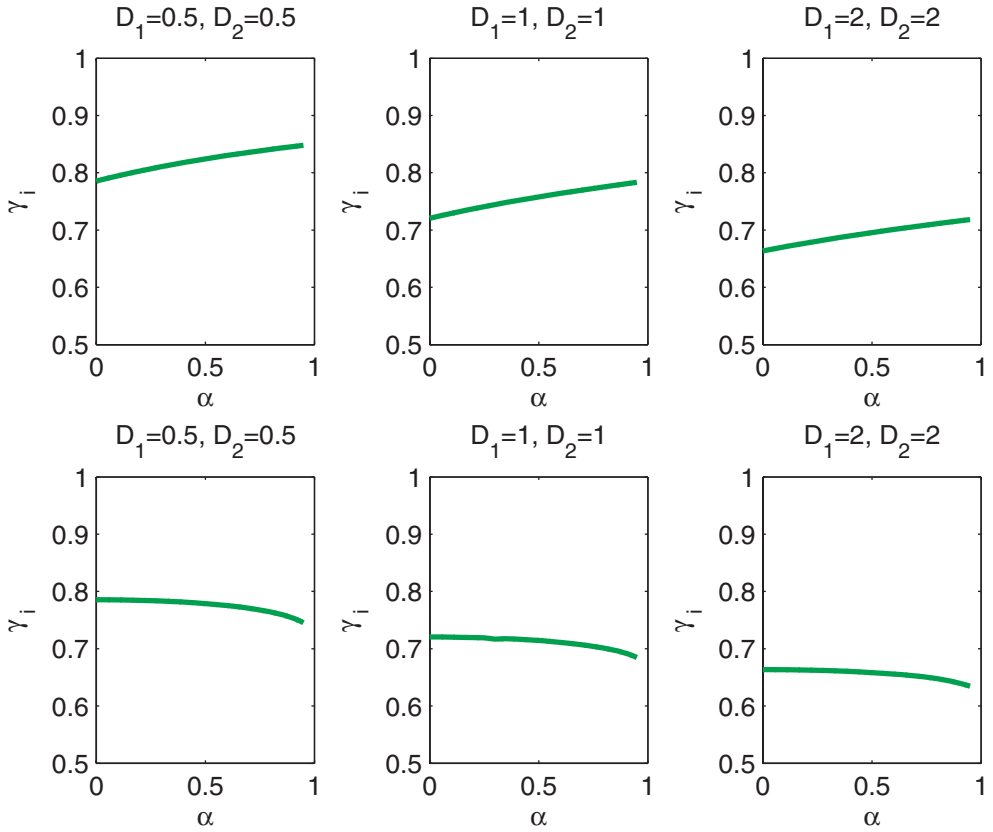


FIG. 9. *In the top plots, $A = 0 = \beta$; in the bottom plots, $A = .5$, $\beta = -.5$. The plots show the range of propagation failure when $D_1 = D_2$. In this case, $a_1 = a_2$. Note how the ranges increase as α increases on the top but decrease as α increases on the bottom. (Note: values for failure were computed at $c = 10^{-2}$.)*

values of the diffusion coefficient are required to achieve a given wave speed (in particular, small wave speeds) as the coupling increases. They give similar results when the nodes are staggered, except that for small wave speeds, coupling appears to have little or no effect on the value of the diffusion coefficient required to achieve a certain wave speed. These results may not be compared directly to ours, however, since the results in [2] relate to the effect of coupling on the diffusion coefficient, whereas our results relate to the effect of coupling on the detuning parameters. In [20] and [21], evidence is given that ephaptic coupling can have a significant effect on the propagation of action potentials in a model for cardiac cells, but these results focus on the length of time required for an action potential on one fiber to affect the action potential on the other fiber, as opposed to the amount of time required for action potentials to proceed along a given fiber. The work [20, 21] considers two types of coupling at once, ephaptic and ohmetric, and that approach may provide further insight into the dynamics of our present problem. The work in [3] is concerned with ephaptic coupling in nonmyelinated nerve fibers that might be modeled with (1.1) with large D. In [16], the authors focus on the effects of demyelination of nerve fibers in large bundles and observe a reduction in the speed of action potential propagation when the nodes of Ranvier are aligned.

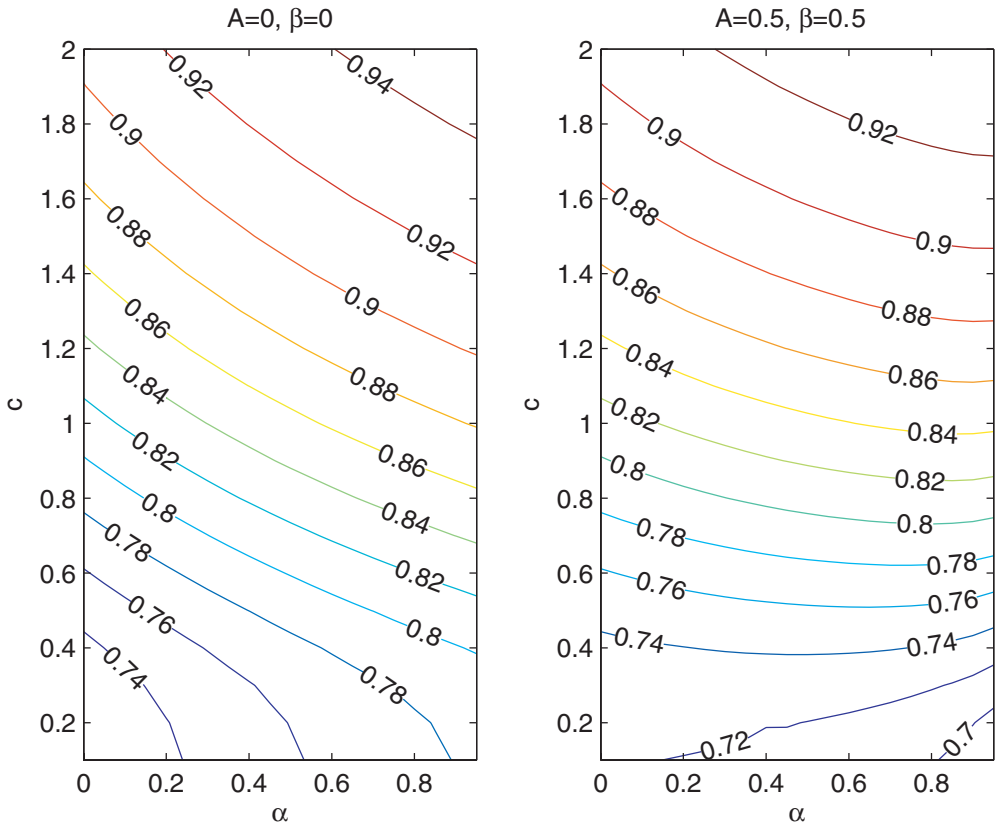


FIG. 10. *Level curves of $a_1(\alpha, c)$ when $D_1 = D_2 = 1$. Note that this implies $a_1 = a_2$ for both plots. In the right plot, the value of $a_i(c)$ decreases with α for small c, whereas in the left plot, the value of $a_i(c)$ increases with α for small c. This is another illustration of the claim that coupling decreases the range of propagation failure when the nodes are staggered but increases the range when the nodes are aligned.*

In addition, we have shown that for any value of A, there is a value of β such that $a_1 = a_2$. This means that regardless of the alignment of the nodes, we can have a situation in which the fibers are the same (i.e., $D_1 = D_2$ and $a_1 = a_2$) and the waves travel together. Perhaps a more surprising result is that with some parameter configurations, there are multiple such β values. The existence of these multiple β values gives rise to questions involving the stability of the solutions that we cannot answer with any certainty. The results in [2], and those in [1], point to the tendency of the two waves to match speeds (although in [1] a stronger type of coupling is considered). Given this fact, it seems reasonable that the same-speed solutions would be stable. However, the earlier results deal with waves that travel together at the same speed, and our results suggest the possibility of waves that travel at the same speed, with one wave lagging behind the other. This difference may give rise to interesting behavior with respect to the stability of the solutions.

Experience with the uncoupled case suggests that the piecewise linear and cubic problems may exhibit qualitatively similar behavior for small values of c and small ($D \leq 1$) values of D. It is reasonable to expect that qualitatively similar results related to propagation failure may hold in the cubic case, provided D is not too large.

7. Appendix A. We now present the details of the construction of our solution.

LEMMA 7.1. *Let* (φ_1, φ_2) *be a solution to* (2.5), (2.7). *Then there exists* $\varepsilon_0 > 0$ *such that for* $i = 1, 2$,

$$|\varphi_i(\xi)| \leq Ke^{\varepsilon_0 \xi} \text{ for } \xi < \min\{-1, \beta\} \tag{7.1}$$

for some $K > 0$.

Proof. Let $\psi_i(\xi) = \varphi_i(-\xi)$, and note that from our boundary conditions, $\psi_i(\xi) \to 0$ as $\xi \to \infty$. Then from the system (2.7), we have

$$\begin{aligned} c\psi_1'(\xi) + \psi_1(\xi) = \frac{1}{1-\alpha^2}[&D_1(L\psi_1)(\xi) - \alpha D_2(L\psi_2)(\xi) \\ &- \alpha A(N^*\psi_2)(\xi) - \alpha^2 A(N * \psi_1)(\xi + 1)], \\ c\psi_2'(\xi) + \psi_2(\xi) = \frac{1}{1-\alpha^2}[&-\alpha D_1(L\psi_1)(\xi) + D_2(L\psi_2)(\xi) \\ &+ \alpha^2 A(N^*\psi_2)(\xi) + \alpha A(N^*\psi_1)(\xi + 1)] \end{aligned}$$

for all $\xi > \max\{1, -\beta\}$, where

$$(N^*\psi_i)(\xi) = c(\psi'(\xi + 1) - \psi'(\xi)) + \psi(\xi + 1) - \psi(\xi). \tag{7.2}$$

For this proof only, let $\hat{\psi}_i$ denote the Laplace transform

$$\hat{\psi}_i(s) = \frac{1}{2\pi} \int_0^\infty e^{-s\xi} \psi_i(\xi) d\xi. \tag{7.3}$$

After applying the transform, we obtain the matrix equation

$$M^*(s) \begin{bmatrix} \hat{\psi}_1(s) \\ \hat{\psi}_2(s) \end{bmatrix} = J^*(s) \begin{bmatrix} \psi_1(0) \\ \psi_2(0) \end{bmatrix} + K^*(s), \tag{7.4}$$

where M^* and J^* are 2×2 matrix functions with entries

$$\begin{aligned} M_{11}^*(s) &= B^*(s) - 2kD_1C^*(s) + k\alpha^2 AE(s)B(s), \\ M_{12}^*(s) &= 2k\alpha D_2C^*(s) - k\alpha AE(-s)B(s), \\ M_{21}^*(s) &= 2k\alpha D_1C^*(s) - k\alpha AE(s)B(s), \\ M_{22}^*(s) &= B^*(s) - 2kD_2C^*(s) + k\alpha^2 AE(-s)B(s), \end{aligned} \tag{7.5}$$

and

$$\begin{aligned} J_{11}^*(s) &= c(1 + k\alpha^2 AE^*(-s)), \\ J_{12}^*(s) &= -ck\alpha AE^*(s), \\ J_{21}^*(s) &= -ck\alpha AE^*(-s), \\ J_{22}^*(s) &= c(1 + k\alpha^2 AE^*(s)), \end{aligned} \tag{7.6}$$

$$\begin{aligned} k &= \frac{1}{1-\alpha^2}, \\ C^*(s) &= \cosh(s) - 1, \\ B^*(s) &= 1 + cs, \\ E^*(s) &= 1 - e^s, \end{aligned} \tag{7.7}$$

and K^* is a vector with entries

$$\begin{aligned} K_1^*(s) &= k(D_1\Delta_1 - \alpha D_2\Delta_2) + k\alpha A(\Omega_2 + \alpha\Omega_1), \\ K_2^*(s) &= k(D_2\Delta_2 - \alpha D_1\Delta_1) - k\alpha A(\Omega_1 + \alpha\Omega_1), \end{aligned} \tag{7.8}$$

where

$$\begin{aligned} \Delta_i &= e^{-s}\int_{-1}^{0} e^{-s\xi}\psi_i(\xi)d\xi - e^{s}\int_{0}^{1} e^{-s\xi}\psi_i(\xi)d\xi, \\ \Omega_i &= \pm e^{\mp s}\left(\int_{\mp 1}^{0} e^{-s\xi}\psi_i'(\xi)d\xi + \int_{\mp 1}^{0} e^{-s\xi}\psi_i(\xi)d\xi\right). \end{aligned}$$

Integrating by parts, we see that

$$\int_0^1 e^{-s\xi}\psi_i(\xi) = \left.\frac{-e^{-s\xi}\psi_i(\xi)}{s}\right|_{\xi=0}^{1} + \frac{1}{s}\int_0^1 e^{-s\xi}\psi_i'(\xi)d\xi \tag{7.9}$$

and

$$\int_0^1 e^{-s\xi}\psi_i'(\xi) = \left.\frac{-e^{-s\xi}\psi_i'(\xi)}{s}\right|_{\xi=0}^{1} + \frac{1}{s}\int_0^1 e^{-s\xi}\psi_i''(\xi)d\xi \tag{7.10}$$

As long as these integrals are defined, we see that $K(s) = O(|s|^{-1})$ as $|\text{Im } s| \to \infty$, uniformly for Re s bounded. Assuming M^* is invertible, we obtain

$$\begin{bmatrix} \hat{\psi}_1(s) \\ \hat{\psi}_2(s) \end{bmatrix} = (M^*)^{-1}J^*(s)\begin{bmatrix} \psi_1(0) \\ \psi_2(0) \end{bmatrix} + O(|s|^{-2}) \quad \text{as} \quad |\text{Im } s| \to \infty \tag{7.11}$$

uniformly for Re s bounded, since $(M^*)^{-1} = O(|s|^{-2})$. This justifies the inversion of the Laplace transform, as well as the shift of contour around singularities on the imaginary axis. This gives us

$$\begin{aligned} \psi_i(\xi) &= \frac{1}{2\pi i}\int_{\varepsilon_0 - i\infty}^{\varepsilon_0 + i\infty} e^{s\xi}\hat{\psi}_i(s)ds \\ &= \frac{1}{2\pi i}\int_{-\varepsilon_0 - i\infty}^{-\varepsilon_0 + i\infty} e^{s\xi}\hat{\psi}_i(s)ds + \frac{1}{2\pi i}\sum_{\text{Re } s=0} e^{s\xi}\text{res}(\hat{\psi}_i, s), \end{aligned} \tag{7.12}$$

where the first equality results from the Laplace inversion formula, and the second is a contour shift around any poles of $\hat{\psi}_i$ in the imaginary axis, and where ε_0 is small enough so that there are no poles of $\hat{\psi}_i$ for $-\varepsilon_0 \le \text{Re } s < 0$. Further, if

$$(M^*)^{-1}J^*(s)\begin{bmatrix} \psi_1(0) \\ \psi_2(0) \end{bmatrix} = O(|s|^{-1}) \text{as} \quad |\text{Im } s| \to \infty \tag{7.13}$$

for $\varepsilon < 0$, then $\hat{\psi}_i = O(|s|^{-1})$ as $|\text{Im } s| \to \infty$.

We know that this condition is met, since $(M^*(s))^{-1} = O(|s|^{-1})$ and J^* oscillates as $|\text{Im } s| \to \infty$. Since we also have, from our boundary conditions, that $\psi_i(\xi) \to 0$ as $\xi \to \infty$, it is clear that

$$\text{res}(\hat{\psi}_i, s) = 0 \quad \text{for} \quad \text{Im } s = 0. \tag{7.14}$$

This fact, together with (7.11) and (7.12), gives us that

$$\psi_i(\xi)| \leq Ke^{-\varepsilon_0\xi} \quad \text{for} \quad \xi > \max\{1, -\beta\} \tag{7.15}$$

for $K > 0$, which is equivalent to the claim made in the statement of the lemma. □

Now define

$$\varphi_{i,\varepsilon}(\xi) = e^{-\varepsilon\xi}\varphi_i(\xi), \tag{7.16}$$

where $\varepsilon > 0$ is sufficiently small. By Lemma 7.1, if (φ_1, φ_2) satisfies (2.5), (2.7), then $\varphi_{i,\varepsilon}(\xi) \to 0$ exponentially fast as $\xi \to +\infty$ and $\xi \to -\infty$.

After writing the system (2.7) in terms of the $\varphi_{i,\varepsilon}(\xi)$, we multiply both sides of each equation by appropriate factors, integrate over $\mathbb{R}$, and solve for the Fourier transforms of the $\varphi_{i,\varepsilon}$. Using the Fourier transform

$$\hat{\varphi}_{i,\varepsilon}(s) = \int_{-\infty}^{+\infty} e^{-is\xi}\varphi_{i,\varepsilon}(\xi)d\xi \tag{7.17}$$

we obtain the matrix equation

$$M(s - i\varepsilon)\begin{bmatrix} \hat{\varphi}_{1,\varepsilon}(s) \\ \hat{\varphi}_{2,\varepsilon}(s) \end{bmatrix} = \frac{1}{is + \varepsilon}N(s), \tag{7.18}$$

where M is a 2×2 matrix function with entries

$$\begin{aligned} M_{11}(s) &= B(s) - 2kD_1C(s) + k\alpha^2AE(s)B(s), \\ M_{12}(s) &= 2k\alpha D_2C(s) - k\alpha AE(-s)B(s), \\ M_{21}(s) &= 2k\alpha D_1C(s) - k\alpha AE(s)B(s), \\ M_{22}(s) &= B(s) - 2kD_2C(s) + k\alpha^2AE(-s)B(s), \end{aligned} \tag{7.19}$$

N is a vector with entries

$$\begin{aligned} N_1(s) &= 1 - k\alpha A[E(-s)e^{-is\beta} - \alpha E(s)], \\ N_2(s) &= e^{-is\beta} - k\alpha A[E(s) - \alpha E(-s)e^{-is\beta}], \end{aligned}$$

and

$$\begin{aligned} k &= \frac{1}{1-\alpha^2}, \\ C(s) &= \cos(s) - 1, \\ B(s) &= 1 - ics, \\ E(s) &= 1 - e^{is}. \end{aligned} \tag{7.20}$$

The result given in Lemma 2.1 allows us to invert the matrix M, and we arrive at the solution of the matrix equation (7.18):

$$\hat{\varphi}_{1,\varepsilon}(s) = \frac{M_{22}(s - i\varepsilon)N_1(s - i\varepsilon) - M_{12}(s - i\varepsilon)N_2(s - i\varepsilon)}{(is + \varepsilon)\det M(s - i\varepsilon)}, \tag{7.21}$$

$$\hat{\varphi}_{2,\varepsilon}(s) = \frac{M_{11}(s - i\varepsilon)N_2(s - i\varepsilon) - M_{21}(s - i\varepsilon)N_1(s - i\varepsilon)}{(is + \varepsilon)\det M(s - i\varepsilon)}. \tag{7.22}$$

Then using the Fourier inversion formula and the definition in (7.16), we have

(7.23)
$$\varphi_1(\xi) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{\xi(is+\varepsilon)}\hat{\varphi}_{i,\varepsilon}(s)ds = \frac{1}{2\pi i}\int_{-i\varepsilon-\infty}^{-i\varepsilon+\infty} \frac{e^{is\xi}}{s}\frac{M_{22}(s)N_1(s) - M_{12}(s)N_2(s)}{\det(M(s))},$$
$$\varphi_2(\xi) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{\xi(is+\varepsilon)}\hat{\varphi}_{i,\varepsilon}(s)ds = \frac{1}{2\pi i}\int_{-i\varepsilon-\infty}^{-i\varepsilon+\infty} \frac{e^{is\xi}}{s}\frac{M_{11}(s)N_2(s) - M_{21}(s)N_1(s)}{\det(M(s))}.$$

Applying the Cauchy integral formula, we shift the contour of integration to the real axis, with a semicircle around the origin. That is, we rewrite the integral in the form

$$\varphi_1(\xi) = \frac{1}{2\pi i}\left(\int_{-\infty}^{-\varepsilon} + \int_{\varepsilon}^{\infty} + \int_{C_\varepsilon}\right)\frac{e^{is\xi}}{s}\frac{M_{22}(s)N_1(s) - M_{12}(s)N_2(s)}{\det(M(s))}, \tag{7.24}$$

where C_ε is lower semicircle of radius ε around the origin. We may use a standard residue calculation for the third of these integrals, and the first two may be combined by using a change of variables in the first integral. After this, we use Euler's formula $\exp(ix) = \cos(x) + i\sin(x)$, and then we simplify by collecting sines and cosines according to their arguments. This leaves us with

$$\varphi_i(\xi) = \frac{1}{2} + \frac{1}{2\pi}\int_0^{\infty}\sum_{j\in M_i} F_{i,j}(s)\cos(s(\xi-\xi_j)) + G_{i,j}(s)\frac{1}{s}\sin(s(\xi-\xi_j))ds, \tag{7.25}$$

where $F_{i,j}$ and $G_{i,j}$ are as in (2.9).

8. Appendix B. We now present the proof of Lemma 2.1.

LEMMA 8.1. *Let $\alpha \in [0,1)$, $A \in [0,1)$, $\beta \in \mathbb{R}$, and $c \neq 0$. Then*

$$|\det M(s)|^2 \geq 1 \tag{8.1}$$

for $s \in \mathbb{R}$.

Proof. Using the definitions above, we may write $|\det M(s)|^2$ as

$$|\det M(s)|^2 = a^2 + b^2 = c^4 s^4 (b_2^2) + c^2 s^2 (2b_2^2 + b_1^2 + 2b_2(b_1 - b_0)) + (b_2 + b_1 + b_0)^2. \tag{8.2}$$

Clearly $b_2^2 \geq 0$ for $s \in \mathbb{R}$. Also,

$$b_2 + b_1 + b_0 = 1 - 2kC(s)[D_1 + D_2 - 2D_1D_2C(s) + \alpha^2 A(1-A)] \geq 1 \tag{8.3}$$

since $C(s) \leq 0$. It remains to show that the coefficient of c^2s^2 is nonnegative. Sorting by powers of $C(s)$, we have

$$\begin{aligned} 2b_2^2 + b_1^2 + 2b_2(b_1 - b_0) = {} & C^3(s)[16k^2 D_1 D_2 \alpha^2 A(1-A)] \\ & + C^2(s)4k\left(\left[k[2\alpha^2 A(1-A)(D_1 + D_2 + \alpha^2 A(1-A)) \right.\right. \\ & \qquad \left.\left. + (D_1 + D_2)^2] - 2D_1D_2\right]\right) \\ & + C(s)4k[-2\alpha^2 A(1-a) - (D_1 + D_2)] \\ & + 2, \end{aligned} \tag{8.4}$$

from which we see that the coefficient of $C^0(s)$ is positive and that of $C^1(s)$ is negative, as desired. It is convenient to express the order 2 and order 3 terms as

$$\begin{aligned} &C^2(s)\left(4k^2\left[2\alpha^2 A(1-A)(D_1+D_2+\alpha^2 A(1-A))\right]\right) \\ &+ C^2(s)\left(4k[(D_1+D_2)^2 k - 2kD_1D_2 + C(s)4kD_1D_2\alpha^2 A(1-A)]\right), \end{aligned} \tag{8.5}$$

where once again the terms on the top row are nonnegative. Finally, we have the estimate for the coefficient on the bottom

$$\begin{aligned} &k[D_1^2+D_2^2+4kD_1D_2C(s)\alpha^2A(1-A)] + 2D_1D_2(k-1) \\ &\geq k[D_1^2+D_2^2+4kD_1D_2C(s)\alpha^2A(1-A)] \\ &\geq D_1^2+D_2^2-2D_1D_2 \geq 0 \end{aligned} \tag{8.6}$$

since $-2 \leq C(s) \leq 0$, $\alpha < 1$, and $4A(1-A) \leq 1$ for $A \in [0,1)$. Hence the coefficient of c^2s^2 in $|\det M(s)|^2$ is greater than or equal to two. This estimate gives us the result stated above. ∎

9. Appendix C. We now prove the claim (5.8) by showing that $\sum_{n=0}^{\infty} E_n \to 0$ as $c \to 0$. Recall

$$\begin{aligned} |E_n| &= \left| \int_{np}^{(n+1)p} F_{i,j}(s)\cos(s(\beta_i - j))ds - \int_{np}^{(n+1)p} F_{i,j}(np,s)\cos(s(\beta_i-\xi_j))ds \right| \\ &\leq \left| \int_{np}^{(n+1)p} [F_{i,j}(s) - F_{i,j}(np,s)]ds \right|. \end{aligned} \tag{9.1}$$

Also, since $|R_i(s)| \leq 2$ and all other terms in the $F_{i,j}$ are constant, it will suffice to look at

$$\left| \int_{np}^{(n+1)p} [W(s) - W(np,s)]ds \right| \quad \text{and} \quad \left| \int_{np}^{(n+1)p} [Y(s) - Y(np,s)]ds \right|. \tag{9.2}$$

The following applies to both cases. First, make a change of variables $s \to \frac{s}{c}$ in both terms of the difference. Also, the differences $W(s) - W(np,s)$ and $Y(s) - Y(np,s)$ may be expressed in the form

$$\frac{\gamma_2(s)s^2 + \gamma_0(s)}{\delta_4(s)s^4 + \delta_2(s)s^2 + \delta_0(s)} - \frac{\gamma_2(s)(npc)^2 + \gamma_0(s)}{\delta_4(s)(npc)^4 + \delta_2(s)(npc)^2 + \delta_0(s)}, \tag{9.3}$$

where the γ and δ are functions of s, but we will suppress this dependence to make the notation lighter. Then for $n \geq 1$,

(9.4)

$$
\begin{aligned}
|E_n| &\le \int_{npc}^{(n+1)pc} ((npc)^2 - s^2) \frac{\gamma_2\delta_4(npc)^2 s^2 + \gamma_0\delta_4((npc)^2 + s^2) + \gamma_0\delta_2 - \gamma_2\delta_0}{(\delta_4 s^4 + \delta_2 s^2 + \delta_0)(\delta_4(npc)^4 + \delta_2(npc)^2 + \delta_0)} ds \\
&\le \int_{npc}^{(n+1)pc} (2n+1)(pc)^2 \frac{\gamma_2\delta_4(npc)^2 s^2 + \gamma_0\delta_4((npc)^2 + s^2) + \gamma_0\delta_2 - \gamma_2\delta_0}{(\delta_4 s^4 + \delta_2 s^2 + \delta_0)(\delta_4(npc)^4 + \delta_2(npc)^2 + \delta_0)} ds \\
&\le \int_{npc}^{(n+1)pc} (2n+1)(pc)^2 \frac{\gamma_2\delta_4(npc)^2 s^2 + \gamma_0\delta_4((npc)^2 + s^2) + \gamma_0\delta_2 - \gamma_2\delta_0}{(\delta_4 s^4 + \delta_2 s^2 + \delta_0)(\frac{\delta_4}{16} s^4 + \frac{\delta_2}{4} s^2 + \delta_0)} ds \\
&\le 16 \int_{npc}^{(n+1)pc} (2n+1)(pc)^2 \frac{\gamma_2\delta_4(npc)^2 s^2 + \gamma_0\delta_4((npc)^2 + s^2) + \gamma_0\delta_2 - \gamma_2\delta_0}{(s^2+1)^4} ds \\
&\le 16 \int_{npc}^{(n+1)pc} (2n+1)(pc)^2 \frac{\gamma_2\delta_4 s^4 + 2\gamma_0\delta_4 s^2 + \gamma_0\delta_2 - \gamma_2\delta_0}{(s^2+1)^4} ds \\
&\le 16 \int_{npc}^{(n+1)pc} 3spc \frac{\gamma_2\delta_4 s^4 + 2\gamma_0\delta_4 s^2 + \gamma_0\delta_2 - \gamma_2\delta_0}{(s^2+1)^4} ds \\
&\le 48pc \int_{npc}^{(n+1)pc} Ks \frac{s^4 + 2s^2 + 1}{(s^2+1)^4} ds \\
&\le 48pcK \int_{npc}^{(n+1)pc} \frac{s}{(s^2+1)^2} ds,
\end{aligned}
$$

where K is a constant. Similarly,

$$
\begin{aligned}
|E_0| &\le (pc)^2 \int_0^{pc} \frac{\gamma_0\delta_4(pc)^2 + \gamma_0\delta_2 - \gamma_2\delta_0}{(\delta_4 s^4 + \delta_2 s^2 + \delta_0)\delta_0} ds \\
&\le K'(pc)^2 \int_0^{pc} \frac{1}{(s^1+1)^2} ds \to 0 \quad \text{as} \quad c \to 0
\end{aligned} \tag{9.5}
$$

where K' is a constant. Using these estimates, we see that

$$
\sum_{n=0}^{\infty} |E_n| \le |E_0| + 48pcK \int_{pc}^{\infty} \frac{s}{(s^2+1)^2} ds \to 0 \quad \text{as} \quad c \to 0, \tag{9.6}
$$

which justifies our claim in (5.8).

Acknowledgments. The authors are grateful to Chris Elmer and Weishi Liu for discussions related to this work and to the referees for helpful comments on an earlier version of this paper.

REFERENCES

[1] J. M. Bilbault, V. B. Kazantsev, P. Marquie, S. Morfu, and V. I. Nekorkin, *Theoretical and experimental study of two discrete coupled Nagumo chains*, Phys. Rev. E, 64 (2001), 036602.

[2] S. Binczak, J. C. Eilbeck, and A. C. Scott, *Ephaptic coupling of myelinated nerve fibers*, Phys. D, 148 (2001), pp. 159–174.

[3] H. Bokil, N. Laaris, K. Blinder, M. Ennis, and A. Keller, *Ephaptic interactions in the mammalian olfactory system*, J. Neurosci., 2001, 21:RC173(1–5).

[4] A. Bose, *Symmetric and antisymmetric pulses in parallel coupled nerve fibres*, SIAM J. Appl. Math., 55 (1995), pp. 1650–1674.

[5] A. Bose and C. K. R. T. Jones, *Stability of the in-phase travelling wave solution in a pair of coupled nerve fibers*, Indiana Univ. Math. J., 44 (1995), pp. 189–220.
[6] J. W. Cahn, J. Mallet-Paret, and E. S. Van Vleck, *Traveling wave solutions for systems of ODEs on a two-dimensional spatial lattice*, SIAM J. Appl. Math., 59 (1998), pp. 455–493.
[7] S. N. Chow, J. Mallet-Paret, W. Shen, *Traveling waves in lattice dynamical systems*, J. Differential Equations, 149 (1998), pp. 248–291.
[8] C. E. Elmer and E. S. Van Vleck, *Traveling wave solutions for bistable differential-difference equations with periodic diffusion*, SIAM J. Appl. Math., 61 (2001), pp. 1648–1679.
[9] C. E. Elmer and E. S. Van Vleck, *Spatially discrete Fitzhugh-Nagumo equations*, SIAM J. Appl. Math., 65 (2005), pp. 1153–1174.
[10] T. Erneux and G. Nicolis, *Propagating waves in discrete bistable reaction-diffusion systems*, Phys. D, 67 (1993), pp. 237–244.
[11] G. Fath, *Propagation failure of traveling waves in discrete bistable medium*, Phys. D, 116 (1998), pp. 176–190.
[12] J. A. Feroe, *Existence and stability of multiple impulse solutions of a nerve equation*, SIAM J. Appl. Math., 42 (1982), pp. 235–246.
[13] J. P. Keener, *Frequency dependent decoupling of parallel excitable fibers*, SIAM J. Appl. Math., 49 (1989), pp. 210–230.
[14] R. S. MacKay and J.-A. Sepulchre, *Multistability in networks of weakly coupled bistable units*, Phys. D, 82 (1995), pp. 243–254.
[15] H. McKean, *Nagumo's Equation*, Adv. Math., 4 (1970), pp. 209–223.
[16] S. Reutskiy, E. Rossoni, and B. Tirozzi, *Conduction in bundles of demyelinated nerve fibers: Computer simulation*, Biol. Cybern., 89 (2003), pp. 439–448.
[17] J. Rinzel and J. B. Keller, *Traveling wave solutions of a nerve conduction equation*, Biophys. J., 13 (1973), pp. 1313–1337.
[18] A. Scott, *Neuroscience: A Mathematical Primer*, Springer, New York, 2002.
[19] L. F. Shampine, R. C. Allen, S. Pruess, *Fundamentals of Numerical Computing*, Wiley, New York, 1997.
[20] N. Sperelakis, *Combined electric field and gap junctions on propagation of action potentials in cardiac muscle and smooth muscle in PSpice simulation*, J. Electrocard., 36 (2003), pp. 279–293.
[21] N. Sperelakis, *An electric field mechanism for transmission of excitation between myocardial cells*, Circulation Res., 91 (2002), pp. 985–987.
[22] W.-P. Wang, *Multiple impulse solutions to McKean's caricature of the nerve equation.* I *Existence*, Comm. Pure Appl. Math., 41 (1988), pp. 71–103.
[23] W.-P. Wang, *Multiple impulse solutions to McKean's caricature of the nerve equation.* II *Stability*, Comm. Pure Appl. Math., 41 (1988), pp. 997–1025.

SIAM J. APPL. MATH.
Vol. 66, No. 3, pp. 977–1001

QUALITATIVE ASPECTS IN DUAL-PHASE-LAG THERMOELASTICITY*

RAMÓN QUINTANILLA† AND REINHARD RACKE‡

Abstract. We consider the system of dual-phase-lag thermoelasticity proposed by Chandrasekharaiah and Tzou. First, we prove that the solutions of the problem are generated by a semigroup of quasi-contractions. Thus, the problem of the third order in time is well-posed. Then the exponential stability is investigated. Finally the spatial behavior of solutions is analyzed in a semi-infinite cylinder and a result on the domain of influence is obtained.

Key words. hyperbolic model in thermoelasticity, well-posedness, spatial evolution in a semi-infinite cylinder, exponential stability

AMS subject classifications. 35L35, 74F05, 74G50

DOI. 10.1137/05062860X

1. Introduction.

It is well known that the usual theory of heat conduction based on Fourier's law predicts infinite heat propagation speed. Heat transmission at low temperature has been observed to propagate by means of waves. These aspects have caused intense activity in the field of heat propagation. Extensive reviews on the so-called second sound theories (hyperbolic heat conduction) are given in Chandrasekharaiah [3] and in the books of Müller and Ruggeri [20] and Jou, Casas-Vazquez, and Lebon [18]. A theory of heat conduction in which the evolution equation contains a third-order derivative with respect to time was proposed in [8]. Several instability results have been obtained for the theory (see, e.g., [7, 23]) as well as proof of the nonexistence of global solutions in the nonlinear theory [29].

In 1995, Tzou [34] proposed a theory of heat conduction in which the Fourier law is replaced by an approximation of the equation

$$\mathbf{q}(\mathbf{x}, t+\tau_q) = -k\nabla\theta(\mathbf{x}, t+\tau_\theta), \quad \tau_q > 0,\ \tau_\theta > 0, \tag{1.1}$$

where τ_q is the phase-lag of the heat flux and τ_θ is the phase-lag of the gradient of temperature. The relation (1.1) states that the gradient of temperature at a point in the material at time $t+\tau_\theta$ corresponds to the heat flux vector at the same point at time $t+\tau_q$. The delay time τ_θ is caused by microstructural interactions such as phonon scattering or phonon-electron interactions. The delay τ_q is interpreted as the relaxation time due to fast-transient effects of thermal inertia. The thermoelastic model was proposed in [3],

$$\mu u_{i,jj} + (\lambda+\mu)u_{j,ji} - m\theta_{,i} = \rho\ddot{u}_i, \tag{1.2}$$

$$-q_{i,i} - m\theta_0\dot{u}_{i,i} = c\dot{\theta}, \tag{1.3}$$

$$q_i(\cdot, t+\tau_q) = -k\theta_{,i}(\cdot, t+\tau_\theta), \tag{1.4}$$

*Received by the editors April 6, 2005; accepted for publication (in revised form) October 10, 2005; published electronically March 3, 2006.
http://www.siam.org/journals/siap/66-3/62860.html
†Department of Applied Mathematics II, UPC Terrassa, Colom 11, 08222 Terrassa, Spain (ramon.quintanilla@upc.edu). This author was supported by project BFM/FEDER 2003-00309.
‡Department of Mathematics and Statistics, University of Konstanz, 78457 Konstanz, Germany (reinhard.racke@uni-konstanz.de). This author was supported by DFG project RA 504/3-1.

where ρ, θ_0, c are positive constants. $\mu > 0$ and λ are the Lamé moduli satisfying $\mu^* > 0$, where μ^* is defined in (2.16) depending on the space dimension.

Here and in what follows we use the Einstein summation convention with indices in the range $1 \dots n$, where $n = 1, 2, 3$ denotes the space dimension. Instead of Fourier's law, being equivalent to assuming

$$\tau_q = \tau_\theta = 0 \tag{1.5}$$

and leading to the classical hyperbolic-parabolic system of thermoelasticity together with the physical paradoxon of infinite propagation speed through the heat conduction part, we consider the model proposed by Chandrasekharaiah [3] and Tzou [34], where

$$\tau_q > 0, \quad \tau_\theta > 0$$

are positive relaxation times and where a second-order approximation for $\mathbf{q}$ and a first-order approximation for θ are used, turning (1.4) into

$$q_i + \tau_q \dot{q}_i + \frac{\tau_q^2}{2} \ddot{q}_i = -k\theta_{,i} - k\tau_\theta \dot{\theta}_{,i}. \tag{1.6}$$

Thus, in this paper, we consider the theory developed by taking a Taylor series expansion on both sides of (1.1) and retaining terms up to the second order in τ_q, but only to the first order in τ_θ. The model that we consider here involves a system of two coupled partial differential equations. It is of hyperbolic type (section 3). One of them is the usual second order in time equation of motion in the major part of thermoelastic systems and the other has a third-order derivative with respect to time. This system of equations has not received much attention in the literature (until now), but Hetnarski and Ignaczak consider it within the nonclassical approach of thermoelasticity in their review [10]. However, we can recall several references in the case that we do not consider mechanical deformations [15, 25]. It is known that when $\tau_\theta = 0$, solutions of heat conduction are not determined by means of a semigroup [9, p. 125]. However in [25], it was established that whenever $\tau_\theta > 0$, one can obtain solutions by means of a semigroup. Thus, the term $\tau_\theta \triangle \dot{\theta}$ plays a role in the stabilization for the equation. In this paper we extend some of the results on existence and stability obtained for the heat conduction to the thermoelastic problem.

The case $\tau_\theta > 0$ but $\tau_q = 0$, also leading to a hyperbolic system, the system of Lord and Shulman, has been studied before, and, for example, the exponential stability has been obtained for bounded reference configurations as well as the nonlinear stability near the equilibrium; see [31, 32].

A natural question is the determination of the time parameters τ_q and τ_θ (see [10]) and our work is motivated by this question. One might expect that mathematical analysis of existence, uniqueness, and stability issues, for example, would furnish certain restrictions on the parameters. One condition to be satisfied by solutions of a heat equation should be exponential stability (or at least stability). In [25], exponential stability (for the heat conduction) was established whenever

$$\tau_\theta > \tau_q/2. \tag{1.7}$$

We also recall that in [10] Hetnarski and Ignaczak asked (p. 474) for a *general domain of influence theorem as well as a principle of Saint-Venant's type* for this theory. We note that results of this kind were obtained in [15] for heat conduction. In this paper

we also extend some of the results concerning the time asymptotic and the spatial behavior obtained for heat conduction in order to include mechanical deformations.

Thus, under condition (1.7), one has a heat theory with a third-order derivative in time in the equation that predicts stability. This is of interest in the light of the results obtained in the theory proposed in [8]. By means of several exact solutions instability of solutions was also established in [25] whenever the condition (1.7) is violated. Thus, one may assume that the condition (1.7) must be satisfied in order to use this model to describe heat transmission. In fact one of the objects of this paper is to extend stability results to the thermoelastic problem. In [26] it was demonstrated for a bounded interval $(0, L) \subset \mathbb{R}$ that for the boundary conditions

$$u = \theta_x = 0 \tag{1.8}$$

which allows a nice series expansion of the solutions into $\sin(nx), \cos(nx)$ terms, exponential stability is to be expected since the relevant spectrum of the associated stationary operator lies strictly in the right-half complex plane.

We shall investigate here the more complicated boundary conditions

$$u = 0, \quad \theta = 0 \tag{1.9}$$

and prove the exponential stability of the associated semigroup.

In this paper we study three kinds of questions. One is to determine the suitable frame where the third-order problem of thermoelasticity of Chandrasekharaiah and Tzou type is well-posed and where the solutions are stable. The second is to prove the exponential stability for bounded reference configurations, and the third is to determine the spatial behavior of the solutions of the thermoelasticity in a semi-infinite cylinder in $\mathbb{R}^3$.

This paper is organized as follows. In section 2 we set down the field equations and the boundary and initial conditions of the problem we consider in this paper. A uniqueness and existence result is proved in section 3. In section 4 we prove the exponential stability for bounded reference configurations. In section 5, we obtain some results of Saint-Venant's type concerning the spatial behavior of solutions in a semi-infinte cylinder and some consequences of them as obtained in section 6. The last section is devoted to the study of the spatial behavior of solutions of a nonstandard problem.

When we study the spatial behavior of solutions of some problems concerning the dual-phase-lag thermoelastic system we shall denote the three-dimensional semi-infinite cylinder R with cross section D. The finite end face of the cylinder is in the plane $x_3 = 0$. The boundary ∂D is supposed regular enough to allow the use of the divergence theorem. We denote by $R(z)$ the set of points of the cylinder R such that x_3 is greater than z and by $D(z)$ the cross section of the points such that $x_3 = z$. The spatial evolution with distance from the end for solutions of elliptic equations is relevant to the study of Saint-Venant's principle in continuum mechanics (see, e.g., [6, 11, 12, 13] for reviews of this work). Such results for parabolic equations have also been obtained (see [6, 11, 12, 13, 14]) and more recently for hyperbolic equations (see [2] and the references cited therein).

2. Preliminaries. We consider the homogeneous isotropic case. In this paper we study solutions $(\mathbf{u}, \theta) = (\mathbf{u}(\mathbf{x}, t), \theta(\mathbf{x}, t))$ of the thermoelastic system for the Chandrasekharaiah–Tzou theory. The equations are

$$\mu u_{i,jj} + (\lambda + \mu) u_{j,ji} - m\theta_{,i} = \rho \ddot{u}_i, \tag{2.1}$$

$$k\hat{\theta}_{,ii} - m\theta_0\dot{\tilde{u}}_{i,i} = c\dot{\tilde{\theta}}. \tag{2.2}$$

We have used the notation

$$\hat{f} = f + \tau_\theta \dot{f}, \;\; \tilde{f} = f + \tau_q \dot{f} + \frac{\tau_q^2}{2}\ddot{f}, \tag{2.3}$$

where $\tau_\theta > 0$, $\tau_q > 0$ are the dimensionless time lag parameters.

We study the qualitative behavior of classical solutions subject to the initial conditions

$$u_i(\mathbf{x},0) = u_i^0(\mathbf{x}), \;\; \dot{u}_i(\mathbf{x},0) = v_i^0(\mathbf{x}), \theta(\mathbf{x},0) = \theta^0(\mathbf{x}), \;\; \dot{\theta}(\mathbf{x},0) = \vartheta^0(\mathbf{x}), \;\; \ddot{\theta}(\mathbf{x},0) = \phi^0(\mathbf{x}) \tag{2.4}$$

and the boundary conditions

$$u_i(\mathbf{x},t) = \theta(\mathbf{x},t) = 0, \;\; \mathbf{x} \in \partial D \;\times\; [0,\infty), \tag{2.5}$$

$$u_i(x_1, x_2\,, 0,\; t) \;= f_i(x_\alpha\,, t) \;\; \theta(x_1, x_2\,, 0,\; t) \;= g(x_\alpha\,, t) \quad \text{on } D(0) \;\times\; [0,\infty), \tag{2.6}$$

where the prescribed boundary data f_i, g on the end $x_3 = 0$ is such that $f_i(x_\alpha, 0) = u_i^0(x_\alpha)$, $g(x_\alpha, 0) = \theta^0(x_\alpha)$, and f_i, g are assumed to vanish on $\partial D(0) \times [0,\infty)$. We will not make any a priori assumption regarding the behavior of solutions as $x_3 \longrightarrow \infty$.

Observe that in the limit as τ_θ and $\tau_q \to 0$, we recover from (2.1) or (2.2) the usual thermoelastic system where, in this limiting case, only the three first of (2.4) are assumed to hold. In this limit the existence, stability and the spatial evolution of solutions have been studied in a variety of contexts (see, e.g., [17, 24, 4] and the references cited therein). When τ_q and τ_θ are positive, the results to be described in what follows will be seen to be similar to those obtained previously for such equations (see, e.g., [2] and the references cited therein).

In the course of our calculations, we will use the fact that the eigenvalues of the real symmetric positive definite matrix

$$\begin{pmatrix} a & b \\ b & l \end{pmatrix} \tag{2.7}$$

are

$$\lambda^{\pm} = \frac{1}{2}\Big(a + l \pm \sqrt{(a-l)^2 + 4b^2}\Big), \tag{2.8}$$

so that the smallest eigenvalue is

$$\lambda^{-} = \frac{1}{2}\Big(a + l - \sqrt{(a-l)^2 + 4b^2}\Big). \tag{2.9}$$

We will use (2.7) in two particular cases. When

$$a = \tau_q + \tau_\theta, \;\; b = \frac{\tau_q^2}{2}, \;\; l = \frac{\tau_q^2 \tau_\theta}{2}, \tag{2.10}$$

it can be easily verified using (1.7) that the matrix (2.7) is indeed positive definite and so its smallest positive eigenvalue, denoted by λ_0, is given by

$$\lambda_0 = \frac{1}{2}\left(\tau_q + \tau_\theta + \frac{1}{2}\tau_q^2\tau_\theta - \sqrt{\tau_q^4 + \tau_q^2 + \tau_\theta^2 + \frac{1}{4}\tau_q^4\tau_\theta^2 + 2\tau_q\tau_\theta - \tau_\theta^2\tau_q^2 - \tau_q^3\tau_\theta}\right). \tag{2.11}$$

When

$$a = \frac{2}{\gamma} + (\tau_q + \tau_\theta), \quad b = \frac{1}{2}\tau_q^2 \ , \quad l = \frac{1}{2}\tau_q^2\tau_\theta + \frac{2}{\gamma}\left(\tau_\theta\tau_q - \frac{1}{2}\tau_q^2\right), \quad \gamma > 0, \tag{2.12}$$

the matrix (2.9) is again positive definite with the smallest eigenvalue, denoted by μ_γ, given by

$$\mu_\gamma = \frac{1}{2\gamma}\left(2 + \gamma(\tau_q + \tau_\theta) + \frac{\gamma}{2}\tau_q^2\tau_\theta + 2\left(\tau_\theta\tau_q - \frac{1}{2}\tau_q^2\right)\right.$$
$$\left. -\sqrt{\left[2 + \gamma(\tau_q + \tau_\theta) - \frac{\gamma}{2}\tau_q^2\tau_\theta - 2\left(\tau_\theta\tau_q - \frac{1}{2}\tau_q^2\right)\right]^2 + \gamma^2\tau_q^4}\right). \tag{2.13}$$

To be used later, it will be worth using the following notation:

$$T_{ij} = \mu u_{i,j} + (\lambda + \mu)\delta_{ij}u_{r,r} - m\delta_{ij}\theta. \tag{2.14}$$

We have that the estimate

$$T_{ji}T_{ji} \le (1 + \epsilon)\mu^*[\mu u_{i,j}u_{i,j} + (\lambda + \mu)u_{r,r}u_{s,s}] + 3m^2(1 + \epsilon^{-1})\theta^2 \tag{2.15}$$

is satisfied, for every positive ϵ, where

$$\mu^* = \begin{cases} 2\mu + \lambda, & n = 1, \\ \max\{\mu, 2\lambda + 3\mu\}, & n = 2, \\ \max\{\mu, 3\lambda + 4\mu\}, & n = 3. \end{cases} \tag{2.16}$$

μ^* is the maximal positive eigenvalue of the quadratic form [4]

$$Q(\zeta) := \mu\zeta_{ij}\zeta_{ij} + (\lambda + \mu)\zeta_{rr}\zeta_{ss}.$$

When we study the qualitative aspects concerning existence, uniqueness, and exponential stability, and without loss of generality, we assume $\rho = c = 1$. However, when we study the spatial behavior of solutions we relax this condition to assume that mass density and thermal capacity are positive because we wish to demonstrate the dependence of the decay parameters on ρ, c explicitly.

3. Well-posedness. We shall formulate the problem for the semi-infinite cylinder R in three space dimensions, but the well-posedness holds for general domains; see the remarks following Theorem 3.3.

The well-posedness result for the third order in time system can be achieved by an appropriately sophisticated choice of variables and spaces which reflect the special structure of the system.

We first transform the system (2.1)–(2.6) to zero boundary conditions on all of ∂R by defining

(3.1)
$$v_i(x_\alpha, 0, t) := u_i(x_\alpha, 0.t) - f_i(x_\alpha, t), \qquad v_i(x_\alpha, x_3, t) := u_i(x_\alpha, x_3.t) \quad \text{for } x_3 > 0,$$

(3.2)
$$\psi(x_\alpha, 0, t) := \theta(x_\alpha, 0, t) - g(x_\alpha, t), \qquad \psi(x_\alpha, x_3, t) := \theta(x_\alpha, x_3, t) \quad \text{for } x_3 > 0,$$

and using (u_i, θ) instead of (v_i, ψ) again, we obtain the initial boundary value problem

$$\mu u_{i,jj} + (\lambda+\mu)u_{j,ji} - m\theta_{,i} = \ddot{u}_i - h_i, \tag{3.3}$$

$$k\hat{\theta}_{,ii} - m\theta_0 \dot{\tilde{u}}_{i,i} = \dot{\tilde{\theta}} - p, \tag{3.4}$$

(3.5)
$$u_i(\mathbf{x},0) = u_i^0(\mathbf{x}),\ \dot{u}_i(\mathbf{x},0) = v_i^0(\mathbf{x}), \theta(\mathbf{x},0) = \theta^0(\mathbf{x}),\ \dot{\theta}(\mathbf{x},0) = \vartheta^0(\mathbf{x}),\ \ddot{\theta}(\mathbf{x},0) = \phi^0(\mathbf{x}),$$

$$u_i(\mathbf{x},t) = \theta(\mathbf{x},t) = 0,\ \ \mathbf{x} \in \partial R\ \times\ [0,\infty), \tag{3.6}$$

where the given external force $\mathbf{h}$ and heat supply p arise from the transformation (3.1), (3.2) in terms of the boundary data $\mathbf{f}$ and g, respectively.

For the transformation to a first-order system that finally will be characterized by a semigroup, we apply the differential operator $\tilde{}$ from (2.3) to the differential equation (3.3) and obtain

$$\mu\tilde{u}_{i,jj} + (\lambda+\mu)\tilde{u}_{j,ji} - m\tilde{\theta}_{,i} = \ddot{\tilde{u}}_i + \tilde{h}_i. \tag{3.7}$$

We remark that finding a solution $(\tilde{\mathbf{u}}, \theta)$ allows to determine the desired solutions $(\mathbf{u}, \theta)$ of the original system.

Defining

$$\mathbf{V} := (\tilde{\mathbf{u}}, \tilde{\mathbf{u}}_t, \theta, \theta_t, \theta_{tt})'$$

we obtain

$$\mathbf{V}_t = A\mathbf{V} + \mathbf{F}, \qquad V(0) = V^0 \tag{3.8}$$

with the (yet formal) differential operator A given by the symbol

$$A_f := \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ \mu\Delta + (\lambda+\mu)\nabla\nabla' & 0 & -m\nabla & -\tau_q m\nabla & -\dfrac{\tau_q^2 m}{2}\nabla \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & -\dfrac{2m\theta_0}{\tau_q^2}\nabla' & \dfrac{2k}{\tau_q^2}\Delta & \dfrac{2}{\tau_q^2}(k\tau_\theta\Delta - 1) & -\dfrac{2}{\tau_q} \end{pmatrix},$$

the right-hand side $\mathbf{F}$ given by

$$\mathbf{F} := (0, \mathbf{h}, 0, 0, 0, p)',$$

and the initial value

$$\mathbf{V}^0(\mathbf{x}) := (\tilde{\mathbf{u}}, \tilde{\mathbf{u}}_t, \theta, \theta_t, \theta_{tt})'(\mathbf{x}, 0)$$

with its components being given in terms of the originally prescribed initial data in (3.5) by using the differential equations.

As underlying Hilbert space we choose

$$\mathcal{H} := (H_0^1(R))^n \times (L^2(R))^n \times H_0^1(R) \times H_0^1(R) \times L^2(R)$$

with inner product

$$\langle V, W, \rangle_{\mathcal{H}} := \frac{4}{\tau_q^4}\left(\langle \theta_0 V^2, W^2\rangle + \langle \theta_0 \mu \nabla V^1, \nabla W^1\rangle + \langle \theta_0(\lambda+\mu)\nabla' V^1, \nabla' W^1\rangle\right)$$
$$+ \left\langle \frac{2}{\tau_q^2} V^4, W^4 \right\rangle + \left\langle \frac{2\tau_\theta k}{\tau_q^2}\nabla V^4, \nabla W^4\right\rangle + \langle V^5, W^5\rangle + \left\langle \frac{2k}{\tau_q^2}\nabla V^3, \nabla W^4\right\rangle$$
$$+ \left\langle \frac{2k}{\tau_q^2}\nabla V^4, \nabla W^3\right\rangle + b_0\langle \nabla V^3, \nabla W^3\rangle,$$

where $\langle\cdot,\cdot\rangle$ denotes the usual $L^2(R)$-inner product and where b_0 is chosen appropriately large in dependence of the coefficients to ensure that the bilinear form $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is positive definite. The operator A is now given as

$$A : D(A) \subset \mathcal{H} \mapsto \mathcal{H}, \qquad AV := A_f V$$

with

$$D(A) := \{V \in \mathcal{H} \mid V^2 \in H_0^1(R)^n,\, V^5 \in H_0^1(R),\, A_f V \in \mathcal{H}\}.$$

The choice of the inner product is special, of course, and extends similar considerations from [25] for the pure heat conduction problem.

LEMMA 3.1. *There exists a constant $c_1 > 0$ such that for all $V \in D(A)$*

$$|\langle AV, V\rangle_{\mathcal{H}}| \le c_1 \|V\|_{\mathcal{H}}^2$$

holds.

Proof. We have

$$\langle AV, V\rangle_{\mathcal{H}} = -\frac{4m\theta_0}{\tau_q^4}\langle \nabla V^3, V^2\rangle - \frac{4m\theta_0}{\tau_q^3}\langle \nabla V^4, V^2\rangle - \frac{2}{\tau_q}\langle V^5, V^5\rangle$$
$$+\frac{2k}{\tau_q^2}\langle \nabla V^4, \nabla V^4\rangle + b_0\langle \nabla V^4, \nabla V^3\rangle,$$

which implies the assertion. ∎

As a consequence we see that for $d > c_1$ the operator $A - d$ is dissipative and invertible.

LEMMA 3.2. *For all $d > c_1$ we have that the range of $A - d$ is all of $\mathcal{H}$.*

Proof. The solvability of $(A-d)V = F$ is equivalent to solving

$$V^2 - dV^1 = F^1, \tag{3.9}$$

$$\mu\Delta V^1 + (\lambda+\mu)\nabla\nabla' V^1 - m\nabla V^3 - \tau_q m\nabla V^4 - \frac{\tau_q^2 m}{2}\nabla V^5 - dV^2 = F^2, \tag{3.10}$$

$$V^4 - dV^3 = F^3, \tag{3.11}$$

$$V^5 - dV^4 = F^4, \tag{3.12}$$

$$-\frac{2m\theta_0}{\tau_q^2}\nabla' V^2 + \frac{2k}{\tau_q^2}\Delta V^3 + \frac{2}{\tau_q^2}(k\tau_\theta\Delta - 1)V^4 - \frac{2}{\tau_q}V^5 - dV^5 = F^5. \tag{3.13}$$

Eliminating V^2, V^4, and V^5 and using

$$E := -\mu\Delta - (\lambda+\mu)\nabla\nabla'$$

we have to solve

$$-EV^1 - d^2V^1 - \underbrace{\left(m + \tau_q md + \frac{\tau_q^2 md^2}{2}\right)}_{=:\alpha_1}\nabla V^3 = F^2 + dF^1 + \left(\tau_q m + \frac{\tau_q^2 md}{2}\right)\nabla F^3$$

$$+ \frac{\tau_q^2 m}{2}\nabla F^4, \tag{3.14}$$

$$-\underbrace{\left(\frac{2k}{\tau_q^2} + \frac{2k\tau_\theta d}{\tau_q^2}\right)}_{=:\gamma_1}\Delta V^3 + \underbrace{\left(\frac{2d}{\tau_q^2} + \frac{2d^2}{\tau_q} + d^3\right)}_{=:\delta_1}V^3 + \underbrace{\frac{2m\theta_0 d}{\tau_q^2}}_{=:\beta_1}\nabla' V^1$$

$$= -F^5 - \frac{2m\theta_0}{\tau_q^2}\nabla' F^1 - \frac{2}{\tau_q^2}F^3 - \left(\frac{2}{\tau_q} + d\right)(dF^3 + F^4) + \frac{2k\tau_\theta}{\tau_q^2}\Delta F^3. \tag{3.15}$$

Hence we consider for $G^1 \in L^2(R)^3$ and $G^2 \in H^{-1}$, the dual space to $H_0^1(R)$, the system

$$EV^1 + d^2V^1 + \alpha_1\nabla V^3 = G^1, \tag{3.16}$$

$$-\gamma_1\Delta V^3 + \delta_1 V^3 + \beta_1\nabla' V^1 = G^2, \tag{3.17}$$

where $\alpha_1, \beta_1, \gamma_1, \delta_1$ are positive. If $(V^1, V^3) \in (H_0^1(R))^n \times H_0^1(R)$ solve (3.16), (3.17), then V^2, V^4, and V^5 can be determined from the equations (3.9), (3.11), and (3.12), respectively, and $V \in D(A)$ will solve $(A - d)V = F$.

$E + d^2$ can be regarded as a positive self-adjoint operator, the inverse of which maps $L^2(R)^3 \mapsto (H^2(R) \cap H_0^1(R))^n$, and hence V^1 should satisfy

$$V^1 = (E + d^2)^{-1}(G^1 - \alpha_1\nabla V^3).$$

Plugging this into (3.17) it remains to determine V^3 as a solution in $H_0^1(R)$ of

$$-\gamma_1\Delta V^3 + \delta_1 V^3 - \alpha_1\beta_1\nabla'(E + d^2)^{-1}\nabla V^3 = G^2 - \beta_1\nabla'(E + d^2)^{-1}G^1. \tag{3.18}$$

But (3.18) can be solved easily because the bilinear form

$$B(g, h) := \gamma_1\langle\nabla g, \nabla h\rangle + \delta_1\langle g, h\rangle + \alpha_1\beta_1\langle(E + d^2)^{-1/2}\nabla g, (E + d^2)^{-1/2}\nabla h\rangle$$

is positive on $H_0^1(R)$, and hence the Lax and Milgram lemma yields the solvability of (3.18) for any right-hand side in H^{-1}. This proves the assertion of the lemma. □

Now we conclude from the last two lemmas that A generates a C_0-semigroup, and hence the initial (boundary) value problem (3.8) is uniquely solvable.

THEOREM 3.3. *For any $F \in C^0([0, \infty), D(A))$ or $F \in C^1([0, \infty), \mathcal{H})$ and any $V^0 \in D(A)$ there is a unique solution V to* (3.8) *with $V \in C^1([0, \infty), \mathcal{H}) \cap C^0([0, \infty), D(A))$.*

The well-posedness consideration in this section extends naturally to other domains $\Omega \subset \mathbb{R}^n$, $n = 1, 2, 3$, instead of the three-dimensional cylinder R, e.g., literally to smoothly bounded domains and to convex domains (where elliptic H^2-regularity up to the boundary holds).

The system under consideration is of hyperbolic type, as we shall demonstrate in the one-dimensional case. Here the differential equations (3.7), (3.4) turn into

$$\tilde{u}_{tt} = \alpha_* \tilde{u}_{xx} - \frac{\tau_q^2 m}{2}\theta_{ttx} - \tau_q m \theta_{tx} - m\theta_x, \tag{3.19}$$

$$\theta_{ttt} = -\frac{2}{\tau_q}\theta_{tt} - \frac{2}{\tau_q^2}\theta_t - \frac{2m\theta_0}{\tau_q^2}\tilde{u}_{tx} + \frac{2\tau_\theta k}{\tau_q^2}\theta_{txx} + \frac{2k}{\tau_q^2}\theta_{xx}, \tag{3.20}$$

where $\alpha_* := 2\mu + \lambda$ and the right-hand sides are assumed to be zero.

Defining

$$\mathbf{W} := (\tilde{u}_x, \tilde{u}_t, \theta_x, \theta_t, \theta_{tx}, \theta_{tt})'$$

we obtain

$$\mathbf{W}_t = B\mathbf{W}_x + D\mathbf{W},$$

where

$$B := \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ \alpha_* & 0 & 0 & 0 & 0 & -\frac{\tau_q^2 m}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & -\frac{2m\theta_0}{\tau_q^2} & \frac{2k}{\tau_q^2} & 0 & \frac{2\tau_\theta k}{\tau_q^2} & 0 \end{pmatrix}$$

and

$$D := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -m & 0 & -\tau_q m & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{2}{\tau_q^2} & 0 & -\frac{2}{\tau_q} \end{pmatrix}.$$

The eigenvalues of B are

$$\lambda_1 = \lambda_2 = 0,$$

$$\lambda_{3|4|5|6} = \pm\frac{1}{\sqrt{2}}\sqrt{\left(m^2\theta_0 + \frac{2\tau_\theta k}{\tau_q^2} + \alpha_*\right) \pm \sqrt{\left(m^2\theta_0 + \frac{2\tau_\theta k}{\tau_q^2} + \alpha_*\right)^2 - \frac{8\alpha_*\tau_\theta k}{\tau_q^2}}},$$

which are all real, thus characterizing a hyperbolic system. The hyperbolicity also becomes apparent in the results on the domains of dependence in the following sections.

4. Exponential stability. We recall that in classical thermoelasticity as well as in several other thermoelastic models like the Lord–Shulman theory or the model of type III, the exponential stability of the system could be proved for bounded domains in one space dimension as well as for radially symmetric situations in higher

dimensions; see, e.g., [17, 31, 32, 28]. We shall demonstrate the exponential stability in one space dimension. Let (u, θ) satisfy ((3.7), (3.4) or (3.19), (3.20))

$$\tilde{u}_{tt} - \alpha_* \tilde{u}_{xx} + m\tilde{\theta}_x = 0, \tag{4.1}$$

$$\tilde{\theta}_t + m\theta_0 \tilde{u}_{tx} - k\hat{\theta}_{xx} = 0 \tag{4.2}$$

with boundary conditions

$$u = \theta = 0 \qquad \text{for } x = 0, L \tag{4.3}$$

and initial conditions given in terms of the original initial conditions $u(\cdot, 0)$, $u_t(\cdot, 0)$, $\theta(\cdot, 0)$. As in the previous section we define

$$\mathbf{V} := (\tilde{\mathbf{u}}, \tilde{\mathbf{u}}_t, \theta, \theta_t, \theta_{tt})'$$

and we have

$$\begin{aligned} \|V(t)\|_{\mathcal{H}} &= \int_0^L \left\{ \frac{4\theta_0}{\tau_q^4}\tilde{u}_t^2 + \frac{4\theta_0\alpha_*}{\tau_q^4}\tilde{u}_x^2 + \frac{2}{\tau_q^2}\theta_t^2 + \frac{2\tau_\theta k}{\tau_q^2}\theta_{tx}^2 + \theta_{tt}^2 + \frac{4k}{\tau_q^2}\theta_x\theta_{tx} + b_0\theta_x^2 \right\} dx \\ &\equiv 2E_{\mathcal{H}}(t) \end{aligned} \tag{4.4}$$

defining the first "energy" term $E_{\mathcal{H}}(t)$. Another energy term is defined by

$$E(t) := \frac{1}{2}\int_0^L \left\{ \theta_0\tilde{u}_t^2 + \theta_0\alpha_*\tilde{u}_x^2 + \tilde{\theta}^2 + \frac{\tau_q^2\tau_\theta k}{2}\theta_{tx}^2 + k(\tau_q + \tau_\theta)\theta_x^2 + k\tau_q^2\theta_x\theta_{tx} \right\} dx. \tag{4.5}$$

The aim will be to find a suitable Lyapunov functional for the energy terms that proves the exponential stability.

Multiplying the differential equation (4.1) by $\theta_0\tilde{u}_t$ and (4.2) by $\tilde{\theta}$, integrating and performing partial integrations we obtain

$$\begin{aligned} \frac{d}{dt}E(t) &= -k\int_0^L \theta_x^2 dx - \tau_q\left(\tau_\theta - \frac{\tau_q}{2}\right)k\int_0^L \theta_{tx}^2 dx \\ &\le -c_1\int_0^L \left\{\theta_x^2 + \theta_{tx}^2\right\} dx \end{aligned} \tag{4.6}$$

for some positive constant c_1, if the condition (1.7) holds. Then (4.6) reflects the dissipative character of the system. We shall assume (1.7) in the rest of this paper, [26], where the sufficiency and necessity of (1.7) were investigated for the boundary conditions (1.8).

Multiplying the differential equation (4.2) by θ_{tt} and integrating we get

$$\begin{aligned} \frac{d}{dt}\frac{1}{2}\int_0^L \left\{ \theta_{tt}^2 + \frac{2}{\tau_q^2}\theta_t^2 + \frac{4k}{\tau_q^2}\theta_x\theta_{tx} + \frac{2k\tau_\theta}{\tau_q^2}\theta_{tx}^2 \right\} dx &= -\frac{2}{\tau_q}\int_0^L \theta_{tt}^2 dx + \frac{2k}{\tau_q^2}\int_0^L \theta_{tx}^2 dx \\ &\quad - \frac{2m\theta_0}{\tau_q^2}\int_0^L \tilde{u}_{tx}\theta_{tt} dx. \end{aligned} \tag{4.7}$$

Moreover,

$$\frac{d}{dt}\frac{1}{2}\int_0^L b_0\theta_x^2 dx = \int_0^L b_0\theta_x\theta_{tx} dx \le \frac{b_0}{2}\int_0^L \theta_x^2 dx + \frac{b_0}{2}\int_0^L \theta_{tx}^2 dx. \tag{4.8}$$

Multiplying (4.1) by $\frac{4\theta_0}{\tau_q^4}\tilde{u}_t$ and integrating we obtain

$$\frac{d}{dt}\frac{1}{2}\int_0^L \left\{\frac{4\theta_0}{\tau_q^4}\tilde{u}_t^2 + \frac{4\theta_0\alpha_*}{\tau_q^4}\tilde{u}_x^2\right\}dx = \frac{2m\theta_0}{\tau_q^2}\int_0^L \theta_{tt}\tilde{u}_{tx}dx - \frac{4m\theta_0}{\tau_q^3}\int_0^L \theta_{tx}\tilde{u}_t dx - \frac{4m\theta_0}{\tau_q^4}\int_0^L \theta_x\tilde{u}_t dx. \tag{4.9}$$

We conclude from (4.7)–(4.9)

$$\begin{aligned}\frac{d}{dt}E_{\mathcal{H}}(t) &\le -\frac{2}{\tau_q}\int_0^L \theta_{tt}^2 dx + \frac{2k}{\tau_q^2}\int_0^L \theta_{tx}^2 dx + \frac{b_0}{2}\int_0^L \theta_x^2 dx + \frac{b_0}{2}\int_0^L \theta_{tx}^2 dx \\ &\quad - \frac{4m\theta_0}{\tau_q^3}\int_0^L \theta_{tx}\tilde{u}_t dx - \frac{4m\theta_0}{\tau_q^4}\int_0^L \theta_x\tilde{u}_t dx \\ &\le -\frac{2}{\tau_q}\int_0^L \theta_{tt}^2 dx + \left(\frac{b_0}{2} + \frac{4m\theta_0}{\tau_q^4\epsilon_1}\right)\int_0^L \theta_x^2 dx + \left(\frac{2k}{\tau_q^2} + \frac{b_0}{2} + \frac{4m\theta_0}{\tau_q^3\epsilon_1}\right)\int_0^L \theta_{tx}^2 dx \\ &\quad + 2\epsilon_1\int_0^L \tilde{u}_t^2 dx, \end{aligned} \tag{4.10}$$

where $\epsilon_1 > 0$ will be chosen later appropriately small. Combining (4.6) and (4.10) we get

$$\begin{aligned}\frac{d}{dt}(E_{\mathcal{H}}(t) + KE(t)) &\le -\frac{2}{\tau_q}\int_0^L \theta_{tt}^2 dx - \left[Kk - \left(\frac{4m\theta_0}{\tau_q^4\epsilon_1} + \frac{b_0}{2}\right)\right]\int_0^L \theta_x^2 dx \\ &\quad - \left[Kk\tau_q\left(\tau_\theta - \frac{\tau_q}{2}\right) - \left(\frac{4m\theta_0}{\tau_q^3\epsilon_1} + \frac{b_0}{2} + \frac{2k}{\tau_q^2}\right)\right]\int_0^L \theta_{tx}^2 dx \\ &\quad + 2\epsilon_1\int_0^L \tilde{u}_t^2 dx, \end{aligned} \tag{4.11}$$

where $K > 0$ will be chosen below appropriately large. Once ϵ_1 will be fixed, we shall fix K such that the coefficients in $[\cdot]$-brackets in front of the two integrals of the right-hand side in (4.11) will be strictly positive.

Now we follow an ansatz described in [17] for classical thermoelasticity but we have to add essential modifications in order to deal with the higher-order system and the different structure under investigation.

If we multiply the differential equation (4.1) by $\frac{1}{\alpha_*}\tilde{u}_{xx}$ and integrate we obtain after partial integrations

$$\frac{1}{\alpha_*}\frac{d}{dt}\int_0^L \tilde{u}_{tx}\tilde{u}_x dx \le -\frac{2}{3}\int_0^L \tilde{u}_{xx}^2 dx + \frac{1}{\alpha_*}\int_0^L \tilde{u}_{tx}^2 dx + C\int_0^L \tilde{\theta}_x^2 dx, \tag{4.12}$$

where capital C will denote a positive constant that may change from line to line in the sequel. Multiplying the differential equation (4.2) by $\frac{3}{\alpha_* m\theta_0}\tilde{u}_{tx}$ and integrating, using (4.1), yields

$$\begin{aligned}\frac{3}{\alpha_*}\int_0^L \tilde{u}_{tx}^2 dx &= -\frac{3}{\alpha_* m\theta_0}\int_0^L \tilde{\theta}_t\tilde{u}_{tx}dx - \frac{3k}{\alpha_* m\theta_0}\frac{d}{dt}\int_0^L \hat{\theta}_x\left(\frac{1}{\alpha_*}\tilde{u}_{tt} + \frac{m}{\alpha_*}\tilde{\theta}_x\right)dx \\ &\quad + \frac{3k}{\alpha_* m\theta_0}\int_0^L \hat{\theta}_{tx}\tilde{u}_{xx} + \frac{3k}{\alpha_* m\theta_0}[\hat{\theta}_x\tilde{u}_{tx}]_{x=0}^{x=L}, \end{aligned}$$

hence

(4.13)
$$\frac{3}{\alpha_*^2 m\theta_0}\frac{d}{dt}\int_0^L\{\hat\theta_x\tilde u_{tt}+m\hat\theta_x\tilde\theta_x\}dx\le-\frac{2}{\alpha_*}\int_0^L\tilde u_{tx}^2dx+\frac16\int_0^L\tilde u_{xx}^2dx+C\int_0^L\{\hat\theta_{tx}^2+\tilde\theta_t^2\}dx + C\|\hat\theta_x\|_{L^\infty(\{0,L\})}\|\tilde u_{tx}\|_{L^\infty(\{0,L\})},$$

where $\|f\|_{L^\infty(\{0,L\})}:=\max\{|f(0)|,|f(L)|\}$ denotes the sup-norm on the boundary. Combining (4.12) and (4.13) we obtain

$$\frac{d}{dt}\int_0^L\left\{\frac{1}{\alpha_*}\tilde u_{tx}\tilde u_x+\frac{3k}{\alpha_*^2 m\theta_0}\hat\theta_x\tilde u_{tt}+\frac{3k}{\alpha_*^2\theta_0}\hat\theta_x\tilde\theta_x\right\}dx\le-\frac{1}{\alpha_*}\int_0^L\tilde u_{tx}^2dx-\frac12\int_0^L\tilde u_{xx}^2dx + C\int_0^L\{\tilde\theta_x^2+\hat\theta_{tx}^2+\tilde\theta_t^2\}dx+\frac{C}{\epsilon_2}\|\hat\theta_x\|_{L^\infty(\{0,L\})}^2+\epsilon_2\|\tilde u_{tx}\|_{L^\infty(\{0,L\})}^2, \tag{4.14}$$

where $\epsilon_2>0$ will be chosen appropriately small later. The differential equation (4.2) yields

$$\int_0^L\hat\theta_{xx}^2dx\le C\int_0^L\{\tilde\theta_t^2+\tilde u_{tx}^2\}dx.$$

Using this and the Sobolev imbedding $W^{1,1}((0,L))\hookrightarrow L^\infty((0,L))$ we arrive at

$$\|\hat\theta_x\|_{L^\infty(\{0,L\})}^2\le\frac{C}{\epsilon_2^2}\int_0^L\{\hat\theta_x^2+\tilde\theta_t^2\}dx+C\epsilon_2^2\int_0^L\tilde u_{tx}^2dx.$$

Inserting this into (4.14) we conclude for sufficiently small ϵ_2

$$\frac{d}{dt}\int_0^L\left\{\frac{1}{\alpha_*}\tilde u_{tx}\tilde u_x+\frac{3k}{\alpha_*^2 m\theta_0}\hat\theta_x\tilde u_{tt}+\frac{3k}{\alpha_*^2\theta_0}\hat\theta_x\tilde\theta_x\right\}dx\le-\frac{1}{2\alpha_*}\int_0^L\tilde u_{tx}^2dx-\frac12\int_0^L\tilde u_{xx}^2dx + \frac{C}{\epsilon_2^3}\int_0^L\{\tilde\theta_x^2+\hat\theta_{tx}^2+\tilde\theta_t^2\}dx+C_1\epsilon_2\|\tilde u_{tx}\|_{L^\infty(\{0,L\})}^2 \tag{4.15}$$

with a constant $C_1>0$. In order to estimate the boundary term, we use a well-known technique exploiting in the multipliers a smooth extension of the normal at the boundary which means in one dimension to use the following function Φ with

$$\Phi(x):=\frac12-\frac{x}{L}. \tag{4.16}$$

Differentiation of (4.1) with respect to t, multiplying with $\Phi\tilde u_{tx}$ and partially integrating yields

$$0=\frac{d}{dt}\int_0^L\tilde u_{tt}\Phi\tilde u_{tx}dx+\frac12\int_0^L\Phi_x(\tilde u_{tt}^2+\alpha_*\tilde u_{tx}^2)dx+\frac{\alpha_*}{4}(\tilde u_{tx}^2(0)+\tilde u_{tx}^2(L)) +m\int_0^L\left(\theta_{tx}+\tau_q\theta_{ttx}+\frac{\tau_q^2}{2}\theta_{tttx}\right)\Phi\tilde u_{tx}dx$$

whence

$$\frac{d}{dt}\int_0^L\tilde u_{tt}\Phi\tilde u_{tx}dx\le-\frac{\alpha_*}{4}(\tilde u_{tx}^2(0)+\tilde u_{tx}^2(L))+C\int_0^L\{\tilde u_{tt}^2+\tilde u_{tx}^2+\theta_{tx}^2+\theta_{ttx}^2\}dx -\frac{m\tau_q^2}{2}\int_0^L\theta_{tttx}\Phi\tilde u_{tx}dx \tag{4.17}$$

follows. Using the differential equation (4.2) again, we obtain

(4.18)
$$
\begin{aligned}
-m\theta_0\int_0^L \theta_{tttx}\Phi\tilde{u}_{tx}dx &= -k\frac{d}{dt}\int_0^L \theta_{ttx}\Phi\hat{\theta}_{xx}dx + k\int_0^L \theta_{ttx}\Phi\hat{\theta}_{txx}dx + \int_0^L \theta_{ttt}\Phi\theta_{tx}dx\\
&\quad + \int_0^L \theta_{ttt}\Phi_x\theta_t dx + \tau_q\int_0^L \theta_{ttt}\Phi\theta_{ttx}dx + \tau_q\int_0^L \theta_{ttt}\Phi_x\theta_{tt}dx\\
&\quad + \frac{\tau_q^2}{4}\int_0^L \Phi_x\theta_{ttt}^2dx\\
&\le -k\frac{d}{dt}\int_0^L \theta_{ttx}\Phi\hat{\theta}_{xx}dx + \frac{k}{\tau_\theta}\int_0^L \hat{\theta}_{tx}\Phi\hat{\theta}_{txx}dx - \frac{k}{\tau_\theta}\int_0^L \theta_{tx}\Phi\hat{\theta}_{txx}dx\\
&\quad + C\int_0^L \{\theta_t^2+\theta_{tt}^2+\theta_{tx}^2+\theta_{ttt}^2+\theta_{ttx}^2\}dx\\
&= -k\frac{d}{dt}\int_0^L \theta_{ttx}\Phi\hat{\theta}_{xx}dx - \frac{k}{4\tau_\theta}[\hat{\theta}_{tx}^2(0)+\hat{\theta}_{tx}^2(L)]\\
&\quad - \frac{d}{dt}\frac{k}{\tau_\theta}\int_0^L \theta_{tx}\Phi\hat{\theta}_{xx}dx + \frac{k}{\tau_\theta}\int_0^L \theta_{ttx}\Phi\hat{\theta}_{xx}dx\\
&\quad + C\int_0^L \{\theta_t^2+\theta_{tt}^2+\theta_{tx}^2+\theta_{ttt}^2+\theta_{ttx}^2\}dx.
\end{aligned}
$$

Inserting (4.18) into (4.17) and using (4.1) again we get

$$
\begin{aligned}
\frac{d}{dt}\int_0^L \left\{\tilde{u}_{tt}\Phi\tilde{u}_{tx} + \frac{k\tau_q^2}{2\theta_0}\theta_{ttx}\Phi\hat{\theta}_{xx} + \frac{k\tau_q^2}{2\theta_0\tau_\theta}\theta_{tx}\Phi\hat{\theta}_{xx}\right\}dx \le\ & -\frac{\alpha_*}{4}(\tilde{u}_{tx}^2(0)+\tilde{u}_{tx}^2(L))\\
& + C\int_0^L \{\tilde{u}_{tx}^2+\tilde{u}_{xx}^2+\theta_t^2+\theta_x^2+\theta_{tt}^2+\theta_{tx}^2+\theta_{ttt}^2+\theta_{ttx}^2+\hat{\theta}_{xx}^2\}dx. \qquad (4.19)
\end{aligned}
$$

We still have to produce a term $-\int_0^L \hat{\theta}_{xx}^2dx$-term on the right-hand side. This is obtained as follows. We have from (4.2)

$$\hat{\theta}_{xx} = \frac{1}{k}\tilde{\theta}_t + \frac{m\theta_0}{k}\tilde{u}_{tx},$$

which, inserted into (4.19), yields

$$
\begin{aligned}
\frac{d}{dt}\int_0^L \left\{\tilde{u}_{tt}\Phi\tilde{u}_{tx} + \frac{k\tau_q^2}{2\theta_0}\theta_{ttx}\Phi\hat{\theta}_{xx} + \frac{k\tau_q^2}{2\theta_0\tau_\theta}\theta_{tx}\Phi\hat{\theta}_{xx}\right\}dx \le\ & -\frac{\alpha_*}{4}(\tilde{u}_{tx}^2(0)+\tilde{u}_{tx}^2(L))\\
& + C_2\int_0^L \{\tilde{u}_{tx}^2+\tilde{u}_{xx}^2+\theta_t^2+\theta_x^2+\theta_{tt}^2+\theta_{tx}^2+\theta_{ttt}^2+\theta_{ttx}^2\}dx \qquad (4.20)
\end{aligned}
$$

with a constant $C_2 > 0$. A multiplication of (4.20) by $\epsilon_3 > 0$ and then a combination

with (4.15) yields

$$\frac{d}{dt}\int_0^L \left\{ \frac{1}{\alpha_*}\tilde{u}_{tx}\tilde{u}_x + \frac{3k}{\alpha_*^2 m\theta_0}\hat{\theta}_x\tilde{u}_{tt} + \frac{3k}{\alpha_*^2\theta_0}\hat{\theta}_x\tilde{\theta}_x + \epsilon_3\tilde{u}_{tt}\Phi\tilde{u}_{tx} \right.$$
$$\left. + \epsilon_3\frac{k\tau_q^2}{2\theta_0}\theta_{ttx}\Phi\hat{\theta}_{xx} + \epsilon_3\frac{k\tau_q^2}{2\theta_0\tau_\theta}\theta_{tx}\Phi\hat{\theta}_{xx} \right\} dx$$
$$\leq -\frac{1}{2\alpha_*}\int_0^L \tilde{u}_{tx}^2 dx - \frac{1}{2}\int_0^L \tilde{u}_{xx}^2 dx + \frac{C}{\epsilon_2^3}\int_0^L \{\tilde{\theta}_t^2 + \hat{\theta}_{tx}^2 + \tilde{\theta_x}^2\} dx$$
$$- \left[\frac{\alpha_*\epsilon_3}{4} - C_1\epsilon_2\right](\tilde{u}_{tx}^2(0) + \tilde{u}_{tx}^2(L))$$
$$+ C_2\epsilon_3\int_0^L \{\tilde{u}_{tx}^2 + \tilde{u}_{xx}^2\} dx + C_2\varepsilon_3\int_0^L \{\theta_t^2 + \theta_{tt}^2 + \theta_{tx}^2 + \theta_{ttt}^2 + \theta_{ttx}^2 + \tilde{\theta}_x^2\} dx.$$

Now choosing

$$\epsilon_3 := \min\left\{\frac{1}{4\alpha_* C_2}, \frac{1}{2C_2}\right\}$$

and then

$$\epsilon_2 := \frac{\alpha_*\epsilon_3}{4C_1}$$

we obtain

$$\frac{d}{dt}\int_0^L \left\{ \frac{1}{\alpha_*}\tilde{u}_{tx}\tilde{u}_x + \frac{3k}{\alpha_*^2 m\theta_0}\hat{\theta}_x\tilde{u}_{tt} + \frac{3k}{\alpha_*^2\theta_0}\hat{\theta}_x\tilde{\theta}_x + \epsilon_3\tilde{u}_{tt}\Phi\tilde{u}_{tx} \right.$$
$$\left. + \epsilon_3\frac{k\tau_q^2}{2\theta_0}\theta_{ttx}\Phi\hat{\theta}_{xx} + \epsilon_3\frac{k\tau_q^2}{2\theta_0\tau_\theta}\theta_{tx}\Phi\hat{\theta}_{xx} \right\} dx$$
$$\leq -\frac{1}{4\alpha_*}\int_0^L \tilde{u}_{tx}^2 dx - \frac{1}{4}\int_0^L \tilde{u}_{xx}^2 dx$$
$$+ C\int_0^L \{\theta_t^2 + \theta_{tt}^2 + \theta_{tx}^2 + \theta_{ttt}^2 + \theta_{ttx}^2 + \tilde{\theta}_t^2 + \tilde{\theta}_x^2\} dx. \tag{4.21}$$

We observe that by Poincaré's estimates and (4.1) we have

$$\int_0^L \{\tilde{u}_t^2 + \tilde{u}_x^2\} dx \leq C\int_0^L \{\tilde{u}_{tx}^2 + \tilde{u}_{xx}^2\} dx, \quad \int_0^L \tilde{u}_{tt}^2 dx \leq C\int_0^L \{\tilde{u}_{xx}^2 + \tilde{\theta}_x^2\} dx. \tag{4.22}$$

Now let $E(t)$ and $E_{\mathcal{H}}(t)$ be given as defined in (4.4) and (4.5), respectively, and define for $K > 0$ (yet to be determined)

$$W_1(t) \equiv E_1(u, \theta; t) := E_{\mathcal{H}}(t) + KE(t), \qquad W_2(t) := W_1(u_t, \theta_t; t),$$

and the final energy term

$$\mathcal{W}(t) := W_1(t) + W_2(t),$$

where we now choose ϵ_1 small enough such that the terms $2\epsilon_1 \int_0^L \{\tilde{u}_t^2 + \tilde{u}_{tt}^2\} dx$ are absorbed by arising corresponding negative terms (see (4.21), (4.22)). Then we choose K large enough to make sure that the coefficients in $[\cdot]$ brackets in (4.11) are positive. Defining for $\epsilon > 0$ the Lyapunov functional L by

$$L(t) := \frac{1}{\epsilon}\mathcal{W}(t) + \int_0^L \left\{ \frac{1}{\alpha_*}\tilde{u}_{tx}\tilde{u}_x + \frac{3k}{\alpha_*^2 m\theta_0}\hat{\theta}_x\tilde{u}_{tt} + \frac{3k}{\alpha_*^2\theta_0}\hat{\theta}_x\tilde{\theta}_x + \epsilon_3\tilde{u}_{tt}\Phi\tilde{u}_{tx} \right.$$
$$\left. + \epsilon_3\frac{k\tau_q^2}{2\theta_0}\theta_{ttx}\Phi\hat{\theta}_{xx} + \epsilon_3\frac{k\tau_q^2}{2\theta_0\tau_\theta}\theta_{tx}\Phi\hat{\theta}_{xx} \right\} dx,$$

we now find from (4.11), (4.21), (4.22), observing

$$\frac{k}{8m\theta_0\alpha_*}\int_0^L \hat{\theta}_{xx}^2 dx \le \frac{1}{8\alpha_*}\int_0^L \tilde{u}_{tx}^2 dx + \frac{1}{8m\theta_0\alpha_*}\int_0^L \tilde{\theta}_t^2 dx$$

and choosing ϵ small enough that

$$\frac{d}{dt}L(t) \le -C_3\mathcal{W}(t) \tag{4.23}$$

for some constant $C_3 > 0$. Moreover, we have for ϵ small enough

$$\exists\, K_1, K_2 > 0\ \forall\, t \ge 0: \quad K_1\mathcal{W}(t) \le L(t) \le K_2\mathcal{W}(t). \tag{4.24}$$

Combining (4.23) and (4.24) we have thus proved the exponential stability

THEOREM 4.1. *The system* (4.1)–(4.3) *is exponentially stable,*

$$\exists\, d_1, d_2 > 0 \forall\, t \ge 0: \quad \mathcal{W}(t) \le d_1 \mathrm{e}^{-d_2 t}\mathcal{W}(0).$$

The Dirichlet–Neumann-type boundary conditions

$$u_x = \theta = 0 \qquad \text{for } x = 0, L$$

or

$$u = \theta_x = 0 \qquad \text{for } x = 0, L$$

could be treated similarly. It is even likely that one can work just with the first energy $W_1(t)$ (instead of $W_1(t) + W_2(t)$). Moreover, the radially symmetric case in two or three space dimensions should be accessible.

The exponential stability result is first a result for θ and $\tilde{u}$. But we obtain an exponential decay result also for u itself observing that for functions $w, h : [0, \infty) \times (0, L) \to \mathbb{R}$ satisfying

$$\ddot{w} + \frac{2}{\tau_q}\dot{w} + \frac{2}{\tau_q^2}w = h \quad (:= \tilde{u}(t, x))$$

and

$$\exists\, d_1, d_2 > 0\ \ \forall\, t \ge 0: \quad \int_0^L |h(x,t)|^2 dx \le d_1 \mathrm{e}^{-2d_2 t}C_0^2,$$

where C_0 depends on the initial data according to Theorem 4.1, we conclude for $\mathbf{z} := (w, \dot{w})'$,

$$\exists\, d_3, d_4 > 0 \;\; \forall\, t \geq 0 : \;\; \int_0^L |\mathbf{z}(x,t)|^2 dx \leq d_3 \mathrm{e}^{-2d_4 t} \left(\int_0^L |\mathbf{z}(x,0)|^2 dx + C_0^2 \right).$$

Here d_4 can be any positive number smaller than $\min\{1/\tau_q, d_2\}$ which becomes apparent observing that the characteristic values for the ODE for w are

$$\beta_{1,2} = -\frac{1}{\tau_q} \pm \mathrm{i}\frac{1}{\tau_q}.$$

5. Some spatial estimates. In this section we establish results on the spatial evolution of solutions of (2.1)–(2.6), provided that the initial data of (2.4) are assumed to be bounded in a certain energy norm.

We begin by considering

$$F(z,t) = -\int_0^t \int_{D(z)} \left(\tilde{T}_{i3} \dot{\tilde{u}}_i + \frac{1}{\theta_0} k \hat{\theta}_{,3} \tilde{\theta} \right) dA ds. \tag{5.1}$$

From (5.2) we find that

$$\frac{\partial F(z,t)}{\partial t} = -\int_{D(z)} \left(\tilde{T}_{i3} \dot{\tilde{u}}_i + \frac{1}{\theta_0} k \hat{\theta}_{,3} \tilde{\theta} \right) dA, \tag{5.2}$$

and, on using (2.1), the divergence theorem on $D(z)$ and (2.4), (2.5), we obtain

$$\begin{aligned} \frac{\partial F(z,t)}{\partial z} = -\frac{1}{2} \int_{D(z)} \Bigg(& \rho \dot{\tilde{u}}_i \dot{\tilde{u}}_i + \mu \tilde{u}_{i,j} \tilde{u}_{i,j} + (\lambda + \mu) \tilde{u}_{r,r} \tilde{u}_{s,s} + \frac{c}{\theta_0} (\tilde{\theta})^2 \\ & + \frac{k}{\theta_0} \Big((\tau_q + \tau_\theta) |\nabla \theta|^2 + \frac{1}{2} \tau_q^2 \tau_\theta |\nabla \dot{\theta}|^2 + \tau_q^2 \nabla \theta \nabla \dot{\theta} \Big) \Bigg) dA \end{aligned} \tag{5.3}$$

$$- \int_0^t \int_{D(z)} \frac{k}{\theta_0} \left(|\nabla \theta|^2 + \left(\tau_\theta \tau_q - \frac{1}{2} \tau_q^2 \right) |\nabla \dot{\theta}|^2 \right) dA ds + E_1(z),$$

where

$$E_1(z) = \frac{1}{2} \int_{D(z)} \Big(\rho \tilde{v}_i^0 \tilde{v}_i^0 + \mu \tilde{u}_{i,j}^0 \tilde{u}_{i,j}^0 + (\lambda + \mu) \tilde{u}_{r,r}^0 \tilde{u}_{s,s}^0 \Big) dA \tag{5.4}$$

$$+ \frac{1}{2\theta_0} \int_{D(z)} \left(c(\tilde{\theta}^0)^2 + k \left((\tau_q + \tau_\theta) |\nabla \theta^0|^2 + \frac{1}{2} \tau_q^2 \tau_\theta |\nabla \vartheta^0|^2 + \tau_q^2 \nabla \theta^0 \nabla \vartheta^0 \right) \right) dA.$$

Note that $E_1(z)$ depends only on the initial data (2.4) and we note that several time derivative at time zero can be obtained assuming the continuity of the solutions at time $t = 0$. Rewriting (5.3) with z replaced by the variable η, and integrating with

respect to η from 0 to z, we get

$$F(z,t)-F(0,t)=-\frac{1}{2}\int_0^z\int_{D(\eta)}\Big(\rho\dot{\tilde{u}}_i\dot{\tilde{u}}_i+\mu\tilde{u}_{i,j}\tilde{u}_{i,j}+(\lambda+\mu)\tilde{u}_{r,r}\tilde{u}_{s,s}\Big)dV \tag{5.5}$$
$$-\frac{1}{2\theta_0}\int_0^z\int_{D(\eta)}\left(c(\tilde{\theta})^2+k\left((\tau_q+\tau_\theta)|\nabla\theta|^2+\frac{k}{2}\tau_q^2\tau_\theta|\nabla\dot{\theta}|^2+k\tau_q^2\nabla\theta\nabla\dot{\theta}\right)\right)dV$$
$$-\frac{k}{\theta_0}\int_0^t\int_0^z\int_{D(\eta)}\left(|\nabla\theta|^2+\left(\tau_\theta\tau_q-\frac{1}{2}\tau_q^2\right)|\nabla\dot{\theta}|^2\right)dVds$$
$$+\frac{1}{2}\int_0^z\int_{D(\eta)}\Big(\rho\tilde{v}_i^0\tilde{v}_i^0+\mu\tilde{u}_{i,j}^0\tilde{u}_{i,j}^0+(\lambda+\mu)\tilde{u}_{r,r}^0\tilde{u}_{s,s}^0\Big)dV$$
$$+\frac{1}{2\theta_0}\int_0^z\int_{D(\eta)}\left(c(\tilde{\theta}^0)^2+k\left((\tau_q+\tau_\theta)|\nabla\theta^0|^2+\frac{1}{2}\tau_q^2\tau_\theta|\nabla\vartheta^0|^2+\tau_q^2\nabla\theta^0\nabla\vartheta^0\right)\right)dV.$$

Our next step is to establish an inequality between the time and spatial derivatives of $F(z,t)$. By virtue of (1.7), the second integral on the right in (5.3) is nonnegative. The last three terms in the integrand in the first integral on the right in (5.3) are a quadratic form and may be bounded below using the smallest positive eigenvalue λ_0 of (2.7), (2.10) given in (2.11). Thus we find that

$$\frac{\partial F(z,t)}{\partial z}\le-\frac{1}{2}\int_{D(z)}\Bigg(\rho\dot{\tilde{u}}_i\dot{\tilde{u}}_i+\mu\tilde{u}_{i,j}\tilde{u}_{i,j}+(\lambda+\mu)\tilde{u}_{r,r}\tilde{u}_{s,s}$$
$$+\frac{c}{\theta_0}(\tilde{\theta})^2+\frac{k\lambda_0}{\theta_0}(|\nabla\theta|^2+|\nabla\dot{\theta}|^2)\Bigg)dA+E_1(z). \tag{5.6}$$

Applying Schwarz's inequality in (5.3) and using (2.3), we get

$$\left|\frac{\partial F}{\partial t}\right|\le\frac{1}{2}\int_{D(z)}\left[\frac{\epsilon_1}{\rho}\tilde{T_{ij}}\tilde{T_{ij}}+\frac{\rho}{\epsilon_1}\dot{\tilde{u}}_i\dot{\tilde{u}}_i+\frac{c}{\epsilon_2\theta_0}(\tilde{\theta})^2+\frac{\epsilon_2k^2}{c\theta_0}\hat{\theta}_{,3}\hat{\theta}_{,3}\right]dA \tag{5.7}$$
$$\le\frac{1}{2}\int_{D(z)}\left[\frac{\epsilon_1}{\rho}(1+\epsilon)\mu^*[\mu\tilde{u}_{i,j}\tilde{u}_{i,j}+(\lambda+\mu)\tilde{u}_{r,r}\tilde{u}_{s,s}\right]+\frac{1}{\epsilon_1}(\rho\dot{\tilde{u}}_i\dot{\tilde{u}}_i)$$
$$+\left(\frac{1}{\epsilon_2}+\frac{3m^2\epsilon_1\theta_0}{\rho c}(1+\epsilon^{-1})\right)\frac{c}{\theta_0}(\tilde{\theta})^2+\frac{\epsilon_2k(1+\tau_\theta^2)}{c\lambda_0}\left[\frac{k\lambda_0}{\theta_0}(|\nabla\theta|^2+|\nabla\dot{\theta}|^2)\right]dA,$$

where the weighted arithmetic-geometric mean inequality has been employed and where ϵ_i are arbitrary positive constants.

Now, we equate the coefficients of the energetic terms in the last integral of the (5.7). We get

$$\frac{1}{\epsilon_1}=\frac{\epsilon_1}{\rho}(1+\epsilon)\mu^*=\frac{1}{\epsilon_2}+\frac{3m^2\epsilon_1\theta_0}{\rho c}(1+\epsilon^{-1})=\frac{\epsilon_2k(1+\tau_\theta^2)}{c\lambda_0}. \tag{5.8}$$

That is,

$$\epsilon_1=\beta^{-1},\epsilon_2=\frac{c\lambda_0\beta}{k(1+\tau_\theta^2)},\ \beta=\sqrt{\frac{(1+\epsilon_0)\mu^*}{\rho}}, \tag{5.9}$$

where ϵ_0 is the positive root of the second-order equation

$$x^2 + \left(1 - \frac{\rho k(1+\tau_0^2)}{\mu^* \lambda_0 c} - \frac{3m^2\theta_0}{\mu^* c}\right) x - \frac{3m^2\theta_0}{\mu^* c} = 0. \tag{5.10}$$

In view of (5.6) we can write (5.7) as

$$\left|\frac{\partial F}{\partial t}\right| + \beta \frac{\partial F}{\partial z} \leq \beta E_1(z), \tag{5.11}$$

where β is defined at (5.9).

The inequality (5.11) implies that

$$\frac{\partial F}{\partial t} + \beta \frac{\partial F}{\partial z} \leq \beta E_1(z) \tag{5.12}$$

and

$$\frac{\partial F}{\partial t} - \beta \frac{\partial F}{\partial z} \geq -\beta E_1(z). \tag{5.13}$$

Integrating (5.12) and recalling the definition of $E_1(z)$ in (5.4) we obtain

$$F(z, \beta^{-1}(z - z^*)) \leq \frac{1}{2}\int_{z^*}^{z}\int_{D(\eta)} \Big(\rho \tilde{v}_i^0 \tilde{v}_i^0 + \mu \tilde{u}_{i,j}^0 \tilde{u}_{i,j}^0 + (\lambda+\mu)\tilde{u}_{r,r}^0 \tilde{u}_{s,s}^0\Big) dV, \tag{5.14}$$

$$+ \frac{1}{2\theta_0}\int_{z^*}^{z}\int_{D(\eta)} \left(c(\tilde{\theta}^0)^2 + k\left((\tau_q+\tau_\theta)|\nabla\theta^0|^2 + \frac{1}{2}\tau_q^2\tau_\theta|\nabla\vartheta^0|^2 + \tau_q^2 \nabla\theta^0\nabla\vartheta^0\right)\right) dV,$$

where $z \geq z^*$. Similarly, by integrating (5.13) we obtain

$$F(z, \beta^{-1}(z^{**} - z)) \geq -\frac{1}{2}\int_{z}^{z^{**}}\int_{D(\eta)} \Big(\rho \tilde{v}_i^0 \tilde{v}_i^0 + \mu \tilde{u}_{i,j}^0 \tilde{u}_{i,j}^0 + (\lambda+\mu)\tilde{u}_{r,r}^0 \tilde{u}_{s,s}^0\Big) dV, \tag{5.15}$$

$$- \frac{1}{2\theta_0}\int_{z}^{z^{**}}\int_{D(\eta)} \left(c(\tilde{\theta}^0)^2 + k\left((\tau_q+\tau_\theta)|\nabla\theta^0|^2 + \frac{1}{2}\tau_q^2\tau_\theta|\nabla\vartheta^0|^2 + \tau_q^2 \nabla\theta^0\nabla\vartheta^0\right)\right) dV,$$

where $z^{**} \geq z$. Let

$$\mathcal{E}(z,t) := \frac{1}{2}\int_{R(z)} \Big(\rho \dot{\tilde{u}}_i \dot{\tilde{u}}_i + \mu \tilde{u}_{i,j}\tilde{u}_{i,j} + (\lambda+\mu)\tilde{u}_{r,r}\tilde{u}_{s,s}\Big) dV \tag{5.16}$$

$$+ \frac{1}{2\theta_0}\int_{R(z)} \left(c(\tilde{\theta})^2 + k\left((\tau_q+\tau_\theta)|\nabla\theta|^2 + \frac{k}{2}\tau_q^2\tau_\theta|\nabla\dot{\theta}|^2 + k\tau_q^2\nabla\theta\nabla\dot{\theta}\right)\right) dV$$

$$+ \frac{k}{\theta_0}\int_0^t\int_{R(z)} \left(|\nabla\theta|^2 + \left(\tau_\theta\tau_q - \frac{1}{2}\tau_q^2\right)|\nabla\dot{\theta}|^2\right) dV ds.$$

If we now assume that the initial data (2.4) is such that

$$\mathcal{E}(0,0) = \frac{1}{2}\int_R \Big(\rho \tilde{v}_i^0 \tilde{v}_i^0 + \mu \tilde{u}_{i,j}^0 \tilde{u}_{i,j}^0 + (\lambda+\mu)\tilde{u}_{r,r}^0 \tilde{u}_{s,s}^0\Big) dV \tag{5.17}$$

$$+\frac{1}{2\theta_0}\int_R\left(c(\tilde{\theta}^0)^2+k\left((\tau_q+\tau_\theta)|\nabla\theta^0|^2+\frac{1}{2}\tau_q^2\tau_\theta|\nabla\vartheta^0|^2+\tau_q^2\nabla\theta^0\nabla\vartheta^0\right)\right)dV<\infty,$$

then the inequalities (5.14), (5.15) imply that for each finite time t,

$$\lim_{z\to\infty}F(z,t)=0. \tag{5.18}$$

Thus, we may rewrite

$$\begin{aligned}F(z,t)&=\frac{1}{2}\int_{R(z)}\left(\rho\dot{\tilde{u}}_i\dot{\tilde{u}}_i+\mu\tilde{u}_{i,j}\tilde{u}_{i,j}+(\lambda+\mu)\tilde{u}_{r,r}\tilde{u}_{s,s}\right)dV\\&+\frac{1}{2\theta_0}\int_{R(z)}\left(c(\tilde{\theta})^2+k\left((\tau_q+\tau_\theta)|\nabla\theta|^2+\frac{k}{2}\tau_q^2\tau_\theta|\nabla\dot{\theta}|^2+k\tau_q^2\nabla\theta\nabla\dot{\theta}\right)\right)dV\\&+\frac{k}{\theta_0}\int_0^t\int_{R(z)}\left(|\nabla\theta|^2+\left(\tau_\theta\tau_q-\frac{1}{2}\tau_q^2\right)|\nabla\dot{\theta}|^2\right)dVds\\&-\frac{1}{2}\int_{R(z)}\left(\rho\tilde{v}_i^0\tilde{v}_i^0+\mu\tilde{u}_{i,j}^0\tilde{u}_{i,j}^0+(\lambda+\mu)\tilde{u}_{r,r}^0\tilde{u}_{s,s}^0\right)dV.\end{aligned} \tag{5.19}$$

$$-\frac{1}{2\theta_0}\int_{R(z)}\left(c(\tilde{\theta}^0)^2+k\left((\tau_q+\tau_\theta)|\nabla\theta^0|^2+\frac{1}{2}\tau_q^2\tau_\theta|\nabla\vartheta^0|^2+\tau_q^2\nabla\theta^0\nabla\vartheta^0\right)\right)dV.$$

Now the inequality (5.12) implies that

$$\mathcal{E}(z,t)\le\mathcal{E}(z^*,0), \tag{5.20}$$

where z, z^*, and t are related by $t=\beta^{-1}(z-z^*)$. In a similar way we get

$$\mathcal{E}(z,t)\ge\mathcal{E}(z^{**},0) \tag{5.21}$$

for $t=\beta^{-1}(z^{**}-z)$. From the inequalities (5.20) and (5.21) we conclude that

$$\mathcal{E}(z,t)\le\mathcal{E}(z^*,t^*) \tag{5.22}$$

for $|t-t^*|\le\beta^{-1}(z-z^*)$. Thus, we have proved the next theorem.

THEOREM 5.1. *Let* $(\mathbf{u},\theta)$ *be a solution of the initial boundary value problem* (2.1)–(2.6). *Then the energy function* $\mathcal{E}(z,t)$ *defined in* (5.16) *satisfies the inequality* (5.22) *whenever* $|t-t^*|\le\beta^{-1}(z-z^*)$, *provided that the initial data satisfy* (5.17).

We note that this result gives an answer to the question proposed by Hetnarski and Ignaczak [10, p. 474] a principle of Saint-Venant's type in this theory.

If one defines the measure

$$\mathcal{E}^*(z,t)=\int_0^t\mathcal{E}(z,s)ds, \tag{5.23}$$

the following inequalities can be obtained as in [2]:

$$\mathcal{E}^*(z,t)\le\beta^{-1}\int_{z-\beta t}^z\mathcal{E}(\eta,0)d\eta,\ \beta t\le z, \tag{5.24}$$

$$\mathcal{E}^*(z,t)\le\beta^{-1}\int_0^z\mathcal{E}(\eta,0)d\eta+\left(1-\frac{z}{\beta t}\right)\mathcal{E}^*(0,t),\ \beta t\ge z. \tag{5.25}$$

6. Consequences of the estimates (5.22), (5.24), (5.25). In this section we show some consequences of the estimates (5.22), (5.24), and (5.25).

First, we assume that the initial conditions (2.4) are homogeneous. In this case we see that $\mathcal{E}(0,0)=0$. Estimate (5.22) implies that $\mathcal{E}(z,t)=0$ whenever $\beta t \leq z$. In view of the definition (5.20) we obtain that

$$\tilde{u}_i = 0, \;\; \theta = 0 \tag{6.1}$$

whenever $\beta t \leq z$. Then for every $\mathbf{x}=(x_1,x_2,z)$ such that $\beta t \leq z$, the functions $u_i(\mathbf{x},t)$ satisfy the ordinary differential equation $\tilde{u}_i = 0$ with null initial conditions. Thus, we also conclude that

$$u_i = 0 \tag{6.2}$$

when $\beta t \leq z$. This is a result of the kind of the domain of dependence of the solutions. We have proved the next theorem.

THEOREM 6.1. *Let $(\mathbf{u},\theta)$ be a solution of the initial boundary value problem* (2.1)–(2.6) *when the initial conditions are null. Then $(\mathbf{u},\theta)=(\mathbf{0},0)$ whenever $\beta t \leq z$.*

We note that this result gives an answer to the question proposed by Hetnarski and Ignaczak [10, p. 474] concerning a general domain of influence theorem in this theory.

In this situation it is natural to look for estimates for

$$\mathcal{H}(z,t) := \int_z^\infty \mathcal{E}(\xi,t)d\xi, \tag{6.3}$$

where $z \leq \beta t$. We have that

$$\mathcal{H}(z,t) = \int_z^{\beta t} \mathcal{E}(\xi,t)d\xi. \tag{6.4}$$

But

$$\mathcal{E}(z,t) \leq \mathcal{E}(0,z^*) \tag{6.5}$$

when $z^* \geq t - \beta^{-1}z \geq 0$. Thus, it follows that

$$\mathcal{H}(z,t) \leq \frac{\beta t - z}{t}\int_0^t \mathcal{E}(0,s)ds = \frac{\beta t - z}{t}\mathcal{E}^*(0,t). \tag{6.6}$$

The second natural question we are interested in is to obtain spatial estimates for some norm of the solutions. We had obtained the estimates (5.22), (5.24), and (5.25), but they are expressed in a combination of the solution and its time derivatives. Now, we give explicit spatial estimates. From (5.20), (5.22), (5.24), and (5.25) we have

$$\begin{aligned}\mathcal{J}(z,t) &= \frac{1}{2\theta_0}\int_{R(z)}\left(k\left((\tau_q+\tau_\theta)|\nabla\theta|^2 + \frac{1}{2}\tau_q^2\tau_\theta|\nabla\dot{\theta}|^2 + \tau_q^2\nabla\theta\nabla\dot{\theta}\right)\right)dV \\ &+ \frac{k}{\theta_0}\int_0^t\int_{R(z)}\left(|\nabla\theta|^2 + \left(\tau_\theta\tau_q - \frac{1}{2}\tau_q^2\right)|\nabla\dot{\theta}|^2\right)dVds \leq \mathcal{E}(z,t)\end{aligned} \tag{6.7}$$

for $|t-t^*| \leq \beta^{-1}(z-z^*)$. Also, we obtain that

$$\mathcal{J}^*(z,t) \leq \beta^{-1}\int_{z-\beta t}^z \mathcal{E}(\eta,0)d\eta, \;\; \beta t \leq z, \tag{6.8}$$

and

$$\mathcal{J}^*(z,t) \leq \beta^{-1}\int_0^z \mathcal{E}(\eta,0)d\eta + (1-\frac{z}{\beta t})\mathcal{E}^*(0,t), \quad \beta t \geq z, \tag{6.9}$$

where

$$\mathcal{J}^*(z,t) = \int_0^t \mathcal{J}(z,s)ds. \tag{6.10}$$

Now, we will obtain estimates for the mechanical part. To this end, we note that

$$\begin{aligned}\int_0^t (\tilde{f})^2 ds = \int_0^t \left(f^2 + \frac{\tau_q^4}{4}(\ddot{f})^2\right) ds + \tau_q\left(f^2(t) + \tau_q f(t)\dot{f}(t) + \frac{\tau_q^2}{2}\dot{f}^2(t)\right)\\ - \tau_q\left(f^2(0) + \tau_q f(0)\dot{f}(0) + \frac{\tau_q^2}{2}\dot{f}^2(0)\right).\end{aligned} \tag{6.11}$$

It is worth noting that the expression

$$\tau_q\left(f^2(t) + \tau_q f(t)\dot{f}(t) + \frac{\tau_q^2}{2}\dot{f}^2(t)\right)$$

is positive in the sense that it is equivalent to the measure defined by $f^2(t) + \dot{f}^2(t)$. Thus, if we define

$$\begin{aligned}\mathcal{M}^*(z,t) = \frac{1}{2}\int_0^t\int_{R(z)}\Bigg(\rho\left(\dot{u}_i\dot{u}_i + \frac{\tau_q^4}{4}\dddot{u}_i\dddot{u}_i\right) + \mu\left(u_{i,j}u_{i,j} + \frac{\tau_q^4}{4}\ddot{u}_{i,j}\ddot{u}_{i,j}\right)\\ + (\lambda+\mu)\left(u_{r,r}u_{s,s} + \frac{\tau_q^4}{4}\ddot{u}_{r,r}\ddot{u}_{s,s}\right)\Bigg)dVds\end{aligned}$$

$$\begin{aligned}+\frac{\tau_q}{2}\int_{R(z)}\Bigg(\rho\left(\dot{u}_i\dot{u}_i + \tau_q\dot{u}_i\ddot{u}_i + \frac{\tau_q^2}{2}\ddot{u}_i\ddot{u}_i\right) + \mu\left(u_{i,j}u_{i,j} + \tau_q u_{i,j}\dot{u}_{i,j} + \frac{\tau_q^2}{2}\dot{u}_{i,j}\dot{u}_{i,j}\right)\\ + (\lambda+\mu)\left(u_{r,r}u_{s,s} + \tau_q u_{r,r}\dot{u}_{s,s} + \frac{\tau_q^2}{2}\dot{u}_{r,r}\dot{u}_{s,s}\right)\Bigg)dV,\end{aligned} \tag{6.12}$$

we obtain the estimates

$$\mathcal{M}^*(z,t) \leq \beta^{-1}\int_{z-\beta t}^z \mathcal{E}(\eta,0)d\eta + \mathcal{P}(z), \quad \beta t \leq z, \tag{6.13}$$

and

$$\mathcal{M}^*(z,t) \leq \beta^{-1}\int_0^z \mathcal{E}(\eta,0)d\eta + (1-\frac{z}{\beta t})\mathcal{E}^*(0,t) + \mathcal{P}(z), \quad \beta t \geq z, \tag{6.14}$$

where

(6.15)
$$\mathcal{P}(z)=\frac{\tau_q}{2}\int_{R(z)}\left(\rho\left(v_i^0v_i^0+\tau_qv_i^0z_i^0+\frac{\tau_q^2}{2}z_i^0z_i^0\right)+\mu\left(u_{i,j}^0u_{i,j}^0+\tau_qu_{i,j}^0v_{i,j}^0+\frac{\tau_q^2}{2}v_{i,j}^0v_{i,j}^0\right)\right.$$
$$\left.+(\lambda+\mu)\left(u_{r,r}^0u_{s,s}^0+\tau_qu_{r,r}^0v_{s,s}^0+\frac{\tau_q^2}{2}v_{r,r}^0v_{s,s}^0\right)\right)dV$$

and

$$z_i^0=(\mu u_{i,j}^0+(\lambda+\mu)u_{r,r}^0\delta_{ij}+m\delta_{ij}\theta)_{,j}. \tag{6.16}$$

It is worth noting that it is also possible to obtain estimates in the L^2-norm of the temperature and its two first times derivatives in a similar way of the estimates (6.13), (6.14).

7. A nonstandard problem for (2.1), (2.2). In this section, we briefly discuss the behavior of solutions of (2.1), (2.2) subject to the boundary condition (2.5), (2.6) and the nonstandard conditions

$$u_i(\mathbf{x},T)=\alpha u_i(\mathbf{x},0),\ \dot{u}_i(\mathbf{x},T)=\alpha\dot{u}_i(\mathbf{x},0),$$
$$\theta(\mathbf{x},T)=\alpha\theta(\mathbf{x},0),\ \dot{\theta}(\mathbf{x},T)=\alpha\dot{\theta}(\mathbf{x},0),\ \ddot{\theta}(\mathbf{x},T)=\alpha\ddot{\theta}(\mathbf{x},0), \tag{7.1}$$

where $\alpha>1$. Such nonstandard conditions have been the subject of much recent attention (see, e.g., [1, 16, 33] in the context of the heat equation, [21] for generalized heat conduction and [22] for viscous flows, [19] for the isothermal elasticity, and [30], [27] for some thermoelastic theories).

The boundary data in (2.6) are assumed compatible with (6.1), (6.2).

The analysis begins by considering the function

$$F_\gamma(z)=-\int_0^T\int_{D(z)}\exp(-\gamma s)\left(\tilde{T}_{i3}\dot{\tilde{u}}_i+\frac{1}{\theta_0}k\hat{\theta}_{,3}\tilde{\theta}\right)dAds, \tag{7.2}$$

where, guided by results established in [1, 16, 33], the positive constant γ is given by

$$\gamma=\frac{2}{T}\ln\alpha. \tag{7.3}$$

We have

(7.4)
$$F_\gamma(z)=F_\gamma(0)+\frac{\gamma}{2}\int_0^T\int_0^z\int_{D(\eta)}\exp(-\gamma s)\left(\rho\dot{\tilde{u}}_i\dot{\tilde{u}}_i+\mu\tilde{u}_{i,j}\tilde{u}_{i,j}+(\lambda+\mu)\tilde{u}_{r,r}\tilde{u}_{s,s}\right.$$
$$\left.+\frac{c}{\theta_0}(\tilde{\theta})^2+\frac{k}{\theta_0}\left((\tau_q+\tau_\theta)|\nabla\theta|^2+\frac{1}{2}\tau_q^2\tau_\theta|\nabla\dot{\theta}|^2+\tau_q^2\nabla\theta\nabla\dot{\theta}\right)\right)dA$$
$$+\gamma\frac{k}{\theta_0}\int_0^T\int_0^z\int_{D(\eta)}\exp(-\gamma s)\left(\frac{1}{\gamma}|\nabla\theta|^2+\frac{1}{\gamma}\left(\tau_\theta\tau_q-\frac{1}{2}\tau_q^2\right)|\nabla\dot{\theta}|^2\right)dVds.$$

An argument similar to the one used in the case of the standard initial conditions leads to the estimate

$$|F_\gamma|\le\gamma\beta_\gamma\frac{\partial F_\gamma}{\partial z}, \tag{7.5}$$

where β_γ is defined in the same form of β defined in (5.9) but with changing the parameter λ_0 by μ_γ defined at (2.13).

This inequality is well known in the study of spatial decay estimates. It implies that

$$F_\gamma \leq \gamma\beta_\gamma \frac{\partial F_\gamma}{\partial z} \text{ and } -F_\gamma \leq \gamma\beta_\gamma \frac{\partial F_\gamma}{\partial z}. \tag{7.6}$$

From (7.6), we can obtain an alternative of Phragmen–Lindelöf type, which states (see [5]) that either the solutions grow exponentially for z sufficiently large or solutions decay exponentially in the form

$$\mathcal{E}_\gamma(z) \leq \mathcal{E}_\gamma(0)\exp\Big(-\gamma^{-1}\beta_\gamma^{-1}z\Big) \tag{7.7}$$

for all $z \geq 0$, where

$$\begin{aligned}\mathcal{E}_\gamma(z) = \frac{\gamma}{2}\int_0^T\int_0^z\int_{D(\eta)} \exp(-\gamma s)\Bigg(&\rho\dot{\tilde{u}}_i\dot{\tilde{u}}_i + \mu\tilde{u}_{i,j}\tilde{u}_{i,j} + (\lambda+\mu)\tilde{u}_{r,r}\tilde{u}_{s,s} \\ &+ \frac{c}{\theta_0}(\tilde{\theta})^2 + \frac{k}{\theta_0}\left((\tau_q+\tau_\theta)|\nabla\theta|^2 + \frac{1}{2}\tau_q^2\tau_\theta|\nabla\dot{\theta}|^2 + \tau_q^2\nabla\theta\nabla\dot{\theta}\right)\Bigg)dVds \\ &+ \frac{k}{\theta_0}\int_0^T\int_0^z\int_{D(\eta)}\exp(-\gamma s)\left(|\nabla\theta|^2 + \left(\tau_\theta\tau_q - \frac{1}{2}\tau_q^2\right)|\nabla\dot{\theta}|^2\right)dVds.\end{aligned} \tag{7.8}$$

The decay rate in (7.7) depends explicitly on γ given in (7.3). Thus, we have proved the next theorem.

THEOREM 7.1. *Let $(\mathbf{u},\theta)$ be a solution of the initial boundary value problem* (2.1), (2.2), (2.5), (2.6), *and* (7.1). *Then either the solutions grow exponentially or the estimate* (7.7) *is satisfied, where $\mathcal{E}_\gamma$ is defined at* (7.8).

Estimate (7.7) implies that

$$\mathcal{J}_\gamma(z) \leq \mathcal{E}_\gamma(0)\exp\left(-\gamma^{-1}\beta_\gamma^{-1}z\right), \tag{7.9}$$

where

$$\begin{aligned}\mathcal{J}_\gamma(z) = \frac{\gamma}{2\theta_0}\int_0^T\int_{R(z)}\exp(-\gamma s)\left(k\left((\tau_q+\tau_\theta)|\nabla\theta|^2 + \frac{1}{2}\tau_q^2\tau_\theta|\nabla\dot{\theta}|^2 + \tau_q^2\nabla\theta\nabla\dot{\theta}\right)\right)dVds \\ + \frac{k}{\theta_0}\int_0^T\int_{R(z)}\exp(-\gamma s)\left(|\nabla\theta|^2 + \left(\tau_\theta\tau_q - \frac{1}{2}\tau_q^2\right)|\nabla\dot{\theta}|^2\right)dVds.\end{aligned} \tag{7.10}$$

To obtain an estimate on the mechanical part, we first note that

$$\begin{aligned}\int_0^T\exp(-\gamma s)(\tilde{f})^2ds = \int_0^T\exp(-\gamma s)\Bigg(&\left(f^2 + \frac{\tau_q^4}{4}(\ddot{f})^2\right) \\ &+ \gamma\tau_q\left(f^2 + \tau_q f\dot{f} + \frac{\tau_q^2}{2}\dot{f}^2\right)\Bigg)ds \\ + \exp(-\gamma T)&\left(\tau_q\left(f^2(T) + \tau_q f(T)\dot{f}(T) + \frac{\tau_q^2}{2}\dot{f}^2(T)\right)\right) \\ &- \tau_q\left(f^2(0) + \tau_q f(0)\dot{f}(0) + \frac{\tau_q^2}{2}\dot{f}^2(0)\right).\end{aligned} \tag{7.11}$$

If we define

$$\mathcal{M}_\gamma(z) = \frac{1}{2}\int_0^T \int_{R(z)} \exp(-\gamma s)\Bigg(\rho\left(\dot{u}_i\dot{u}_i + \frac{\tau_q^4}{4}\dddot{u}_i\dddot{u}_i\right) + \mu\left(u_{i,j}u_{i,j} + \frac{\tau_q^4}{4}\ddot{u}_{i,j}\ddot{u}_{i,j}\right)$$
$$+ (\lambda+\mu)\left(u_{r,r}u_{s,s} + \frac{\tau_q^4}{4}\ddot{u}_{r,r}\ddot{u}_{s,s}\right)\Bigg)dVds$$
$$+ \frac{\gamma\tau_q}{2}\int_0^T \int_{R(z)} \exp(-\gamma s)\left(\rho\left(\dot{u}_i\dot{u}_i + \tau_q\dot{u}_i\ddot{u}_i + \frac{\tau_q^2}{2}\ddot{u}_i\ddot{u}_i\right)\right.$$
$$+ \mu\left(u_{i,j}u_{i,j} + \tau_q u_{i,j}\dot{u}_{i,j} + \frac{\tau_q^2}{2}\dot{u}_{i,j}\dot{u}_{i,j}\right)$$
$$\left.+ (\lambda+\mu)\left(u_{r,r}u_{s,s} + \tau_q u_{r,r}\dot{u}_{s,s} + \frac{\tau_q^2}{2}\dot{u}_{r,r}\dot{u}_{s,s}\right)\right)dVds, \tag{7.12}$$

we obtain the estimate

$$\mathcal{M}_\gamma(z) \le \mathcal{E}_\gamma(0)\exp\left(-\gamma^{-1}\beta_\gamma^{-1}z\right), \tag{7.13}$$

which is a spatial decay estimate.

REFERENCES

[1] K. A. Ames, L. E. Payne, and P. W. Schaefer, *On a nonstandard problem for heat conduction in a cylinder*, Appl. Anal., 83 (2004), pp. 125–133.

[2] F. Bofill and R. Quintanilla, *Spatial estimates for dynamical problems in several elasticity theories*, Ricerche. Mat., 46 (1997), pp. 425–441.

[3] D. S. Chandrasekharaiah, *Hyperbolic thermoelasticity: A review of recent literature*, Appl. Mech. Rev., 51 (1998), pp. 705–729.

[4] S. Chiriţă and M. Ciarletta, *Some further growth and decay results in linear thermoelastodynamics*, J. Thermal Stresses, 26 (2003), pp. 889–903.

[5] J. N. Flavin, R. J. Knops, and L. E. Payne, *Decay estimates for the constrained elastic cylinder of variable cross section*, Quart. Appl. Math., 47 (1989), pp. 325–350.

[6] J. N. Flavin and S. Rionero, *Qualitative Estimates for Partial Differential Equations*, CRC Press, Boca Raton, FL, 1996.

[7] F. Franchi and B. Straughan, *Continuous dependence on the relaxation time and modelling, and unbounded growth in theories of heat conduction with finite propagation speeds*, J. Math. Anal. Appl., 185 (1994), pp. 726–746.

[8] A. F. Ghaleb and M. Sh. El-Deen Mohamedein, *A heat conduction equation with three relaxation times*, Internat. J. Engrg. Sci., 27 (1989), pp. 1367–1377.

[9] J. Goldstein, *Semigroups of Linear Operators and Applications*, Oxford University Press, Oxford, UK, 1985.

[10] R. B. Hetnarski and J. Ignaczak, *Generalized thermoelasticity*, J. Thermal Stresses, 22 (1999), pp. 451–470.

[11] C. O. Horgan, *Recent developments concerning Saint-Venant's principle: An update*, Appl. Mech. Rev., 42 (1989), pp. 295–303.

[12] C. O. Horgan, *Recent developments concerning Saint-Venant's principle: A second update*, Appl. Mech. Rev., 49 (1996), pp. 101–111.

[13] C. O. Horgan and J. K. Knowles, *Recent developments concerning Saint-Venant's principle*, Adv. Appl. Mech., 23 (1983), pp. 179–269.

[14] C. O. Horgan, L. E. Payne, and L. T. Wheeler, *Spatial decay estimates in transient heat conduction*, Quart. Appl. Math., 42 (1984), pp. 119–127.

[15] C. O. Horgan and R. Quintanilla, *Spatial behaviour of solutions of the dual-phase-lag heat equation*, Math. Meth. Appl. Sci., 28 (2005), pp. 43–57.
[16] C. O. Horgan and R. Quintanilla, *Spatial decay of transient end effects for non-standard linear diffusion problems*, IMA J. Appl. Math., 70 (2005), pp. 119–128.
[17] S. Jiang and R. Racke, *Evolution Equations in Thermoelasticity*, π Monogr. Surv., Pure Appl. Math. 112. Chapman & Hall/CRC, Boca Raton, 2000.
[18] D. Jou, J. Casas-Vazquez, and G. Lebon, *Extended Irreversible Thermodynamics*, Springer-Verlag, Berlin, 1996.
[19] R. J. Knops and L. E. Payne, *Alternative spatial growth and decay for constrained motion in an elastic cylinder*, Math. Mech. Solids, in press.
[20] I. Müller and T. Ruggeri, *Rational and Extended Thermodynamics*, Springer-Verlag, New York, 1998.
[21] L. E. Payne, P. W. Schaefer, and J. C. Song, *Improved bounds for some nonstandard problems in generalized heat conduction*, J. Math. Anal. Appl., 298 (2004), pp. 325–340.
[22] L. E. Payne, P. W. Schaefer, and J. C. Song, *Some nonstandard problems in viscous flow*, Math. Meth. Appl. Sci., 27 (2004), pp. 2045–2053.
[23] R. Quintanilla, *Spatial bounds and growth estimates for the heat equation with three relaxation times*, Math. Meth. Appl. Sci., 20 (1997), pp. 1335–1344.
[24] R. Quintanilla, *End effects in thermoelasticity*, Math. Meth. Appl. Sci., 24 (2001), pp. 93–102.
[25] R. Quintanilla, *Exponential stability in the dual-phase-lag heat conduction theory*, J. Non-Equilibrium Thermodynamics, 27 (2002), pp. 217–227.
[26] R. Quintanilla, *A condition on the delay parameters in the one-dimensional dual-phase-lag thermoelastic theory*, J. Thermal Stresses, 26 (2003), pp. 713–721.
[27] R. Quintanilla, *A note on semigroup arguments in nonstandard problems*, J. Math. Anal. Appl., 310 (2005), pp. 690–698.
[28] R. Quintanilla and R. Racke, *Stability in thermoelasticity of type* III, Discrete Contin. Dyn. Syst., Ser. B, 3 (2003), pp. 383–400.
[29] R. Quintanilla and B. Straughan, *Explosive instabilities in heat transmission*, Proc. Roy. Soc. London Ser. A, 458 (2002), pp. 2833–2838.
[30] R. Quintanilla and B. Straughan, *Energy bounds for some non-standard problems in thermoelasticity*, Proc. Roy. Soc. London Ser. A, 461 (2005), pp. 1147–1162.
[31] R. Racke, *Thermoelasticity with second sound—exponential stability in linear and nonlinear* 1-d, Math. Meth. Appl. Sci., 25 (2002), pp. 409–441.
[32] R. Racke, *Asymptotic behavior of solutions in linear* 2- *or* 3-d *thermoelasticity with second sound*, Quart. Appl. Math., 61 (2003), pp. 315–328.
[33] J. C. Song, *Spatial decay for solutions of Cauchy problems for perturbed heat equations*, Math. Models Methods Appl. Sci., 11 (2001), pp. 797–808.
[34] D. Y. Tzou, *A unified approach for heat conduction from macro to micro-scales*, J. Heat Transfer, 117 (1995), pp. 8–16.

SIAM J. APPL. MATH.
Vol. 66, No. 3, pp. 1002–1026

ACOUSTIC SCATTERING BY MILDLY ROUGH UNBOUNDED SURFACES IN THREE DIMENSIONS*

SIMON N. CHANDLER-WILDE†, ERIC HEINEMEYER‡, AND ROLAND POTTHAST‡

Abstract. For a nonlocally perturbed half-space we consider the scattering of time-harmonic acoustic waves. A second kind boundary integral equation formulation is proposed for the sound-soft case, based on a standard ansatz as a combined single- and double-layer potential but replacing the usual fundamental solution of the Helmholtz equation with an appropriate half-space Green's function. Due to the unboundedness of the surface, the integral operators are noncompact. In contrast to the two-dimensional case, the integral operators are also strongly singular, due to the slow decay at infinity of the fundamental solution of the three-dimensional Helmholtz equation. In the case when the surface is sufficiently smooth (Lyapunov) we show that the integral operators are nevertheless bounded as operators on $L^2(\Gamma)$ and on $L^2(\Gamma) \cap BC(\Gamma)$ and that the operators depend continuously in norm on the wave number and on Γ. We further show that for *mild* roughness, i.e., a surface Γ which does not differ too much from a plane, the boundary integral equation is uniquely solvable in the space $L^2(\Gamma) \cap BC(\Gamma)$ and the scattering problem has a unique solution which satisfies a limiting absorption principle in the case of real wave number.

Key words. boundary integral equation method, rough surface scattering, Helmholtz equation

AMS subject classifications. 35J05, 35J25, 45E10, 45E99, 78A45

DOI. 10.1137/050635262

1. Introduction.

The simulation of scattering of acoustic or electromagnetic waves is of great importance for a large number of application areas ranging from medical imaging to seismic exploration. To carry out this simulation, boundary integral equation (BIE) methods have become very popular in recent decades. For scattering by bounded obstacles in two or three dimensions a very complete theory of the boundary integral equation method has been developed (e.g., [14, 20]), and the method forms the basis of very effective numerical algorithms (e.g., [12]).

This paper is concerned with the problem of scattering by *unbounded* surfaces for which the mathematical theory is much less well developed. More precisely, we are concerned with what are termed *rough surface scattering problems* in the engineering literature. We use the phrase *rough surface*, as is the practice in this literature, to denote a surface which is a (usually nonlocal) perturbation of an infinite plane surface such that the whole surface lies within a finite distance of the original plane. In particular we have in mind what is the usual case in the engineering literature where the scattering surface Γ is the graph of some bounded continuous function $f : \mathbb{R}^2 \to \mathbb{R}$, i.e.,

$$\Gamma := \big\{x = (x_1, x_2, x_3) \in \mathbb{R}^3 : x_3 = f(x_1, x_2)\big\}. \tag{1.1}$$

We will focus on a typical problem of this type, namely acoustic scattering by a rough, sound soft surface, the acoustic medium of propagation occupying the perturbed half-

*Received by the editors July 6, 2005; accepted for publication (in revised form) November 29, 2005; published electronically March 3, 2006.
http://www.siam.org/journals/siap/66-3/63526.html
†Department of Mathematics, University of Reading, Whiteknights, P.O. Box 220, Berkshire RG6 6AX, UK (s.n.chandler-wilde@reading.ac.uk).
‡Institute for Numerical and Applied Mathematics, University of Göttingen, Lotzestr. 16-18, 37083 Göttingen, Germany (potthast@scienceatlas.de, heinemey@math.uni-goettingen.de).

space

$$D := \{x = (x_1, x_2, x_3) : x_3 > f(x_1, x_2)\} \tag{1.2}$$

above the scattering surface Γ. The paper is concerned, particularly, with the theory of BIE methods for such problems, in the case when f is a sufficiently smooth function (Γ is Lyapunov).

Rough surface scattering problems arise frequently in applications, for example modelling acoustic and electromagnetic wave propagation over outdoor ground and sea surfaces or, at a very different scale, optical scattering from the surface of materials in nanotechnology. The mathematical and computational modelling of these problems has a large literature; see, e.g., the reviews and monographs by Ogilvy [22], Voronovich [28], Saillard and Sentenac [25], Warnick and Chew [29], and DeSanto [15]. The simulation of these scattering problems, requiring discretizations of sections of three-dimensional (3D) surfaces of diameter large compared to the wavelength, is a substantial scientific computing problem for which BIE methods are very popular, with many effective, specialized numerical algorithms developed [27, 25, 29, 31].

Although BIE methods are applied widely to rough surface scattering problems, the mathematical basis of the method is still poorly developed, especially in the 3D case. In fact, there are a number of severe difficulties in extending the theory of BIE methods from bounded to unbounded scatterers.

The first of these difficulties is that, due to the slow decay at infinity of the standard fundamental solution, $\Phi(x, y)$, of the Helmholtz equation (like $|x-y|^{-(n-1)/2}$ in n dimensions), the standard boundary integral operators are not bounded on any of the standard function spaces when the surface is unbounded. We will see that this difficulty can be overcome by modifying the usual kernels so as to obtain bounded integral operators and corresponding novel BIE formulations.

A second difficulty is that of loss of compactness of boundary integral operators associated with the noncompactness of the unbounded scattering surface. This is a severe barrier to establishing existence of solution to the BIEs. We recall that, in the case of scattering by smooth bounded obstacles, compactness arguments (the Riesz–Fredholm theory) lead directly to proofs of well-posedness for second kind boundary integral equation formulations (e.g., [14]). In the case of nonsmooth (Lipschitz) obstacles, compactness arguments are no longer sufficient but still play an essential role in establishing well-posedness (e.g., [26]).

For the two-dimensional (2D) rough surface scattering case much progress has been made in terms of deriving well-posed BIEs for a variety of acoustic, electromagnetic, and elastic wave problems [9, 8, 32, 2]. Surprisingly, none of the analysis for the 2D case extends straightforwardly to three dimensions; indeed most of the 2D analysis appears to be unsuitable in the 3D case.

In more detail, in the 2D case bounded integral operators have been obtained by replacing the standard fundamental solution by the Dirichlet or impedance Green's function for a half-plane that contains the domain D of propagation (see, e.g., [8, 32]). This modification leads to kernels of boundary integral operators that are weakly singular in their asymptotic behavior at infinity so that the integral operators are bounded on $L^p(\Gamma)$ for $1 \le p \le \infty$ and on $BC(\Gamma)$, the space of bounded continuous functions on Γ. In this paper we will employ the analogous modification for the 3D case, replacing the standard fundamental solution with the Dirichlet Green's function for a half-space that contains D. But this modification leads to kernels of the integral operators that are strongly rather than weakly singular. As a consequence,

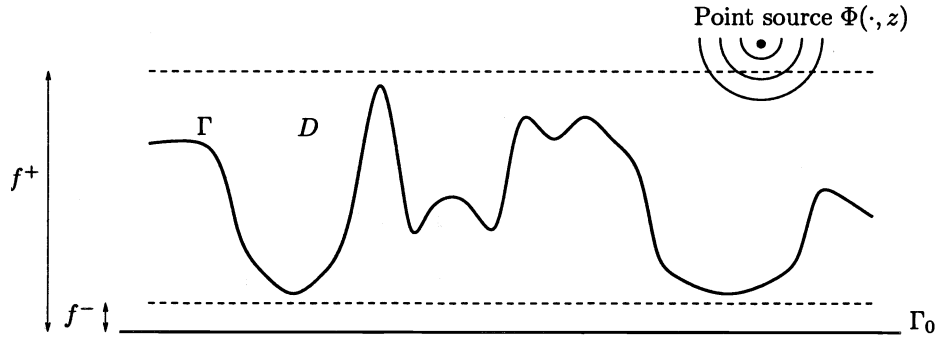


FIG. 1.1. *Geometrical setting of the scattering problem.*

the boundary integral operators are no longer well defined as operators on $BC(\Gamma)$ or $L^\infty(\Gamma)$. We are, however, able to show the boundedness of the operators on $L^2(\Gamma)$ by more elaborate arguments, expressing each integral operator as the sum of products of convolution and multiplication operators plus a well-behaved remainder, each multiplication operator a multiplication by an L^∞ function. To complete the proof of boundedness of the integral operators, one of the main results of the paper, we show, by explicit computations, that the Fourier transform of each convolution kernel is bounded. We note that our technique of expressing the kernel as the sum of products of convolution and multiplication operators plus a short-range remainder has been used previously [27, 31] but as a computational rather than a theoretical tool, as a device for matrix compression and acceleration of matrix-vector multiplications in iterative solvers.

To establish existence of solution and well-posedness in the 2D case, generalizations of part of the Riesz theory of compact operators have been developed [24, 11, 10] which require only local compactness rather than compactness and enable existence of solution in $BC(\Gamma)$ to be deduced from uniqueness of solution. In fact, injectivity of the second kind BIE in $BC(\Gamma)$ implies well-posedness in $BC(\Gamma)$ and in the space $L^p(\Gamma)$, $1 \leq p \leq \infty$ [3]. But this theory does not seem relevant for 3D rough surface scattering problems given that the corresponding boundary integral operators are not well defined as operators on $BC(\Gamma)$. In the absence of these tools we will prove existence of solution to the BIE (and the corresponding scattering problem) by perturbation arguments, used for the much simpler 2D case in [7]. The perturbation arguments we employ will prove to be sufficient to establish existence in the case when Γ is sufficiently close to a flat plane.

The results contained in this paper are as follows. We suppose that the rough surface is given by (1.1) with f continuously differentiable with Hölder continuous first derivative ($f \in BC^{1,\alpha}(\mathbb{R}^2)$ for some $\alpha \in (0, 1]$). As mentioned above, we investigate the mapping properties of the single- and double-layer potentials when the kernel is the Dirichlet Green's function for a half-space, consisting of the standard fundamental solution minus the same function with a point source mirrored in a plane. By *Fourier techniques* on a 2D plane and appropriate *decompositions* of the operators into a local and a global part we show that these layer potentials exist as bounded operators on $L^2(\Gamma)$. After these results, of significant interest in their own right, we consider the problem of acoustic scattering by the rough surface Γ in the case when the surface is sound soft (the field vanishes on Γ) and the incident field is due to a source distribution

with compact support in D. We reduce this scattering problem to a second kind boundary integral equation via an ansatz for the solution as a combined single- and double-layer potential. (The analogous ansatz was used for the 2D rough surface scattering case in [32], based on the analogous approach for scattering by bounded obstacles dating back to [4].) For a *flat surface*, unique solvability of the BIE is shown by explicit computation of a symbol. We then prove *continuous dependence* of the boundary integral operators on *variations of the boundary* and use these results and *perturbation arguments* to show that the scattering problem has a solution for all surfaces in a neighborhood of the plane, i.e., for *mildly rough* unbounded surfaces. Moreover, we show that the solution we compute by the BIE method satisfies a *limiting absorption principle*. For the convenience of the reader our main results are collected together and precisely stated at the end of section 2.

We should point out that a rigorous mathematical theory for BIE methods for 3D rough surface scattering has been developed previously for two special cases, both instances where the integral equation can be reduced to one on a finite domain so that compactness arguments can be applied. The first is the case of scattering by a locally perturbed plane, where the unbounded surface coincides with a plane in the exterior of some ball. This case can be reduced to a BIE on a finite domain, related to the local perturbation; we refer the reader to [30, 19, 5] and the references therein. The second is the case when the surface is a diffraction grating (the function f in (1.1) is biperiodic) and the incident field is a plane wave. In this case the BIE can be reduced to one on a finite part of the surface that is a single period; see [21, 17].

Finally, we note that, since our results assume boundary data in the space $L^2(\Gamma)$, they do not include the interesting and problematic case of plane wave incidence, which is included in the analogous theory that has been developed for the 2D problem [8, 32]. For a partial theoretical justification for BIE methods for 3D rough surface scattering with plane wave incidence, namely a justification, with some provisos, of Green's representation formula, see [16].

Notation. Throughout the paper x and y will denote points in $\mathbb{R}^3$ with components $x = (x_1, x_2, x_3)$ and $y = (y_1, y_2, y_3)$. The reflection of $y \in \mathbb{R}^3$ in the plane $\Gamma_0 := \{x \in \mathbb{R}^3 : x_3 = 0\}$ will be denoted by $y' := (y_1, y_2, -y_3)$. By $\mathbf{x}$ we will denote $(x_1, x_2) \in \mathbb{R}^2$, as well as the projection $(x_1, x_2, 0)$ of x onto the plane Γ_0. Similarly $\mathbf{y}$ denotes (y_1, y_2) and the projection of y onto Γ_0. The standard scalar product in $\mathbb{R}^2$ is denoted by $\mathbf{x} \cdot \mathbf{y}$ and $|\cdot|$ is the Euclidean norm in $\mathbb{R}^n$. Let $H^+ := \{z \in \mathbb{C} : \mathrm{Im} z \geq 0, \mathrm{Re} z > 0\}$. Given an unbounded closed set $S \subset \mathbb{R}^n$, $n = 2, 3$, $BC(S)$ will denote the set of bounded continuous real- or complex-valued functions on S, a Banach space with the norm $||\cdot||_{BC(S)}$ defined by $||F||_{BC(S)} = \sup_{x \in S} |F(x)|$. We will employ this notation particularly often in the cases $S = \Gamma \subset \mathbb{R}^3$ and $S = \mathbb{R}^2$. Similarly, for $0 < \alpha \leq 1$, let $BC^{1,\alpha}(\mathbb{R}^2)$ denote the set of those bounded continuously differentiable functions $F : \mathbb{R}^2 \to \mathbb{R}$ that have the property that ∇F is bounded and uniformly Hölder continuous with index α, so that

$$||F||_{BC^{1,\alpha}(\mathbb{R}^2)} := \sup_{\mathbf{x} \in \mathbb{R}^2} |F(\mathbf{x})| + \sup_{\mathbf{x} \in \mathbb{R}^2} |\nabla F(\mathbf{x})| + \sup_{\mathbf{x}, \mathbf{y} \in \mathbb{R}^2,\, \mathbf{x} \neq \mathbf{y}} \frac{|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})|}{|\mathbf{x} - \mathbf{y}|^\alpha} < \infty.$$

$BC^{1,\alpha}(\mathbb{R}^2)$ is a Banach space under the norm $||\cdot||_{BC^{1,\alpha}(\mathbb{R}^2)}$. It is convenient also to have a shorthand for the intersection of the sets $L^2(\Gamma)$ and $BC(\Gamma)$, so we define

$$X := L^2(\Gamma) \cap BC(\Gamma).$$

Since $L^2(\Gamma)$ and $BC(\Gamma)$ are Banach spaces equipped with their respective norms, so also is X, equipped with the norm $||\cdot||_X$ defined by

$$||F||_X := \max(||F||_{L^2(\Gamma)}, ||F||_{BC(\Gamma)}).$$

2. Scattering by rough surfaces in $\mathbb{R}^3$. Time-harmonic ($e^{-i\omega t}$ time dependence) acoustic waves are modelled by the Helmholtz equation

$$\triangle u + \kappa^2 u = 0, \tag{2.1}$$

where $\kappa = \kappa_0 + i\kappa_1$ denotes the *wave number* for which we will assume that $\kappa \in H^+$, i.e., $\kappa_0 > 0$ and $\kappa_1 \geq 0$. We define the *domain of propagation* D by (1.2), where $f \in BC^{1,\alpha}(\mathbb{R}^2)$ is a strictly positive function, so that there exist constants $f^+ > f^- > 0$ with

$$f^- \leq f(\mathbf{x}) \leq f^+, \quad \mathbf{x} \in \mathbb{R}^2.$$

We denote the boundary of D by Γ, so that Γ is given by (1.1). (For a sketch of the geometry see Figure 1.1.) Whenever we wish to denote explicitly the dependence of the domain on the boundary function f we will write D_f for D and Γ_f for Γ. This of course includes the case of the constant function $f \equiv h \in \mathbb{R}^+ := (0, \infty)$.

We will consider the scattering of an *incident acoustic wave* u^i by the surface Γ. For the *total field*

$$u := u^i + u^s, \tag{2.2}$$

which is the sum of the incident field and the *scattered field* u^s, we assume on Γ the *Dirichlet* boundary condition

$$u(x) = 0, \quad x \in \Gamma. \tag{2.3}$$

We require that the scattered field is bounded in D, i.e.,

$$|u^s(x)| \leq c, \quad x \in D, \tag{2.4}$$

for some constant $c > 0$. In the case $\kappa > 0$ we also require that u satisfies the following *limiting absorption principle*: denoting u temporarily by $u^{(\kappa)}$ to indicate its dependence on κ, we suppose that for all sufficiently small $\epsilon > 0$ a solution $u^{(\kappa+i\epsilon)}$ exists and that, for all $x \in D$,

$$u^{(\kappa+i\epsilon)}(x) \to u^{(\kappa)}(x), \quad \epsilon \to 0. \tag{2.5}$$

The limiting absorption principle plays the role of a radiation condition for real κ to single out the physical solution.

Before proceeding further to define the scattering problem precisely, we want to take a look, in the important case when the wave number is real, at the fundamental solution

$$\Phi(x, y) := \frac{1}{4\pi} \frac{e^{i\kappa|x-y|}}{|x-y|}, \quad x, y \in \mathbb{R}^3,\ x \neq y, \tag{2.6}$$

of the Helmholtz equation in $\mathbb{R}^3$ and the ordinary boundary layer potentials, e.g., the single-layer potential

$$\int_\Gamma \Phi(x, y)\varphi(y)\, ds(y), \quad x \in \mathbb{R}^3. \tag{2.7}$$

For an unbounded surface Γ the integral (2.7) converges (for $\kappa > 0$) only if φ decreases sufficiently rapidly at infinity. This is due to the slow decay of the fundamental solution in $\mathbb{R}^3$ at infinity. Letting $B_R(x) := \{y \in \mathbb{R}^3 : |x - y| < R\}$ denote the open ball of radius R centered at x, an easy calculation yields that, for $x \in D$, $p > 0$, and $\kappa > 0$,

$$\int_{\Gamma \cap B_R(x)} |\Phi(x,y)|^p \, ds(y) = \int_{\Gamma \cap B_R(x)} \frac{1}{|x-y|^p} \, ds(y) \to \infty, \quad R \to \infty,$$

for $p \leq 2$. Thus we observe that the trace of $\Phi(x, \cdot)$ on Γ is not integrable; indeed $\Phi(x, \cdot) \notin L^p(\Gamma)$ for $p \leq 2$. Thus, for every $x \in D$, the single-layer potential (2.7) is not well defined for all $\varphi \in L^2(\Gamma)$.

In order to get a faster decaying kernel we will, following what has been proposed for the analogous 2D rough surface scattering case [32], replace $\Phi(x, y)$ by an appropriate half-space Green's function for the Helmholtz equation. Specifically, we will work with the function

$$G(x,y) := \Phi(x,y) - \Phi(x,y'), \tag{2.8}$$

with $y' = (y_1, y_2, -y_3)$, which is the Dirichlet Green's function for the half-space $\{x : x_3 > 0\}$. Thus we will use layer potentials with $\Phi(x, y)$ replaced by $G(x, y)$, so that we define the *single-layer potential operator* by

$$(S\varphi)(x) := 2 \int_\Gamma G(x,y)\varphi(y) \, ds(y), \quad x \in \Gamma, \tag{2.9}$$

and the *double-layer potential operator* by

$$(K\varphi)(x) := 2 \int_\Gamma \frac{\partial G(x,y)}{\partial \nu(y)} \varphi(y) \, ds(y), \quad x \in \Gamma, \tag{2.10}$$

where the normal $\nu(y)$ is directed into D. Whenever we wish to denote explicitly the dependence of S and K on the boundary function f we will write S_f and K_f for S and K, respectively.

It is a straightforward calculation (cf. (3.7) below) to see that, for $y \in \Gamma$,

$$|G(x,y)| \sim \frac{x_3 y_3 |\kappa|}{2\pi} \, \frac{e^{-\kappa_1 |x-y|}}{|x-y|^2}, \quad |y| \to \infty. \tag{2.11}$$

This decay and the analogous decay at infinity (3.11) that we show for the kernel of the double-layer potential operator are fast enough for (2.9) and (2.10) to be well defined as improper integrals, for every $x \in \bar{D}$ and $\varphi \in C(\Gamma) \cap L^2(\Gamma)$, in particular in the case $\kappa_1 = 0$. Further, we will show, via Fourier techniques in section 5, as a main result of the paper, that this decay is fast enough for S and K to be bounded operators on $L^2(\Gamma)$.

Because, for $x \in \Gamma$,

$$\int_{\Gamma \cap B_R(x) \setminus B_1(x)} \frac{1}{|x-y|^2} \, ds(y) \to \infty, \quad R \to \infty,$$

the decay of $G(x, y)$ as $y \to \infty$ is not fast enough when $\kappa > 0$ for S to be well defined as an operator on the space of bounded continuous functions. Thus integral equation

methods for the 3D rough surface scattering problem are essentially different from the 2D case studied in [7, 9, 8, 32, 3].

Returning to the scattering problem, we wish to develop an analysis that is applicable whenever the incident wave is due to sources of the acoustic field located in some compact set $M \subset D$. Since waves with sources in a bounded set $M \subset \mathbb{R}^3$ can be represented as superpositions of point sources located in the same set, we will concentrate on the case when the incident field is due to a point source located at some point $z \in D$, i.e., $u^i = \Phi(\cdot, z)$. Thus the following is the specific problem that we will consider in this paper:

PROBLEM 1 (point source rough surface scattering problem). *Let $u^i = \Phi(\cdot, z)$ be the incident field due to a point source at $z \in D$. Then we seek a scattered field $u^s \in C^2(D) \cap C(\bar{D})$ such that u^s is a solution to the Helmholtz equation* (2.1) *in D, the total field satisfies the sound-soft boundary condition* (2.3)*, and the bound* (2.4) *holds. In the case $\kappa > 0$, we also require that the limiting absorption principle* (2.5) *holds.*

We will convert this scattering problem to a boundary value problem. To do this we will seek the scattered field as the sum of a mirrored point source $\Phi'(\cdot, z) := -\Phi(\cdot, z')$, where z' is the reflection of z in the flat plane Γ_0, plus some unknown remainder v, i.e., $u^s = v + \Phi'(\cdot, z)$. Note that $\Phi'(\cdot, z)$ is a solution to the scattering problem in the special case that $\Gamma = \Gamma_0$. Using the boundary condition $u^s + \Phi(\cdot, z) = 0$ on $\Gamma = \partial D$, we obtain the boundary condition on v that

$$v(x) = -\{\Phi(x, z) - \Phi(x, z')\} = -G(x, z) =: g(x), \quad x \in \Gamma. \tag{2.12}$$

Clearly, $g \in BC(\Gamma)$ and it follows from (2.11) that $g \in L^2(\Gamma)$, so that $g \in X = L^2(\Gamma) \cap BC(\Gamma)$. Thus u^s satisfies the above scattering problem if and only if v satisfies the following Dirichlet problem with g given by (2.12).

PROBLEM 2 (BVP). *Given $g \in X$, find $v \in C^2(D) \cap C(\bar{D})$, which satisfies the Helmholtz equation* (2.1) *in D, the Dirichlet boundary condition $v = g$ on Γ, the bound* (2.4)*, and, for $\kappa > 0$, the limiting absorption principle* (2.5)*.*

In this paper we will look for a solution to this boundary value problem as the *combined single- and double-layer potential*

$$v(x) := u_2(x) - i\eta\, u_1(x), \quad x \in D, \tag{2.13}$$

with some parameter $\eta \geq 0$, where for a given function $\varphi \in X$ we define the *single-layer potential*

$$u_1(x) := \int_\Gamma G(x, y)\varphi(y)\, ds(y), \quad x \in \mathbb{R}^3, \tag{2.14}$$

and the *double-layer potential*

$$u_2(x) := \int_\Gamma \frac{\partial G(x, y)}{\partial \nu(y)} \varphi(y)\, ds(y), \quad x \in \mathbb{R}^3. \tag{2.15}$$

Seeking the solution in this form we will see that the boundary condition (2.12) is satisfied if and only if the BIE

$$(I + K - i\eta S)\varphi = 2g \tag{2.16}$$

holds on Γ, where I is the identity operator.

The following are the main results in the remainder of the paper. In the next two sections we examine the kernels of the integral operators K and S and recall relevant properties of convolution and more general integral operators that we need to study K and S. The properties we discuss are exploited in section 5. In particular we show the following result.

THEOREM 2.1. *The single- and double-layer potential operators S and K, defined by* (2.9) *and* (2.10)*, are bounded operators on $L^2(\Gamma)$ and on X.*

We also establish that the single- and double-layer potential operators S and K depend continuously on $\kappa_1 = \operatorname{Im}\kappa$ in the norm topology on the set of bounded linear operators on $L^2(\Gamma)$; this result is needed to establish the limiting absorption principle. Moreover, we show, for S and K, continuous dependence in norm on the boundary Γ, in a sense we make precise.

In the final section, section 6, we establish existence and uniqueness of solution of the BIE and boundary value problem, at least in certain cases. As the first step we justify the integral equation (2.16) as a reformulation of the boundary value problem, showing the following result.

THEOREM 2.2. *Suppose that v is defined by* (2.13)–(2.15) *with $\varphi \in X$. Then, in the case $\kappa_1 > 0$, v satisfies the boundary value problem if and only if φ satisfies the BIE* (2.16)*. In the case $\kappa_1 = 0$ (i.e., $\kappa > 0$), if v satisfies the boundary value problem, then φ satisfies* (2.16)*. Conversely, if $\kappa > 0$, $\varphi^{(\kappa+i\epsilon)} \in X$ satisfies the integral equation* (2.16) *with κ replaced by $\kappa + i\epsilon$, for all sufficiently small $\epsilon > 0$, and $||\varphi - \varphi^{(\kappa+i\epsilon)}||_{L_2(\Gamma)} \to 0$ as $\epsilon \to 0$, then v satisfies the boundary value problem.*

We further establish the following result.

THEOREM 2.3. *The boundary value problem has at most one solution.*

Then we study the invertibility of the operator $I + K - i\eta S$, first for the case when Γ is flat and the operator $I + K - i\eta S$ is a convolution operator and then for the case when Γ is mildly rough by perturbation arguments. Our main result is the following.

THEOREM 2.4. *Suppose that $h > 0$ and that either $\eta > 0$ or $\eta = 0$ and $\operatorname{Im}\kappa = \kappa_1 > 0$. Then, provided $\|f - h\|_{BC^{1,\alpha}(\mathbb{R}^2)}$ is sufficiently small (so that Γ_f is sufficiently close to the flat surface $f \equiv h$), it holds that the integral equation* (2.16) *has a unique solution $\varphi \in L^2(\Gamma)$ for every $g \in L^2(\Gamma)$, so that $(I + K - i\eta S)^{-1}$ exists and is bounded as an operator on $L^2(\Gamma)$. If, further, $g \in X$, then $\varphi \in X$, so that $(I + K - i\eta S)^{-1}$ is also a bounded operator on X.*

Combining these results we have a final corollary concerning the solvability of the boundary value problem.

THEOREM 2.5. *If $h > 0$ and $\|f - h\|_{BC^{1,\alpha}(\mathbb{R}^2)}$ is sufficiently small (so that Γ_f is sufficiently close to the flat surface $f \equiv h$), then the boundary value problem has exactly one solution. Further, for some constant $c > 0$, independent of g,*

$$|v(x)| \le c\, ||g||_X, \quad x \in \bar{D}.$$

3. Properties of the 3D fundamental solution. We start with an investigation of properties of the fundamental solution $\Phi(x, y)$ and its derivatives. The key results are the expansions (3.7) and (3.11) needed to prove mapping properties of the boundary integral operators S and K in section 5.

For the first derivative of $\Phi(x, y)$ with respect to y_3 we calculate

$$\frac{\partial \Phi(x,y)}{\partial y_3} = -\frac{i\kappa}{4\pi}\frac{(x_3 - y_3)}{|x-y|^2}e^{i\kappa|x-y|} + \frac{1}{4\pi}\frac{(x_3 - y_3)}{|x-y|^3}e^{i\kappa|x-y|}. \tag{3.1}$$

The second derivative is given by

$$\frac{\partial^2\Phi(x,y)}{\partial y_3^2} = \frac{1}{4\pi}\left\{ i\kappa\frac{e^{i\kappa|x-y|}}{|x-y|^2} - \kappa^2\frac{(x_3-y_3)^2}{|x-y|^3}e^{i\kappa|x-y|} - 2i\kappa\frac{(x_3-y_3)^2}{|x-y|^4}e^{i\kappa|x-y|} \right.$$
$$\left. - \frac{e^{i\kappa|x-y|}}{|x-y|^3} - i\kappa\frac{(x_3-y_3)^2}{|x-y|^4}e^{i\kappa|x-y|} + 3\frac{(x_3-y_3)^2}{|x-y|^5}e^{i\kappa|x-y|} \right\}. \tag{3.2}$$

For the third derivative with respect to y_3 we obtain

$$\frac{\partial^3\Phi(x,y)}{\partial y_3^3} = \frac{3\kappa^2}{4\pi}\frac{(x_3-y_3)}{|x-y|^3}e^{i\kappa|x-y|} + O\left(\frac{1}{|x-y|^4}\right). \tag{3.3}$$

This holds in the sense that, given $c>0$ and a compact subset S of H^+, there exists a constant $C>0$ such that

$$\left|\frac{\partial^3\Phi(x,y)}{\partial y_3^3} - \frac{3\kappa^2}{4\pi}\frac{(x_3-y_3)}{|x-y|^3}e^{i\kappa|x-y|}\right| \le \frac{C}{|x-y|^4}$$

for all $x,\ y\in\mathbb{R}^3$, $x\neq y$, with $x_3,\ y_3\in[0,c]$ and all $\kappa\in S$. The similar equations below, in particular (3.7) and (3.11), are to be understood in an analogous fashion.

We use Taylor's expansion for the fundamental solution $\Phi(x,y)$ with respect to variations of x_3 and y_3. From Taylor's theorem, if $g\in C^3[0,\infty)$, then

$$g(s) = g(0) + g'(0)s + \frac{1}{2}g^{(2)}(0)s^2 + \frac{1}{3!}\int_0^s (s-t)^2 g^{(3)}(t)\,dt, \quad s>0. \tag{3.4}$$

Applying (3.4) to $g(s) := \Phi(x,\mathbf{y}+se_3)$, where e_3 is the unit vector in the x_3-direction, with $\mathbf{y}=(y_1,y_2,0)\in\Gamma_0$ and $s\in[0,c]$ with some constant c, we obtain

$$\Phi(x,\mathbf{y}+se_3) = \frac{1}{4\pi}\frac{e^{i\kappa|x-\mathbf{y}|}}{|x-\mathbf{y}|} - \frac{i\kappa}{4\pi}\frac{x_3\,e^{i\kappa|x-\mathbf{y}|}}{|x-\mathbf{y}|^2}\,s$$
$$+ \frac{i\kappa}{4\pi}\frac{e^{i\kappa|x-\mathbf{y}|}}{|x-\mathbf{y}|^2}\,\frac{s^2}{2} + O\left(\frac{1}{|x-\mathbf{y}|^3}\right). \tag{3.5}$$

To estimate the properties of single- and double-layer potentials on $L^2(\Gamma)$ we need to use Taylor's expansion also with respect to x_3. We treat all the terms of (3.5) separately and obtain, after some calculations,

$$\Phi(\mathbf{x}+he_3,\mathbf{y}+se_3) = \frac{1}{4\pi}\frac{e^{i\kappa|\mathbf{x}-\mathbf{y}|}}{|\mathbf{x}-\mathbf{y}|}$$
$$+ \frac{1}{4\pi}\frac{i\kappa\,e^{i\kappa|\mathbf{x}-\mathbf{y}|}}{|\mathbf{x}-\mathbf{y}|^2}\,\frac{(h-s)^2}{2} + O\left(\frac{1}{|\mathbf{x}-\mathbf{y}|^3}\right). \tag{3.6}$$

Altogether we obtain

$$G(\mathbf{x}+he_3,\mathbf{y}+se_3) = -\frac{1}{4\pi}\frac{i\kappa\,e^{i\kappa|\mathbf{x}-\mathbf{y}|}}{|\mathbf{x}-\mathbf{y}|^2}\,2hs + O\left(\frac{1}{|\mathbf{x}-\mathbf{y}|^3}\right), \tag{3.7}$$

in the sense that, given $c>0$ and a compact subset S of H^+, there exists a constant $C>0$ such that

$$\left|G(\mathbf{x}+he_3,\mathbf{y}+se_3) + \frac{2hs}{4\pi}\frac{i\kappa\,e^{i\kappa|\mathbf{x}-\mathbf{y}|}}{|\mathbf{x}-\mathbf{y}|^2}\right| \le \frac{C}{|\mathbf{x}-\mathbf{y}|^3}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$ with $\mathbf{x} \neq \mathbf{y}$, all $\kappa \in S$, and all $h, s \in [0, c]$. Arguing precisely as in [7] in the case $|x - y| > 1$, we can also show the bound that (cf. [7, equations (3.6) and (3.8)]), given a compact subset $S \subset H^+$, there exists a constant $C > 0$ such that

$$|G(x, y)| \leq \frac{C(1 + x_3)(1 + y_3)}{|x - y|^2} \tag{3.8}$$

for all $x, y \in \mathbb{R}^3$ with $x, y \neq 0$ and $x_3, y_3 \geq 0$ and all $\kappa \in S$.

For the normal derivative of G, noting that $\partial\Phi(x, y')/\partial\nu(y) = \partial\Phi(x', y)/\partial\nu(y)$ and introducing the notation $\boldsymbol{\nu}(y) := (\nu_1(y), \nu_2(y))$, we derive

$$\begin{aligned} 4\pi \frac{\partial G(x, y)}{\partial \nu(y)} = &- i\kappa\, \boldsymbol{\nu}(y) \cdot (\mathbf{x} - \mathbf{y}) \left\{ \frac{e^{i\kappa|x-y|}}{|x-y|^2} - \frac{e^{i\kappa|x-y'|}}{|x-y'|^2} \right\} \\ &+ \boldsymbol{\nu}(y) \cdot (\mathbf{x} - \mathbf{y}) \left\{ \frac{e^{i\kappa|x-y|}}{|x-y|^3} - \frac{e^{i\kappa|x-y'|}}{|x-y'|^3} \right\} \\ &- i\kappa \frac{\nu_3(y)(x_3 - y_3)}{|x-y|^2}\, e^{i\kappa|x-y|} + \frac{\nu_3(y)(x_3 - y_3)}{|x-y|^3}\, e^{i\kappa|x-y|} \\ &- i\kappa \frac{\nu_3(y)(x_3 + y_3)}{|x-y'|^2}\, e^{i\kappa|x-y'|} + \frac{\nu_3(y)(x_3 + y_3)}{|x-y'|^3}\, e^{i\kappa|x-y'|}. \end{aligned} \tag{3.9}$$

We proceed as in (3.6) and calculate

$$\frac{e^{i\kappa|x-y|}}{|x-y|^2} = \frac{e^{i\kappa|\mathbf{x}-\mathbf{y}|}}{|\mathbf{x}-\mathbf{y}|^2} + \frac{i\kappa e^{i\kappa|\mathbf{x}-\mathbf{y}|}}{|\mathbf{x}-\mathbf{y}|^3} \frac{(x_3 - y_3)^2}{2} + O\left(\frac{1}{|\mathbf{x}-\mathbf{y}|^4}\right). \tag{3.10}$$

We use this to transform (3.9) into

$$\begin{aligned} 4\pi \frac{\partial G(\mathbf{x} + he_3, \mathbf{y} + se_3)}{\partial \nu(y)} = &-\kappa^2 \boldsymbol{\nu}(y) \cdot \frac{(\mathbf{x} - \mathbf{y})}{|\mathbf{x} - \mathbf{y}|} \frac{e^{i\kappa|\mathbf{x}-\mathbf{y}|}}{|\mathbf{x}-\mathbf{y}|^2}\, 2hs \\ &- i\kappa\nu_3(y) \frac{e^{i\kappa|\mathbf{x}-\mathbf{y}|}}{|\mathbf{x}-\mathbf{y}|^2}\, 2h + O\left(\frac{1}{|\mathbf{x}-\mathbf{y}|^3}\right), \end{aligned} \tag{3.11}$$

this equation holding in the same sense as (3.7).

4. Convolution and related integral operators. To establish that S and K are bounded operators on $L^2(\Gamma)$ and on X we need tools from the theory of convolution operators and the Fourier and Hankel transforms. In this section we briefly recall the relevant results and compute explicitly certain Fourier transforms that we will need. The results in the first three paragraphs are contained, for example, in [23].

For $\ell \in L^1(\mathbb{R}^2) \cup L^2(\mathbb{R}^2)$ we define the Fourier transform of ℓ, $\mathcal{F}\ell$, by

$$(\mathcal{F}\ell)(\mathbf{k}) = \frac{1}{2\pi} \int_{\mathbb{R}^2} e^{-i\mathbf{k}\cdot\mathbf{y}} \ell(\mathbf{y})\, d\mathbf{y}, \quad \mathbf{k} \in \mathbb{R}^2. \tag{4.1}$$

In the case $\ell \in L^1(\mathbb{R}^2)$ the integral (4.1) exists in the ordinary Lebesgue sense, and $\mathcal{F}\ell \in BC(\mathbb{R}^2)$. If $\ell \in L^2(\mathbb{R}^2)$, then the integral (4.1) exists for almost all $\mathbf{k} \in \mathbb{R}^2$ as the limit

$$\lim_{R\to\infty} \frac{1}{2\pi} \int_{\mathbb{R}^2 \cap B_R(0)} e^{-i\mathbf{k}\cdot\mathbf{y}} \ell(\mathbf{y})\, d\mathbf{y},$$

and $\mathcal{F}\ell \in L^2(\mathbb{R}^2)$, with $||\mathcal{F}\ell||_{L^2(\mathbb{R}^2)} = ||\ell||_{L^2(\mathbb{R}^2)}$. Further, the mapping $\mathcal{F} : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2)$ is surjective and thus an isometric isomorphism. If $\ell \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$, then the two interpretations of (4.1) coincide and $\mathcal{F}\ell \in X_0 := L^2(\mathbb{R}^2) \cap BC(\mathbb{R}^2)$.

For $\ell, \psi \in L^2(\mathbb{R}^2)$ we define $\ell * \psi$, the *convolution* of ℓ and ψ, by

$$(\ell * \psi)(\mathbf{x}) := \int_{\mathbb{R}^2} \ell(\mathbf{x} - \mathbf{y})\psi(\mathbf{y})\, d\mathbf{y}, \quad \mathbf{x} \in \mathbb{R}^2.$$

The integral is defined in the ordinary Lebesgue sense, and $\ell * \psi \in BC(\mathbb{R}^2)$. If also $\mathcal{F}\ell \in L^\infty(\mathbb{R}^2)$, then $\ell * \psi \in X_0$, with

$$\ell * \psi = 2\pi\mathcal{F}^{-1}((\mathcal{F}\ell)(\mathcal{F}\psi)) \tag{4.2}$$

so that the *convolution operator* L, defined by $L\psi = \ell * \psi$, maps $L^2(\mathbb{R}^2)$ to $L^2(\mathbb{R}^2)$ and is bounded, with norm

$$||L||_{L^2(\mathbb{R}^2)\to L^2(\mathbb{R}^2)} \le 2\pi||\mathcal{F}\ell||_{L^\infty(\mathbb{R}^2)}.$$

We shall need in our arguments to also consider integral operators with kernels of a more general type. Suppose that $l : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{C}$ is such that $l(\mathbf{x}, \cdot)$ is measurable for all $\mathbf{x} \in \mathbb{R}^2$, and let L be the integral operator with kernel l, so that

$$(L\psi)(\mathbf{x}) = \int_{\mathbb{R}^2} l(\mathbf{x}, \mathbf{y})\psi(\mathbf{y})\, d\mathbf{y}, \quad \mathbf{x} \in \mathbb{R}^2. \tag{4.3}$$

One case of relevance to our later arguments is that in which

$$l(\mathbf{x}, \mathbf{y}) = m_1(\mathbf{x})\ell(\mathbf{x} - \mathbf{y})m_2(\mathbf{y}), \tag{4.4}$$

with $m_1, m_2 \in BC(\mathbb{R}^2)$, $\ell \in L^2(\mathbb{R}^2)$, $\mathcal{F}\ell \in L^\infty(\mathbb{R}^2)$. In this case, if $\psi \in L^2(\mathbb{R}^2)$, then (4.3) exists in the Lebesgue sense for all $\mathbf{x} \in \mathbb{R}^2$, $L\psi \in X_0$, and L is a bounded operator on $L^2(\mathbb{R}^2)$ with norm

$$||L||_{L^2(\mathbb{R}^2)\to L^2(\mathbb{R}^2)} \le 2\pi||m_1||_{BC(\mathbb{R}^2)}\, ||\mathcal{F}\ell||_{L^\infty(\mathbb{R}^2)}\, ||m_2||_{BC(\mathbb{R}^2)}. \tag{4.5}$$

Clearly, L is also a bounded operator on $L^2(\mathbb{R}^2)$ if it is a sum of operators of this form.

Another case of relevance is that in which

$$|l(\mathbf{x}, \mathbf{y})| \le \ell(\mathbf{x} - \mathbf{y}), \tag{4.6}$$

with $\ell \in L^p(\mathbb{R}^2)$, for some $p \in [1, \infty)$. In this case, if ψ is continuous and compactly supported, then (4.3) exists in the Lebesgue sense for all $\mathbf{x} \in \mathbb{R}^2$, $L\psi \in L^s(\mathbb{R}^2)$, for $s \ge 1$, and, from Young's inequality [23], it follows that

$$||L\psi||_{L^s(\mathbb{R}^2)} \le ||\ell||_{L^p(\mathbb{R}^2)}\, ||\psi||_{L^r(\mathbb{R}^2)}, \tag{4.7}$$

where $r^{-1} = 1 + s^{-1} - p^{-1}$. Since the set of continuous compactly supported functions is dense in $L^r(\mathbb{R}^2)$, we can extend the domain of L by density so that L is a bounded operator from $L^r(\mathbb{R}^2)$ to $L^s(\mathbb{R}^2)$ with norm $\le ||\ell||_{L^p(\mathbb{R}^2)}$. Further, if $\ell \in L^1(\mathbb{R}^2)$ and $\psi \in L^\infty(\mathbb{R}^2)$, then, trivially, (4.3) exists in the Lebesgue sense for all $\mathbf{x} \in \mathbb{R}^2$ and (4.7) holds.

We will use the bound (4.7) particularly often in the case $\ell \in L^1(\mathbb{R}^2)$, in which case it implies that

$$||L||_{L^q(\mathbb{R}^2)\to L^q(\mathbb{R}^2)} \le ||\ell||_{L^1(\mathbb{R}^2)}, \tag{4.8}$$

for $1 \le q \le \infty$. A further consequence of (4.7) is the following result that will be used to prove Lemma 5.1.

LEMMA 4.1. *Suppose that L is the integral operator given by* (4.3)*, and that the bound* (4.6) *holds with $\ell \in L^1(\mathbb{R}^2) \cap L^p(\mathbb{R}^2)$, for some $p \in (1,2)$. Then, for some $n \in \mathbb{N}$, L^n is a bounded operator from $L^2(\mathbb{R}^2)$ to $L^\infty(\mathbb{R}^2)$.*

Proof. Note first that $\ell \in L^1(\mathbb{R}^2) \cap L^p(\mathbb{R}^2)$ implies that $\ell \in L^{\tilde{p}}(\mathbb{R}^2)$ for $1 < \tilde{p} < p$. Let $\omega := p/(p-1) > 2$. Define the finite or infinite sequence (r_j) iteratively by

$$r_0 := 2, \quad r_{j+1} := \left(\frac{1}{p} + \frac{1}{r_j} - 1\right)^{-1} = \frac{r_j}{1 - \frac{r_j}{\omega}}, \quad j = 0, 1, 2, \ldots, \tag{4.9}$$

continuing the definition (4.9) for as long as $r_j < \omega$. Let $J \subset \mathbb{N}_0 := \mathbb{N} \cup \{0\}$ denote the set of indices j for which r_j is defined. We will show that the set J is finite, so that $J = \{0, 1, \ldots, N\}$ with $r_N \ge \omega$. Then, by (4.7), it follows that L is a bounded operator from $L^{r_{j-1}}(\mathbb{R}^2)$ to $L^{r_j}(\mathbb{R}^2)$ for $j = 1, \ldots, N$. Further, defining $\tilde{p} := r_N/(r_N - 1)$, we observe that $1 < \tilde{p} < p$ and $\frac{1}{\tilde{p}} + \frac{1}{r_N} = 1$, so that, by (4.7), L is a bounded operator from $L^{r_N}(\mathbb{R}^2)$ to $L^\infty(\mathbb{R}^2)$, and so L^{N+1} is a bounded operator from $L^2(\mathbb{R}^2)$ to $L^\infty(\mathbb{R}^2)$.

We complete the proof by showing that J is finite. Suppose otherwise. Then $r_j < \omega$ for all $j \in J = \mathbb{N}_0$. It follows from (4.9), by induction, that the sequence (r_j) is monotonically increasing. Thus the sequence (r_j) is convergent to some limit r, with $2 \le r \le \omega$. Rearranging (4.9) and taking limits, we see that $(1 - r/\omega)r = r$, so that $r = 0$, a contradiction. ∎

Examining (3.7) and (3.9) we see that large parts of the kernels of the operators S and K have the form (4.4), where, moreover, ℓ has certain symmetries that simplify the calculation of its Fourier transform. For the remainder of this section, for $\mathbf{y} \in \mathbb{R}^2$ let $r := |\mathbf{y}|$ and $\hat{\mathbf{y}} := \mathbf{y}/|\mathbf{y}|$. The specific symmetries that arise are those where ℓ has the form

$$\ell(\mathbf{y}) = F(r) Y_n^j(\hat{\mathbf{y}}), \tag{4.10}$$

with $n = 0$ or 1, and $j = 0, \ldots, n$, where the functions Y_n^j are spherical harmonics of order n defined on the unit circle $\Omega \subset \mathbb{R}^2$ by

$$Y_0^0(\hat{\mathbf{y}}) := 1, \quad Y_1^0(\hat{\mathbf{y}}) := \cos\theta, \quad Y_1^1(\hat{\mathbf{y}}) := \sin\theta, \quad \hat{\mathbf{y}} = (\cos\theta, \sin\theta) \in \Omega. \tag{4.11}$$

Integrating the product of the *Jacobi–Anger expansion* [13, equation (3.66)] (a Fourier series of $e^{ir\mathbf{k}\cdot\hat{\mathbf{y}}}$) with the spherical harmonics of order n over the unit circle we deduce the *Funk–Hecke formulae* in $\mathbb{R}^2$,

$$\int_\Omega e^{-ir\mathbf{k}\cdot\hat{\mathbf{y}}}\, Y_n^j(\hat{\mathbf{y}})\, ds(\hat{\mathbf{y}}) = 2\pi\, i^n\, J_n(rk)\, Y_n^j(\hat{\mathbf{k}}), \tag{4.12}$$

where we define $k := |\mathbf{k}|$ and $\hat{\mathbf{k}} := \mathbf{k}/k$ and J_n denotes the Bessel function of order n.

If $\ell \in L^1(\mathbb{R}^2) \cup L^2(\mathbb{R}^2)$ has the form (4.10), then it holds for almost all $\mathbf{k} \in \mathbb{R}^2$ that

$$\begin{aligned}(\mathcal{F}\ell)(\mathbf{k}) &= \lim_{R\to\infty} \frac{1}{2\pi} \int_{|\mathbf{y}|<R} e^{-i\mathbf{k}\cdot\mathbf{y}} \ell(\mathbf{y})\, d\mathbf{y} \\ &= \frac{1}{2\pi} \lim_{R\to\infty} \int_0^R \left(\int_\Omega e^{-ir\mathbf{k}\cdot\hat{\mathbf{y}}}\, Y_n^j(\hat{\mathbf{y}})\, ds(\hat{\mathbf{y}}) \right) F(r)\, r\, dr \\ &= i^n Y_n^j(\hat{\mathbf{k}}) \lim_{R\to\infty} \int_0^R F(r) J_n(kr)\, r\, dr \qquad (4.13) \\ &= \frac{i^n Y_n^j(\hat{\mathbf{k}})}{\sqrt{k}}\, \mathcal{H}_n\big(\sqrt{\cdot}\, F(\cdot)\big)(k), \qquad (4.14)\end{aligned}$$

where $\mathcal{H}_n$ denotes the *Hankel transform of order* n, i.e.,

$$(\mathcal{H}_n F)(k) := \lim_{R\to\infty} \int_0^R F(r)\, J_n(kr)\sqrt{kr}\, dr, \quad k \in \mathbb{R}^+, \quad n \in \mathbb{N}. \qquad (4.15)$$

We note that (4.13) can be used to extend the domain of the Fourier transform. Precisely, whenever ℓ has the form (4.10) and the limit (4.13) exists, (4.13) can be used to define a Fourier transform of ℓ, this definition coinciding with the usual one if $\ell \in L^1(\mathbb{R}^2) \cup L^2(\mathbb{R}^2)$. An example is the function ℓ defined, for some $h > 0$, by

$$\ell(\mathbf{y}) := W_h(|\mathbf{y}|), \quad W_h(r) := \frac{1}{4\pi} \frac{e^{i\kappa\sqrt{r^2+4h^2}}}{\sqrt{r^2+4h^2}}, \quad r > 0. \qquad (4.16)$$

The relevance of this example is that, defining $x_h := (0, 0, 2h)$, $\ell = \Phi(\cdot, x_h)$ is the trace of $\Phi(\cdot, x_h)$ on the plane Γ_0. It is not difficult to see, from the asymptotic behavior of the Bessel function J_n in (4.20), that, for $F = W_h$, the limit (4.13) is well defined except for $k = \kappa$ in the case $\kappa > 0$. Explicitly, from (4.14) with $n = 0$ and the Hankel transforms in section 8.2 of [18], namely formula (24) for $\kappa_1 > 0$ and formulae (41) and (50) for $\kappa_1 = 0$, we find after some elementary calculations that, for all $\mathbf{k} \in \mathbb{R}^2$ with $k = |\mathbf{k}| \neq \kappa$,

$$(\mathcal{F}\ell)(\mathbf{k}) = (\mathcal{F}\Phi(\cdot, x_h))(\mathbf{k}) = \int_0^\infty W_h(r) J_0(kr) r\, dr = \frac{1}{4\pi} \frac{e^{-2h\sqrt{k^2-\kappa^2}}}{\sqrt{k^2-\kappa^2}}, \qquad (4.17)$$

where the square root is chosen so that its argument lies in $[-\pi/2, 0]$.

We now use the representation (4.13) to calculate the Fourier transforms of parts of the kernels of the operators S and K. We suppose that ℓ is given by (4.10), with $n = 0$ or 1 and

$$F(r) := \frac{e^{i\kappa r}}{\beta + r^2}, \quad r \geq 0, \qquad (4.18)$$

for some $\beta > 0$. The relevance of this example to the operators S and K is that the explicitly written terms on the right-hand side of (3.7) and (3.9) all take the form (4.4) if ℓ is given by (4.10) and (4.18) with $\beta = 0$ and if $\mathbf{x} + he_3$ and $\mathbf{y} + se_3$ lie on Γ.

Clearly, $\ell \in L^2(\mathbb{R}^2)$ (for $\beta > 0$). We will show also that the Fourier transform of ℓ is bounded, so that the operation of convolution with ℓ is bounded on $L^2(\mathbb{R}^2)$. To this end we show that the improper integral

$$I(k) := \int_0^\infty F(r) J_n(kr)\, r\, dr \qquad (4.19)$$

is bounded on $[0, \infty)$.

With the help of the asymptotic expansion of the Bessel function (see, e.g., [1]),

$$J_n(z) = \sqrt{\frac{2}{\pi z}} \cos\left(z - \frac{n\pi}{2} - \frac{\pi}{4}\right)\left\{1 + O\left(\frac{1}{z}\right)\right\}, \quad |z| \to \infty, \tag{4.20}$$

we see that $e^{iz}J_n(z)$ is bounded in $0 \le \arg z \le \theta$ for every $\theta \in (0, \pi/2)$. Since $\mathrm{Re}(i(\kappa - k)z) = -(\kappa_0 - k)\mathrm{Im}z - \kappa_1 \mathrm{Re}z$ and $F(z)$ is a holomorphic function in $\mathrm{Re}\, z > 0$, we see that for $0 \le k < \kappa_0$ we may transform the integral

$$I(k) = \int_0^\infty \frac{e^{i(\kappa - k)z}}{\beta + z^2}\, e^{ikz} J_n(kz)\, z\, dz \tag{4.21}$$

into

$$I(k) = \int_\gamma \frac{e^{i(\kappa - k)z}}{\beta + z^2}\, e^{ikz} J_n(kz)\, z\, dz \tag{4.22}$$

with $\gamma = \{(1+i)t : t \ge 0\}$. This integral is bounded for $0 \le k \le \kappa_0/2$.

For $k \ge \kappa_0/2$ we can use (4.20) and that $J_n(z)$ is continuous and thus, by (4.20), bounded on $[0, \infty)$ to estimate that, for some constants C_1 and C_2,

$$\left| \int_0^\infty F(r) J_n(kr)\, r\, dr \right| \le C_1 + C_2 \int_1^\infty \frac{1}{r^{3/2}}\, dr. \tag{4.23}$$

We conclude that I is bounded on $[0, \infty)$ for $n = 0, 1$, so that, by (4.13), $\mathcal{F}\ell \in L^\infty(\mathbb{R}^2)$ for $n = 0, 1$.

We will be interested in the last section of the paper, in order to establish a limiting absorption principle, in the dependence of ℓ on κ_1. Denote, temporarily, ℓ and I by ℓ_{κ_1} and I_{κ_1} to indicate their dependence on κ_1. Then, from (4.22), since $e^{iz}J_n(z)$ is bounded on $\gamma = \{(1+i)t : t \ge 0\}$, we see that, for some constant $C > 0$,

$$|I_{\kappa_1}(k) - I_0(k)| \le C \int_0^\infty e^{-\kappa_0 t/2}(1 - e^{-\kappa_1 t})\, dt$$

for $0 \le k \le \kappa_0/2$, $\kappa_1 \ge 0$, so that $I_{\kappa_1}(k) \to I_0(k)$ as $\kappa_1 \to 0$, uniformly on $[0, \kappa_0/2]$. Similarly, using (4.21) and (4.20) (cf. (4.23)), we can show that $I_{\kappa_1}(k) \to I_0(k)$ as $\kappa_1 \to 0$, uniformly on $[\kappa_0/2, \infty]$. Thus the following lemma holds.

LEMMA 4.2. *If ℓ is given by* (4.10) *and* (4.18) *with $\beta > 0$ and $n = j = 0$, or $n = 1$ and $j = 0$, or 1, then $\mathcal{F}\ell \in L^\infty(\mathbb{R}^2)$ so that the convolution integral operator L, with kernel $\ell(\mathbf{x} - \mathbf{y})$, is a bounded operator on $L^2(\mathbb{R}^2)$. Further, denoting ℓ and L by ℓ_{κ_1} and L_{κ_1} to indicate their dependence on κ_1, we have that $||\mathcal{F}\ell_{\kappa_1} - \mathcal{F}\ell_0||_{L^\infty(\mathbb{R}^2)} \to 0$ as $\kappa_1 \to 0$, so that L_{κ_1} tends to L_0 in norm as $\kappa_1 \to 0$.*

5. Properties of single- and double-layer potentials. In this section we prove that the single- and double-layer operators are well defined when considered as operators on $L^2(\Gamma)$. We further investigate the jump relations for unbounded regions and show continuity properties of the boundary operators with respect to variations of the boundary.

We first prove Theorem 2.1, stated at the end of section 2. To prove this result we split the operators into a local and a global part with the help of an appropriate *cut-off function*. To this end let $\chi : [0, \infty) \to \mathbb{R}$ be a continuous function with

$$\chi(t) := \begin{cases} 0, & t < 1/2, \\ 1, & t \ge 1, \end{cases} \qquad \text{and} \qquad 0 \le \chi(t) \le 1 \quad \forall t \ge 0. \tag{5.1}$$

Let A with kernel a denote one of the operators S or K, respectively. We define the *global part*

$$(A_1\varphi)(x) := \int_\Gamma \chi(|x-y|)a(x,y)\varphi(y)\,ds(y), \quad x \in \Gamma, \tag{5.2}$$

and the *local part*

$$(A_2\varphi)(x) := \int_\Gamma \Big(1-\chi(|x-y|)\Big)a(x,y)\varphi(y)\,ds(y), \quad x \in \Gamma. \tag{5.3}$$

This yields the decomposition $A = A_1 + A_2$, and we can study the mapping properties of A_1 and A_2 as operators on $L^2(\Gamma)$ and on X separately. We start by proving the following lemma.

LEMMA 5.1. *A_2 is a bounded operator on $L^q(\Gamma)$ for $1 \le q \le \infty$, is a bounded operator from $L^\infty(\Gamma)$ to $BC(\Gamma)$, and is a bounded operator on X. Further, for some $n \in \mathbb{N}$, A_2^n is a bounded operator from $L^2(\Gamma)$ to X.*

Proof. The kernel a_2 of A_2 has compact support and is weakly singular. Precisely, since (cf. [7, equation (4.23)])

$$|\nu(y)\cdot(x-y)| \le |\mathbf{x}-\mathbf{y}|^{1+\alpha}||f||_{BC^{1,\alpha}(\mathbb{R}^2)}, \quad x,y \in \Gamma, \tag{5.4}$$

it holds in the double-layer case $A = K$ that, for some constant $C > 0$,

$$|a_2(x,y)| \le C\ell(\mathbf{x}-\mathbf{y}), \quad x,y \in \Gamma,\ x \ne y, \tag{5.5}$$

where

$$\ell(\mathbf{y}) := \begin{cases} |\mathbf{y}|^{\alpha-2}, & |\mathbf{y}| \le 1, \\ 0, & |\mathbf{y}| > 1. \end{cases} \tag{5.6}$$

The same bound holds (but is not sharp) in the single-layer case $A = S$.

Since $\ell \in L^p(\mathbb{R}^2)$, for $1 \le p < 2/(2-\alpha)$, we see from (4.8) that A_2 is a bounded operator on $L^q(\Gamma)$ for $1 \le q \le \infty$. Since $a_2(x,y)$ is also continuous for $x \ne y$, it follows, moreover, that A_2 maps $L^\infty(\Gamma)$ to $BC(\Gamma)$. Thus A_2 is a bounded operator on X. By Lemma 4.1, A_2^m is a bounded operator from $L^2(\Gamma)$ to $L^\infty(\Gamma)$, for some $m \in \mathbb{N}$, so that A_2^{m+1} is a bounded operator from $L^2(\Gamma)$ to X. ∎

We now consider the global part and prove a lemma on the mapping properties of A_1. Together, Lemmas 5.1 and 5.2 provide a proof of Theorem 2.1.

LEMMA 5.2. *A_1 is a bounded operator on $L^2(\Gamma)$ and is a bounded operator from $L^2(\Gamma)$ to X.*

Proof. From the decompositions (3.7) and (3.11) it follows that the kernel a_1 of A_1 can be written, in both the cases $A = S$ and $A = K$, in the form

$$a_1(x,y) = l^*(\mathbf{x},\mathbf{y}) + l(\mathbf{x},\mathbf{y}), \tag{5.7}$$

where l^* is a sum of terms each of the form (4.4), with $m_1, m_2 \in BC(\mathbb{R}^2)$ and ℓ given by (4.10) and (4.18) with $\beta = 1$, and with $n = 0$ or 1. Further, l^* can be chosen so that l satisfies the bound, for some constant $C > 0$,

$$|l(\mathbf{x},\mathbf{y})| \le C\tilde{\ell}(\mathbf{x}-\mathbf{y}), \quad \mathbf{x},\mathbf{y} \in \mathbb{R}^2, \tag{5.8}$$

where $\tilde{\ell}(\mathbf{y}) := (1+|\mathbf{y}|)^{-3}$, so that $\tilde{\ell} \in L^1(\mathbb{R}^2)$. In detail, in the case $A = S$ we see from (3.7) that an appropriate choice is to take

$$l^*(\mathbf{x},\mathbf{y}) = -\frac{i\kappa f(\mathbf{x}) f(\mathbf{y})}{2\pi} \frac{e^{i\kappa|\mathbf{x}-\mathbf{y}|}}{1+|\mathbf{x}-\mathbf{y}|^2}, \tag{5.9}$$

while, in the case $A = K$ we see from (3.11) that we can take

$$l^*(\mathbf{x},\mathbf{y}) = -\frac{\kappa^2 f(\mathbf{x}) f(\mathbf{y})}{2\pi} \boldsymbol{\nu}(y) \cdot \frac{\mathbf{x}-\mathbf{y}}{|\mathbf{x}-\mathbf{y}|} \frac{e^{i\kappa|\mathbf{x}-\mathbf{y}|}}{1+|\mathbf{x}-\mathbf{y}|^2} - \frac{i\kappa f(\mathbf{x})\nu_3(y)}{2\pi} \frac{e^{i\kappa|\mathbf{x}-\mathbf{y}|}}{1+|\mathbf{x}-\mathbf{y}|^2}. \tag{5.10}$$

It follows from (4.5) and Lemma 4.2 applied to the integral operator with kernel l^*, and (4.8) applied to the integral operator with kernel l, that A_1 is a bounded operator on $L^2(\Gamma)$.

Note, from the above representation (5.7), that also

$$|a_1(x,y)| \le \ell^*(\mathbf{x}-\mathbf{y}), \quad x, y \in \Gamma,$$

for some $\ell^* \in L^2(\mathbb{R}^2)$; this is true since $\tilde{\ell} \in L^2(\mathbb{R}^2)$ and since each term of (5.9) and (5.10) can be bounded in this way. It follows from (4.7) that A_1 maps $L^2(\Gamma)$ to $L^\infty(\Gamma)$; in fact, since also a_1 is continuous, it holds that A_1 maps $L^2(\Gamma)$ to $BC(\Gamma)$. ∎

Remark 5.3. For $C_2 > C_1 > 0$ let

$$B = B(C_1, C_2) := \big\{ f \in BC^{1,\alpha}(\mathbb{R}^2) : C_1 \le f(\mathbf{y}),\ \mathbf{y} \in \mathbb{R}^2, \quad \|f\|_{BC^{1,\alpha}(\mathbb{R}^2)} \le C_2 \big\}. \tag{5.11}$$

We note that, given $C_2 > C_1 > 0$ and $\kappa_0 > 0$, we can choose $C > 0$ such that the estimates (5.5) and (5.8) hold for all $f \in B$ and all $\kappa_1 \ge 0$. (For (5.8) this follows from (3.7) and (3.11).) This observation will be helpful in establishing continuous dependence of A on f and on κ_1.

Combining the above lemmas we deduce Theorem 2.1 and have also the following corollary.

COROLLARY 5.4. *For all sufficiently large $n \in \mathbb{N}$ it holds that A^n is a bounded map from $L^2(\Gamma)$ to X.*

As part of the proof of Theorem 2.2 we need to show that our modified single- and double-layer potentials u_1 and u_2, over the unbounded surface Γ, behave in a similar way to the corresponding standard layer potentials supported on a smooth bounded surface. This is done in the following theorem, in which $M := \{x : 0 < x_3 < f(\mathbf{x})\}$ denotes the region between Γ and Γ_0.

THEOREM 5.5. *Let u_1 and u_2 denote the single- and double-layer potentials with density $\varphi \in X$, defined by* (2.14) *and* (2.15)*, respectively. The following hold:*

(i) *For $n = 1, 2$, $u_n \in C^2(D \cup M)$ and $\Delta u_n + k^2 u_n = 0$ in $D \cup M$.*

(ii) *u_1 and u_2 can be continuously extended from D to $\bar{D}$ and from M to $\bar{M}$, with limiting values*

$$u_{1,\pm}(x) = \int_\Gamma G(x,y)\varphi(y)\, ds(y), \quad x \in \Gamma, \tag{5.12}$$

and

$$u_{2,\pm}(x) = \int_\Gamma \frac{\partial G(x,y)}{\partial \nu(y)} \varphi(y)\, ds(y) \pm \frac{1}{2}\varphi(x), \quad x \in \Gamma, \tag{5.13}$$

where $u_{n,\pm}(x) := \lim_{\epsilon\to 0+} u_n(x \pm \epsilon\nu(x))$*, for* $n = 1, 2$ *and* $x \in \Gamma$*, and* $\nu(x)$ *denotes the unit normal at* $x \in \Gamma$ *directed into* D*.*

(iii) *Given constants* $C_2 > C_1 > 0$ *and a compact subset* S *of* H^+*, there exists a constant* $C > 0$ *such that*

$$|u_n(x)| \le C||\varphi||_X, \quad x \in D \cup M,\ n = 1, 2, \tag{5.14}$$

for all $\varphi \in X$*,* $\kappa \in S$*, and* $f \in B = B(C_1, C_2)$*.*

(iv) *Given constants* $C_2 > C_1 > 0$ *and* $\epsilon > 0$ *and a compact subset* S *of* H^+*, there exists a constant* $C > 0$ *such that*

$$|u_n(x)| \le C||\varphi||_{L^2(\Gamma)}, \quad n = 1, 2, \tag{5.15}$$

for all $x \in D \cup M$ *with* $|x_3 - f(x_1, x_2)| > \epsilon$*, all* $\varphi \in X$*, all* $\kappa \in S$*, and all* $f \in B = B(C_1, C_2)$*.*

Proof. We first show that $u \in C(D \cup M)$ and establish (ii) and (iii). We use the cut-off function χ given by (5.1). Let u denote one of u_1 and u_2, and let a denote the kernel of u so that $a(x, y) := G(x, y)$ and $a(x, y) := \partial G(x, y)/\partial\nu(y)$ in the respective cases. We have, for $x \in D \cup M$, that

$$u(x) = \int_\Gamma \chi(|x - y|)a(x, y)\varphi(y)\, ds(y) + \int_\Gamma \big[1 - \chi(|x - y|)\big]a(x, y)\varphi(y)\, ds(y).$$

The first term has a continuous kernel that is bounded at infinity by the estimate (3.7) or (3.9) and, since $\varphi \in L^2(\Gamma)$, is continuous in $\{x : x_3 > 0\}$. The second term is clearly continuous in $D \cup M$; to see that it can be continuously extended up to Γ from above and below and to compute its limiting values we observe that, keeping x within some ball centered at some $x_0 \in \Gamma$, it holds that the integrand is supported in a finite patch of the surface. We can extend this surface patch to a bounded obstacle with boundary of class $C^{1,\alpha}$ and, since $\varphi \in C(\Gamma)$, use the jump relations for bounded obstacles as presented in [14].

To show that the first term satisfies the bound (5.14) we recall that G satisfies the bound (3.8) and point out that, by interior elliptic regularity estimates for solutions of the Helmholtz equation (e.g., [9, Lemma 2.7]), it follows that $\nabla_y G(x, y)$ satisfies the same bound for all $\kappa \in S$ (with a different constant C), provided $x_3, y_3 > 0$ and $|x - y| > 1/4$; further we calculate directly that C can be chosen so that this bound also holds for $0 < |x - y| \le 1/4$. Thus, for some constant $C' > 0$, whether a is the kernel of the single- or double-layer potential, it holds for all $\kappa \in S$ that

$$|\chi(|x - y|)a(x, y)| \le C' \frac{(1 + x_3)(1 + y_3)}{1 + |x - y|^2}, \quad x, y \in \mathbb{R}^3,\ x_3, y_3 \ge 0. \tag{5.16}$$

Applying the Cauchy–Schwarz inequality we have that the first term is bounded, for $x \in \{x : x_3 > 0\}$, by $C'(1 + f_+)I(x)\|\varphi\|_{L^2(\Gamma)}$, where

$$\begin{aligned}[I(x)]^2 &= (1 + x_3)^2 \int_\Gamma \frac{ds(y)}{(1 + |x - y|^2)^2} \\ &\le (1 + x_3)^2(1 + ||\nabla f||_{BC(\Gamma)})^{1/2} \int_{\mathbb{R}^2} \frac{d\mathbf{y}}{(1 + |\mathbf{x} - \mathbf{y}|^2 + (x_3 - f(\mathbf{y}))^2)^2}.\end{aligned}$$

Thus, for some constant $c > 0$ it holds, for all $x \in \{y : y_3 > 0\}$ and all $f \in B$, that $[I(x)]^2 \le cF(x_3)$, where

$$F(x_3) := (1 + x_3)^2 \int_0^\infty \frac{r\, dr}{(1 + x_3^2 + r^2)^2} = \frac{(1 + x_3)^2}{x_3^2} \int_0^\infty \frac{s\, ds}{(x_3^{-2} + 1 + s^2)^2}.$$

Clearly, F is bounded on $[0, \infty)$. Thus the first term satisfies the bound (5.14).

To treat the second term we argue analogously to the corresponding 2D case [7]. We remark that $1-\chi(|x-y|)$ is zero for $|x-y| \geq 1$. We consider only the double-layer case $u = u_1$. (The argument is similar but simpler in the single-layer case.) Directly from the definitions (see (3.9)) we see that there exists a constant $C > 0$ such that

$$|(1-\chi(|x-y|))a(x,y)| \leq C\frac{|\nu(y)\cdot(x-y)|}{|x-y|^3}, \quad x \in D \cup M,\ y \in \Gamma,$$

for all $\kappa \in S$ and all $f \in B$. Define $x^* := (\mathbf{x}, f(\mathbf{x})) \in \Gamma$, $\delta := |f(\mathbf{x}) - x_3|$, and note that, by the triangle inequality,

$$\left(|\mathbf{x}-\mathbf{y}|^2+\delta^2\right)^{1/2} \leq |x-y| + |f(\mathbf{x})-f(\mathbf{y})| \leq \left(1+||\nabla f||_{BC(\Gamma)}\right)|x-y|.$$

Using this inequality, and (5.4) to bound $|\nu(y)\cdot(x^*-y)|$, we see that, for some $C' > 0$,

$$|(1-\chi(|x-y|))a(x,y)| \leq C'\,\frac{|\mathbf{x}-\mathbf{y}|^{1+\alpha}+\delta}{\left(|\mathbf{x}-\mathbf{y}|^2+\delta^2\right)^{3/2}}, \quad x \in D \cup M,\ y \in \Gamma,$$

for all $\kappa \in S$ and $f \in B$. Thus, defining $C'' = C'\left(1+||\nabla f||_{BC(\Gamma)}\right)^{1/2}$, the second term is bounded by

$$C''\,||\varphi||_{BC(\Gamma)} \int_{|\mathbf{y}|<1} \frac{|\mathbf{y}|^{1+\alpha}+\delta}{\left(|\mathbf{y}|^2+\delta^2\right)^{3/2}}\,d\mathbf{y} \leq 2\pi C''\,||\varphi||_{BC(\Gamma)} \int_0^1 \frac{r^{1+\alpha}+\delta}{r^2+\delta^2}\,dr,$$

for all $\kappa \in S$ and $f \in B$, so that the second term satisfies the bound (5.14).

To establish (iv) we modify the argument used to show (iii). We have remarked above that both $G(x,y)$ and $\nabla_y G(x,y)$ satisfy the bound (3.8). Thus (cf. (5.16)), for every $\epsilon > 0$ there exists $C_\epsilon > 0$ such that

$$|a(x,y)| \leq C_\epsilon \frac{(1+x_3)(1+y_3)}{1+|x-y|^2} \tag{5.17}$$

for all $x, y \in \mathbb{R}^3$ with $x_3, y_3 \geq 0$ and $|x-y| \geq \epsilon$ and all $\kappa \in S$. Applying the Cauchy–Schwarz inequality, as in the proof of (5.14), we see that it holds, for some constant $C'_\epsilon > 0$, that

$$|u_n(x)| \leq C'_\epsilon(1+f_+)I(x)\|\varphi\|_{L^2(\Gamma)}, \quad n = 1, 2,$$

for all $x \in D \cup M$ with $|x_3 - f(x_1,x_2)| \geq \epsilon$ and all $\kappa \in S$ and $f \in B$. In view of the bound on $I(x)$ already shown above, we see that we have established (5.15).

We complete the proof by establishing (i). This is clear when φ is compactly supported. The general case follows from the density in $L^2(\Gamma)$ of the set of those elements of X that are compactly supported, from the bound (5.15), and from the fact that limits of uniformly convergent sequences of solutions of the Helmholtz equation satisfy the Helmholtz equation (e.g., [9, Remark 2.8]). □

We continue this section by proving that the single- and double-layer potential operators depend continuously on variations in the boundary Γ. In the statement of the following theorem, $B = B(C_1, C_2)$ is the set defined in Remark 5.3 for some constants $C_2 > C_1 > 0$. We use the notation A_f for either S or K defined on a surface Γ_f given by some $f \in B$. With the help of the isomorphism

$$I_f : L^2(\Gamma_f) \to L^2(\mathbb{R}^2), \quad (I_f\varphi)(\mathbf{y}) = \varphi((\mathbf{y}, f(\mathbf{y}))), \quad \mathbf{y} \in \mathbb{R}^2, \tag{5.18}$$

we associate A_f with the element $\tilde{A}_f = I_f A_f I_f^{-1}$ of the set of bounded linear operators on $L^2(\mathbb{R}^2)$ for each $f \in B$. Denoting the kernel of $\tilde{A}_f$ by a_f, we see that, where $x = (\mathbf{x}, f(\mathbf{x}))$, $y = (\mathbf{y}, f(\mathbf{y}))$, and $a(x,y) := G(x,y)$ or $a(x,y) := \partial G(x,y)/\partial\nu(y)$, in the respective cases $A_f = S$ and $A_f = K$, it holds that

$$a_f(\mathbf{x},\mathbf{y}) = a(x,y)J_f(\mathbf{y}), \quad J_f(\mathbf{y}) := \sqrt{1+|\nabla f(\mathbf{y})|^2}.$$

THEOREM 5.6. *The single- and double-layer potential operators depend continuously on the boundary Γ_f of the unbounded domain D_f in the sense that*

$$\sup_{\substack{f,g\in B \\ \|f-g\|_{BC^{1,\alpha}(\mathbb{R}^2)}\le\epsilon}} \|\tilde{A}_f - \tilde{A}_g\|_{L^2(\mathbb{R}^2)\to L^2(\mathbb{R}^2)} \to 0, \quad \epsilon \to 0. \tag{5.19}$$

Proof. Similarly to how we proceeded when proving Theorem 2.1, we decompose the operator $\tilde{A}_f - \tilde{A}_g$ into a global and a local part, i.e., $\tilde{A}_f - \tilde{A}_g = A_1 + A_2$ with A_1, A_2 defined similarly to (5.2) and (5.3). We now carry out the proof for the case of the single-layer operator. The necessary changes for the double-layer operator are straightforward.

The global operator. The kernel of the global operator A_1 is given by

$$a_1(\mathbf{x},\mathbf{y}) := \chi(|\mathbf{x}-\mathbf{y}|)[a_f(\mathbf{x},\mathbf{y}) - a_g(\mathbf{x},\mathbf{y})]. \tag{5.20}$$

We use the expansion (5.7) and (5.9), denoting l by l_f, to indicate its dependence on f. We obtain

$$\begin{aligned} a_1(\mathbf{x},\mathbf{y}) = {} & \frac{i\kappa}{2\pi}\frac{e^{i\kappa|\mathbf{x}-\mathbf{y}|}}{1+|\mathbf{x}-\mathbf{y}|^2}\Big\{f(\mathbf{x})[f(\mathbf{y})-g(\mathbf{y})] + [f(\mathbf{x})-g(\mathbf{x})]g(\mathbf{y})\Big\}J_f(\mathbf{y}) \\ & + \Big(l_f(\mathbf{x},\mathbf{y}) - l_g(\mathbf{x},\mathbf{y})\Big)J_f(\mathbf{y}) \\ & + \Big(\frac{i\kappa g(\mathbf{x})g(\mathbf{y})}{2\pi}\frac{e^{i\kappa|\mathbf{x}-\mathbf{y}|}}{1+|\mathbf{x}-\mathbf{y}|^2} + l_g(\mathbf{x},\mathbf{y})\Big)(J_f(\mathbf{y}) - J_g(\mathbf{y})) \end{aligned} \tag{5.21}$$

for $\mathbf{x},\mathbf{y} \in \mathbb{R}^2$, $\mathbf{x} \neq \mathbf{y}$.

The integral operator whose kernel is the first term of (5.21) can be bounded using Lemma 4.2 and (4.5). Similarly, the integral operator whose kernel is the last term of (5.21) can be bounded using Lemma 4.2, (4.5), (5.8), and (4.8), noting that Remark 5.3 guarantees the uniformity of (5.8) for $f \in B$. To bound the integral operator whose kernel is the second term of (5.21), we construct, for every $\eta \in (0,1)$, a function $\ell_\eta \in L^1(\mathbb{R}^2)$ such that

$$\Big|\Big(l_f(\mathbf{x},\mathbf{y}) - l_g(\mathbf{x},\mathbf{y})\Big)J_f(\mathbf{y})\Big| \le \ell_\eta(\mathbf{x}-\mathbf{y}), \quad \mathbf{x},\mathbf{y}\in\mathbb{R}^2, \tag{5.22}$$

whenever $f, g \in B$ and $\|f-g\|_{BC^{1,\alpha}(\mathbb{R}^2)}$ is sufficiently small, and such that $||\ell_\eta||_{L^1(\mathbb{R}^2)} \to 0$ as $\eta \to 0$, and then we use the estimate (4.8). Together, the bounds on the three parts of A_1 show (5.19) for the global part of the operator.

The construction of ℓ_η is as follows. We choose (possible by Remark 5.3) a constant $C > 0$ so that (5.8) holds for all $f \in B$. We choose another constant $C' > 0$ that is a bound for $||J_f||_{L^\infty(\mathbb{R}^2)}$ for $f \in B$. Then, where $\tilde{\ell} \in L^1(\mathbb{R}^2)$ is defined as in Lemma 5.2, we set

$$\ell_\eta(\mathbf{y}) := \begin{cases} \eta^3, & \eta < |\mathbf{y}| < \eta^{-1}, \\ 2C\,C'\tilde{\ell}(\mathbf{y}) & \text{otherwise.} \end{cases}$$

Clearly, this satisfies that $||\ell_\eta||_{L^1(\mathbb{R}^2)} \to 0$ as $\eta \to 0$. Since, for every $\eta \in (0,1)$, $|l_f(\mathbf{x},\mathbf{y}) - l_g(\mathbf{x},\mathbf{y})| \to 0$ as $\|f-g\|_{BC^{1,\alpha}(\mathbb{R}^2)} \to 0$, uniformly in f and g for $f,g \in B$, and uniformly in $\mathbf{x}$ and $\mathbf{y}$ for $\eta \le |\mathbf{x}-\mathbf{y}| \le \eta^{-1}$, the bound (5.22) holds for all $f,g \in B$ with $\|f-g\|_{BC^{1,\alpha}(\mathbb{R}^2)}$ sufficiently small.

The local operator. For the local operator we argue in a similar way as for the global operator, in particular in a similar way as for the integral operator corresponding to the second term in (5.21). In particular, where a_2 is the kernel of the local operator, it holds for every $\eta > 0$ that $|a_2(\mathbf{x},\mathbf{y})| \to 0$ as $\|f-g\|_{BC^{1,\alpha}(\mathbb{R}^2)} \to 0$, uniformly in f and g for $f,g \in B$, and uniformly in $\mathbf{x}$ and $\mathbf{y}$ for $|\mathbf{x}-\mathbf{y}| \ge \eta$, and (5.5) takes the role of (5.8). ∎

We have just established continuous dependence of the single- and double-layer potential operators on the boundary Γ_f. To show later that the limiting absorption condition (2.5) is satisfied in the case $\kappa > 0$ we need to also establish continuous dependence on κ, which we do by similar arguments.

LEMMA 5.7. *Denote S and K temporarily by S_{κ_1} and K_{κ_1} to indicate their dependence on κ_1. Then, where A_{κ_1} denotes either S_{κ_1} or K_{κ_1}, it holds that*

$$||A_{\kappa_1} - A_0||_{L^2(\Gamma)\to L^2(\Gamma)} \to 0 \tag{5.23}$$

as $\kappa_1 \to 0$.

Proof. As we did when proving Theorem 2.1 we split A_{κ_1} into global and local parts, as $A_{\kappa_1} = A_1 + A_2$, with A_1, A_2 defined by (5.2) and (5.3). As in the proofs of Lemma 5.1 and 5.2 we denote the kernel of A_j by a_j.

To show (5.23) for the local part A_2 we note that $a_2(x,y)$ depends continuously on κ_1, uniformly in x and y for $|x-y| \ge \eta$ and every $\eta > 0$, and that, by Remark 5.3, the bound (5.5) holds uniformly in κ_1 for $\kappa_1 \in [0,1]$. We then argue as for the local part in the proof of Theorem 5.6, showing that the kernel of the local part of $A_{\kappa_1} - A_0$ is bounded by an L^1 convolution kernel $\ell(\mathbf{x}-\mathbf{y})$ with $||\ell||_{L^1(\mathbb{R}^2)} \to 0$ as $\kappa_1 \to 0$. Finally, we apply (4.8).

To show (5.23) for the global part A_2 we use the representation (5.7) for $a_1(x,y)$, which splits a_1 into a weakly singular part $l(\mathbf{x},\mathbf{y})$, bounded by (5.8), and a strongly singular part $l^*(\mathbf{x},\mathbf{y})$, given explicitly by (5.9) or (5.10). To show (5.23) for the weakly singular part of A_2 we argue exactly as we did in the proof of Theorem 5.6, noting that, by Remark 5.3, (5.8) holds uniformly in κ_1 for $\kappa_1 \in [0,1]$, and that $l(\mathbf{x},\mathbf{y})$ depends continuously on κ_1, uniformly in $\mathbf{x}$ and $\mathbf{y}$ for $\eta \le |\mathbf{x}-\mathbf{y}| \le \eta^{-1}$, for every $\eta \in (0,1)$. That (5.23) holds for the strongly singular part of A_2 follows from Lemma 4.2 and (4.5). ∎

6. Uniqueness and existence results. In this section we prove, for the case when the surface is mildly rough, uniqueness and existence for our integral equation formulation and for the boundary value problem and the scattering problem defined in section 2. As the first step in this argument we prove Theorem 2.2 on the equivalence of the integral equation (2.16) and the boundary value problem.

Proof of Theorem 2.2. Let v be the combined single- and double-layer potential v, defined in (2.13), with density $\varphi \in X$. By Theorem 5.5, $v \in C^2(D) \cap C(\bar{D})$ and satisfies the Helmholtz equation in D. Further, due to the jump relations (5.12) and (5.13), $v = g \in X$ on Γ if and only if the density φ satisfies the boundary integral equation (2.16). Applying Theorem 5.5 again, we see that v satisfies the bound (2.4). This yields the equivalence statement for $\kappa_1 > 0$.

For real κ, in addition, we need to show the limiting absorption principle (2.5).

Let $a(x,y) = \partial G(x,y)/\partial\nu(y) - i\eta G(x,y)$, so that

$$v(x) = \int_\Gamma a(x,y)\varphi(y)\,ds(y), \quad x \in D. \tag{6.1}$$

Suppose, as stated in the theorem, that $\varphi^{(\kappa+i\epsilon)} \in X$ satisfies the integral equation (2.16) with κ replaced by $\kappa + i\epsilon$, for all sufficiently small $\epsilon > 0$, and that $||\varphi - \varphi^{(\kappa+i\epsilon)}||_{L_2(\Gamma)} \to 0$ as $\epsilon \to 0$. Let $a^{(\kappa+i\epsilon)}$ denote a with κ replaced with $\kappa + i\epsilon$, and define $v^{(\kappa+i\epsilon)}$ by (6.1) with a, φ replaced by $a^{(\kappa+i\epsilon)}$, $\varphi^{(\kappa+i\epsilon)}$, respectively. We have shown in the previous paragraph that $v^{(\kappa+i\epsilon)}$ satisfies Problem 2 (with κ replaced by $\kappa + i\epsilon$). To show the limiting absorption principle (2.5) we need to show that $v^{(\kappa+i\epsilon)}(x) \to v(x)$ as $\epsilon \to 0$. We have

$$\begin{aligned} v^{(\kappa+i\epsilon)}(x) - v(x) = &\int_\Gamma \left(a^{(\kappa+i\epsilon)}(x,y) - a(x,y)\right)\varphi^{(\kappa+i\epsilon)}(y)\,ds(y) \\ &+ \int_\Gamma a(x,y)\left(\varphi^{(\kappa+i\epsilon)}(y) - \varphi(y)\right)\,ds(y). \end{aligned}$$

We see that the second term tends to zero as $\epsilon \to 0$ by the bound (5.15). Clearly, $a^{(\kappa+i\epsilon)}(x,y) - a(x,y) \to 0$ as $\epsilon \to 0$ for every $y \in \Gamma$. Thus, applying the Cauchy–Schwarz inequality and then the dominated convergence theorem, noting that the bound (5.17) holds uniformly in κ, we see that the first term tends to 0 as $\epsilon \to 0$. ∎

We obtain uniqueness of solution for the boundary value problem (proving Theorem 2.3) as follows. Due to [6, Theorem 1] (see also [24, Theorem 3.1]), a solution $u \in C^2(G) \cap C(\overline{G})$ to the Helmholtz equation (2.1) with $\mathrm{Im}(\kappa) > 0$ on an open set $G \subset \mathbb{R}^n$ which satisfies the growth condition $|u(x)| \le Ce^{\theta|x|}$, with some constant $\theta < \mathrm{Im}(\kappa)$, and the boundary condition $u(x) = 0$ for $x \in \partial G$ will vanish identically on G. This result directly implies uniqueness for the scattering problem and the boundary value problem for $\kappa_1 > 0$. For $\kappa_1 = 0$ uniqueness is a consequence of the limiting absorption principle we require, i.e., of the convergence (2.5).

Next we turn to establishing existence of solution in the mildly rough case. But first we prove a preliminary lemma which shows that to establish unique solvability of the integral equation in the space X it is enough to study solvability in $L^2(\Gamma)$.

LEMMA 6.1. *Suppose that the integral equation* (2.16) *has exactly one solution* $\varphi \in L^2(\Gamma)$ *for every* $g \in L^2(\Gamma)$. *Then also* (2.16) *has exactly one solution* $\varphi \in X$ *for every* $g \in X$, *so that* $(I + K - i\eta S)^{-1}$ *exists and is bounded as an operator on* X.

Proof. If the assumptions of the lemma hold, then (2.16) has exactly one solution $\varphi \in L^2(\Gamma)$ for every $g \in X \subset L^2(\Gamma)$. Further, defining $A = K - i\eta S$, it holds that $\varphi = A\varphi + 2g$ and, by induction, that, for every $n \in \mathbb{N}$,

$$\varphi = A^n\varphi + 2(A^{n-1} + \cdots + A^0)g.$$

Now, by Theorem 2.1, A is a bounded operator on X and, by Corollary 5.4, A^n is a bounded operator from $L^2(\Gamma)$ to X for some $n \in \mathbb{N}$. Thus $\varphi \in X$. We have shown that (2.16) has exactly one solution $\varphi \in X$ for every $g \in X$, so that $(I + K - i\eta S)^{-1}$ exists as an operator on X. Since X is a Banach space it follows as a standard corollary of the open mapping theorem that $(I + K - i\eta S)^{-1}$ is bounded. ∎

As a corollary of Theorems 2.2 and 2.3 and Lemmas 5.7 and 6.1 we have the following result.

COROLLARY 6.2. *If $(I+K-i\eta S)^{-1}$ exists as a bounded operator on $L^2(\Gamma)$, then the boundary value problem and scattering problem have exactly one solution.*

Proof. In the case $\kappa_1 > 0$ this result is clear from Theorems 2.2 and 2.3 and Lemma 6.1.

In the case $\kappa_1 = 0$ we note that, by Lemma 5.7 and standard operator perturbation arguments (e.g., [23]), if $(I+K-i\eta S)^{-1}$ exists as a bounded operator on $L^2(\Gamma)$ for $\kappa = \kappa_0 > 0$, then $(I+K-i\eta S)^{-1}$ exists and is a bounded operator on $L^2(\Gamma)$ for $\kappa = \kappa_0 + i\kappa_1$, $0 \le \kappa_1 \le c$, for some $c > 0$. Moreover, $(I+K-i\eta S)^{-1}$ depends continuously in the norm topology on κ_1 for $\kappa_1 \in [0,c]$. Thus, provided $g \in L^2(\Gamma)$ depends continuously in norm on κ_1, for $\kappa_1 \in [0,c]$, it holds that $(I+K-i\eta S)^{-1}g$ depends continuously in norm on $\kappa_1 \in [0,c]$ in $L^2(\Gamma)$. If g is given by (2.12), then, from the continuity of $\Phi(x,y)$ as a function of κ_1, uniformly in $x, y \in \mathbb{R}^3$, $x \neq y$, the bound (3.8), and the dominated convergence theorem, it follows that $g \in L^2(\Gamma)$ depends continuously in norm on κ_1 for $\kappa_1 \in [0,c]$. Thus the result follows by Theorems 2.2 and 2.3 and Lemma 6.1. ∎

To make use of this result it remains to establish that $(I+K-i\eta S)$ is invertible as an operator on $L^2(\Gamma)$. We will show this first for the case of a flat surface $\Gamma_h = \{y = (\mathbf{y}, h) : \mathbf{y} \in \mathbb{R}^2\}$, with $h > 0$. In this case the kernels of K and S depend only on the difference $\mathbf{x} - \mathbf{y}$, and thus, identifying Γ_h with $\mathbb{R}^2$, the operators are convolution operators on $L^2(\mathbb{R}^2)$.

In terms of the function W_h defined by (4.16), it follows from (3.9) that we can write the kernel of the double-layer potential operator as $P_h(\mathbf{x}-\mathbf{y})$, where $P_h(\mathbf{y}) := p_h(|\mathbf{y}|)$ and

$$p_h(r) := -\frac{i\kappa h}{\pi}\frac{e^{i\kappa\sqrt{r^2+4h^2}}}{(\sqrt{r^2+4h^2}\,)^2} + \frac{h}{\pi}\frac{e^{i\kappa\sqrt{r^2+4h^2}}}{(\sqrt{r^2+4h^2}\,)^3} = -\frac{\partial}{\partial h}\{W_h(r)\}, \quad r > 0.$$

The kernel of the single-layer potential operator is $Q_h(\mathbf{x}-\mathbf{y})$, where $Q_h(\mathbf{y}) := q_h(|\mathbf{y}|)$ and

$$q_h(r) := 2\left\{\frac{1}{4\pi}\frac{e^{i\kappa r}}{r} - \frac{1}{4\pi}\frac{e^{i\kappa\sqrt{r^2+4h^2}}}{\sqrt{r^2+4h^2}}\right\} = 2W_0(r) - 2W_h(r), \quad r > 0.$$

Hence, the integral equation (2.16) is transformed into

$$\varphi(\mathbf{x}) + \int_{\mathbb{R}^2} \big\{P_h(\mathbf{x}-\mathbf{y}) - i\eta\, Q_h(\mathbf{x}-\mathbf{y})\big\}\varphi(\mathbf{y})\, d\mathbf{y} \;=\; 2g(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^2. \tag{6.2}$$

To prove invertibility in $L^2(\Gamma_h)$ we compute the 2D Fourier transforms of P_h and Q_h. From (4.13), for almost all $\mathbf{k} \in \mathbb{R}^2$, where $k := |\mathbf{k}|$,

$$(\mathcal{F}P_h)(\mathbf{k}) = -\int_0^\infty \frac{\partial W_h(r)}{\partial h} J_0(kr) r\, dr.$$

To evaluate this integral we reverse the order of integration and differentiation and use (4.17) to get that, for almost all $\mathbf{k} \in \mathbb{R}^2$,

$$(\mathcal{F}P_h)(\mathbf{k}) = -\frac{\partial}{\partial h}\left\{\frac{1}{4\pi}\frac{e^{-2h\sqrt{k^2-\kappa^2}}}{\sqrt{k^2-\kappa^2}}\right\} = \frac{1}{2\pi}\, e^{-2h\sqrt{k^2-\kappa^2}}. \tag{6.3}$$

The interchange of integration and differentiation with respect to h is certainly justified whenever $k > 0$ and $k \neq \kappa$. For then the integral (4.17) is well defined, and,

using (4.20), we see that for every $H > 0$ there exists a constant $C > 0$ such that

$$\left|\frac{\partial W_h(r)}{\partial h} J_0(kr) r\right| = \left|\frac{\partial}{\partial h}\left\{\frac{1}{4\pi}\frac{e^{i\kappa\sqrt{r^2+4h^2}}}{\sqrt{r^2+4h^2}}\right\} J_0(kr) r\right| \le \frac{C}{r^{3/2}} \tag{6.4}$$

for $r \ge 1$ and $0 \le h \le H$.

The Fourier transform of Q_h can also be evaluated using (4.17). We obtain that

$$(\mathcal{F}Q_h)(\mathbf{k}) = \frac{1}{2\pi}\left\{\frac{1}{\sqrt{k^2-\kappa^2}} - \frac{e^{-2h\sqrt{k^2-\kappa^2}}}{\sqrt{k^2-\kappa^2}}\right\} \tag{6.5}$$

for all $\mathbf{k} \in \mathbb{R}^2$ with $k \ne \kappa$. We combine the Fourier transforms of P_h and Q_h to derive for the Fourier transform of the kernel $R_h := P_h - i\eta\, Q_h$ of $K - i\eta S$ the formula $(\mathcal{F}R_h)(\mathbf{k}) = \hat{r}_h(k)$, for almost all $\mathbf{k} \in \mathbb{R}^2$, where

$$\hat{r}_h(k) := \frac{1}{2\pi}\left\{e^{-2h\sqrt{k^2-\kappa^2}} - i\eta\,\frac{1 - e^{-2h\sqrt{k^2-\kappa^2}}}{\sqrt{k^2-\kappa^2}}\right\}, \quad k \ge 0. \tag{6.6}$$

From (4.2) we see that, for $\psi \in L^2(\mathbb{R}^2)$,

$$(I + K - i\eta S)\psi = \mathcal{F}^{-1}((1 + 2\pi\mathcal{F}R_h)(\mathcal{F}\psi)).$$

Since $\mathcal{F}$ is an isomorphism on $L^2(\mathbb{R}^2)$ it follows that the inverse of $I + K - i\eta S$ exists as a bounded operator from $L^2(\Gamma_h)$ into $L^2(\Gamma_h)$ if and only if

$$\operatorname{ess.}\inf_{\mathbf{k}\in\mathbb{R}^2} |1 + 2\pi\,(\mathcal{F}R_h)(\mathbf{k})| = \inf_{k\ge 0} |1 + 2\pi\,\hat{r}_h(k)| > 0. \tag{6.7}$$

We need to investigate $K(k) := 1 + 2\pi\,\hat{r}_h(k) = A(h\sqrt{k^2-\kappa^2})$, for $k \ge 0$, where

$$A(z) := 1 + e^{-2z} - \frac{ih\eta}{z}\left(1 - e^{-2z}\right). \tag{6.8}$$

We recall that (see (4.17)) the square root is to be taken with $\sqrt{k^2-\kappa^2} \in V := \{z \in \mathbb{C} : \mathrm{Re}z \ge 0, \mathrm{Im}z \le 0\}$. Indeed, in the case that $\kappa_1 > 0$, so that $\mathrm{Im}(k^2-\kappa^2) < 0$, it is clear that $\sqrt{k^2-\kappa^2}$ lies in the interior of V. Now A is an entire function (the singularity at 0 is removable) so that K is continuous on $[0,\infty)$. Further, $K(k) \to 1$ as $k \to \infty$. Thus, to show (6.7) it is enough to show that $K(k) \ne 0$ for $k \ge 0$, which holds if $A(z) \ne 0$ for $z \in V$; indeed, in the case $\kappa_1 > 0$, we need only show that $A(z) \ne 0$ for all z in the interior of V.

So suppose $\eta \ge 0$, and consider first the case when $z = z_0 - iz_1$, with $z_0 > 0$, $z_1 \ge 0$. It holds that

$$A(z) = -i\,\left(1 + e^{-2z}\right)\left(\frac{h\eta\tanh z}{z} + i\right),$$

and straightforward calculations yield

$$\mathrm{Im}\left(\frac{\tanh z}{z}\right) = \frac{z_0\sin(2z_1) + z_1\sinh(2z_0)}{2[\sinh^2 z_0 + \cos^2 z_1](z_0^2 + z_1^2)} \ge 0,$$

since $|\sin t| \le t \le \sinh t$ for $t \ge 0$. Thus (6.7) holds if $\eta \ge 0$ and $\kappa_1 > 0$.

In the case $\kappa_1 = 0$ we need to show, additionally, that $A(z) \neq 0$ when $z = -iz_1$ with $z_1 \geq 0$, in order to establish that $A(z) \neq 0$ for all $z \in V$. Now $A(0) = 2 - 2i\eta h$ and, for $z_1 > 0$, from (6.8), $A(-iz_1) = 2\cos z_1 - \frac{2ih\eta}{z_1}\sin z_1$. Thus, provided $\eta > 0$, $A(-iz_1) \neq 0$ for $z_1 \geq 0$, so $A(z) \neq 0$ for $z \in V$. Thus (6.7) holds if $\eta > 0$.

We have proven, in the case $\eta > 0$ and in the case $\eta = 0$, $\kappa_1 > 0$, that (6.7) holds, and thus we have shown the invertibility of $I + K - i\eta S$ and the boundedness of the inverse operator in $L^2(\Gamma_h)$. Thus we have established the solvability of (2.16) in the space $L^2(\Gamma)$ for flat surfaces. If Γ_f is mildly rough, we may use a perturbation argument to show that the integral equation remains solvable. We state our result precisely in the following theorem.

THEOREM 6.3. *Suppose that $h > 0$ and that either $\eta > 0$ or $\eta = 0$ and $\kappa_1 > 0$. Then, provided $\|f - h\|_{BC^{1,\alpha}(\mathbb{R}^2)}$ is sufficiently small (so that Γ_f is sufficiently close to the flat surface $f \equiv h$), it holds that $(I + K - i\eta S)^{-1}$ exists and is bounded as an operator on $L^2(\Gamma_f)$.*

Proof. Let $A = I + K - i\eta S$, and then denote A by A_f to denote its dependence on f. With the help of the isomorphism $I_f : L^2(\Gamma_f) \to L^2(\mathbb{R}^2)$ defined by (5.18) we associate A_f with the element $\tilde{A}_f = I_f A_f I_f^{-1}$ of the set of bounded linear operators on $L^2(\mathbb{R}^2)$. Now $\tilde{A}_h$ is invertible with bounded inverse, by our analysis above for the flat plane case. Moreover, by the continuity of $\tilde{A}_f$ with respect to f as proven in Theorem 5.6 it follows from standard arguments that $\tilde{A}_f$ is boundedly invertible on $L^2(\mathbb{R}^2)$ for $\|f - h\|_{BC^{1,\alpha}(\mathbb{R}^2)}$ sufficiently small, and so A_f is boundedly invertible on $L^2(\Gamma_f)$. ∎

Combining Theorem 6.3 with Lemma 6.1 we deduce Theorem 2.4. Combining Theorem 6.3 with Corollary 6.2 we establish Theorem 2.5.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] T. ARENS, *Existence of solution in elastic wave scattering by unbounded rough surfaces*, Math. Methods Appl. Sci., 25 (2002), pp. 507–528.

[3] T. ARENS, S. N. CHANDLER-WILDE, AND K. O. HASELOH, *Solvability and spectral properties of integral equations on the real line.* II. *L^p-spaces and applications*, J. Integral Equations Appl., 15 (2003), pp. 1–35.

[4] H. BRAKHAGE AND P. WERNER, *Über das Dirichletsche Außenraumproblem für die Helmholtzsche Schwingungsgleichung*, Arch. Math., 16 (1965), pp. 325–329.

[5] S. N. CHANDLER-WILDE AND A. T. PEPLOW, *A boundary integral equation formulation for the Helmholtz equation in a locally perturbed half-plane*, ZAMM Z. Angew. Math. Mech., 85 (2005), pp. 79–88.

[6] S. N. CHANDLER-WILDE AND C. R. ROSS, *Uniqueness results for direct and inverse scattering by infinite surfaces in a lossy medium*, Inverse Problems, 11 (1995), pp. 1063–1067.

[7] S. N. CHANDLER-WILDE AND C. R. ROSS, *Scattering by rough surfaces: The Dirichlet problem for the Helmholtz equation in a nonlocally perturbed half-plane*, Math. Methods Appl. Sci., 19 (1996), pp. 959–976.

[8] S. N. CHANDLER-WILDE, C. R. ROSS, AND B. ZHANG, *Scattering by infinite one-dimensional rough surfaces*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 455 (1999), pp. 3767–3787.

[9] S. N. CHANDLER-WILDE AND B. ZHANG, *Electromagnetic scattering by an inhomogenous conducting or dielectric layer on a perfectly conducting plate*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 454 (1998), pp. 519–542.

[10] S. N. CHANDLER-WILDE AND B. ZHANG, *A generalised collectively compact operator theory with an application to second kind integral equations on unbounded domains*, J. Integral Equations Appl., 14 (2002), pp. 11–52.

[11] S. N. CHANDLER-WILDE, B. ZHANG, AND C. R. ROSS, *On the solvability of second kind integral equations on the real line*, J. Math. Anal. Appl., 245 (2000), pp. 28–51.

[12] W. C. Chew, J. M. Song, T. J. Cui, S. Velarnparambil, M. L. Hastriter, and B. Hu, *Review of large scale computing in electromagnetics with fast integral equation solvers*, CMES Comput. Model. Eng. Sci., 5 (2004), pp. 361–372.
[13] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer-Verlag, Berlin, 1998.
[14] D. L. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.
[15] J. A. DeSanto, *Scattering by rough surfaces*, in Scattering: Scattering and Inverse Scattering in Pure and Applied Science, R. Pike and P. Sabatier, eds., Academic Press, San Diego, 2002, pp. 15–36.
[16] J. A. DeSanto and P. A. Martin, *On the derivation of boundary integral equations for scattering by an infinite two-dimensional rough surface*, J. Math. Phys., 39 (1998), pp. 894–912.
[17] D. Dobson and A. Friedman, *The time harmonic Maxwell equations in a doubly-periodic structure*, J. Math. Anal. Appl., 166 (1992), pp. 507–528.
[18] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Tables of Integral Transforms*, vol. II, McGraw–Hill, London, 1954.
[19] R. Kress and T. Tran, *Inverse scattering for a locally perturbed half-plane*, Inverse Problems, 16 (2000), pp. 1541–1559.
[20] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.
[21] J.-C. Nedelec and F. Starling, *Integral equation methods in a quasi-periodic diffraction problem for the time-harmonic Maxwell's equations*, SIAM J. Math. Anal., 22 (1991), pp. 1679–1701.
[22] J. A. Ogilvy, *Theory of Wave Scattering from Random Rough Surfaces*, Adam Hilger, Bristol, UK, 1991.
[23] M. Reed and B. Simon, *Methods of modern mathematical physics: Part* II. *Fourier Analysis, Self-Adjointness*, Academic Press, New York, 1975.
[24] C. R. Ross, *Direct and Inverse Scattering by Rough Surfaces*, Ph.D. thesis, Brunel University, Uxbridge, UK, 1996.
[25] M. Saillard and A. Sentenac, *Rigorous solutions for electromagnetic scattering from rough surfaces*, Waves Random Media, 11 (2001), pp. R103–R137.
[26] R. H. Torres and G. V. Welland, *The Helmholtz-equation and transmission problems with Lipschitz interfaces*, Indiana Univ. Math. J., 42 (1993), pp. 1457–1485.
[27] L. Tsang, C. H. Chan, K. Pak, and H. Sangani, *Monte-Carlo simulations of large-scale problems of random rough surface scattering and applications to grazing incidence with the BMIA/canonical grid method*, IEEE Trans. Antennas and Propagation, 43 (1995), pp. 851–859.
[28] A. G. Voronovich, *Wave Scattering from Rough Surfaces*, 2nd ed., Springer, Berlin, 1998.
[29] K. F. Warnick and W. C. Chew, *Numerical simulation methods for rough surface scattering*, Waves Random Media, 11 (2001), pp. R1–R30.
[30] A. Willers, *The Helmholtz equation in disturbed half-spaces*, Math. Methods Appl. Sci., 9 (1987), pp. 312–323.
[31] M. Xia, C. H. Chan, S. Li, B. Zhang, and L. Tsang, *An efficient algorithm for electromagnetic scattering from rough surfaces using a single integral equation and multilevel sparse-matrix canonical-grid method*, IEEE Trans. Antennas and Propagation, 51 (2003), pp. 1142–1149.
[32] B. Zhang and S. N. Chandler-Wilde, *Integral equation methods for scattering by infinite rough surfaces*, Math. Methods Appl. Sci., 26 (2003), pp. 463–488.

SIAM J. APPL. MATH.
Vol. 66, No. 3, pp. 1027–1049

MATERIAL SURFACE DESIGN TO COUNTER ELECTROMAGNETIC INTERROGATION OF TARGETS*

H. T. BANKS†, K. ITO†, G. M. KEPLER†, AND J. A. TOIVANEN†

Abstract. Utilization of controllable ferromagnetic layers coating a conducting object to provide an attenuation capability against electromagnetic interrogation is discussed. The problem is formulated as a differential game and/or a robust optimization. The scattered field due to interrogation can be attenuated with the assumption of an uncertainty in the interrogation wave numbers. The controllable layer composed of ferromagnetic materials [H. How and C. Vittoria, *Implementation of Microwave Active Nulling*, private communication; H. How and C. Vittoria, *IEEE Trans. Microwave Theory Tech.*, 52 (2004), pp. 2177–2182] is incorporated in a mathematical formulation based on the time-harmonic Maxwell equation. Fresnel's law for the reflectance index is extended to the electromagnetic propagation in anisotropic composite layers of ferromagnetic and electronic devices and is used to demonstrate feasibility of control of reflections. Our methodology is also tested for a nonplanar geometry of the conducting object (an NACA airfoil) in which we report our findings in the form of reduced radar cross sections (RCS).

Key words. electromagnetic, inverse scattering, attenuation

AMS subject classifications. 35Q60, 78A46, 78M10, 78M50

DOI. 10.1137/040621430

1. Introduction.

In this paper we discuss an optimal attenuation problem; i.e., we attempt to maximize attenuation capabilities of interrogating signals by utilizing a controllable dielectric layer on the surface of a conducting object. The objective of the interrogator is to detect and identify the location and shape of the conducting object based on the scattered field from an interrogation incident field, i.e., the solution of an inverse scattering problem [2, 7]. In the plane wave case the incident electromagnetic (EM) field has the form $(\vec{E}^{(i)}, \vec{H}^{(i)})\, e^{i\vec{k}\cdot\vec{x}}$ and the interrogator has control over the wave numbers $\vec{k}$. The attenuation problem is to minimize or diminish the detection capability of interrogation by either grating [14] or fabrications on the object surface. In this paper we consider the utilization of thin controllable dielectric surface layers on the object as a method for achieving the attenuation capability. Here "controllable" means that we have the capability of adjusting the material properties of the surface layers parametrically. The technical ideas developed here also have the potential to aid in the design of *medical shields* employed to protect parts of an irradiated target or to focus radiation to pinpoint specific regions of the target.

In these initial investigations, we investigate a design case in which one determines values of the dielectric permittivity and magnetic permeability of the controllable layer in order to attenuate reflections. From a control theoretic viewpoint this is a "passive" or open loop control strategy. But our efforts here lay the foundations for "active" or closed loop control strategies in which one combines controllable layer dynamics with a sensor for incoming interrogating signals to develop real time feedback controls for adaptive choice of the permittivity and permeability of the controllable layer.

*Received by the editors December 24, 2004; accepted for publication (in revised form) December 6, 2005; published electronically March 3, 2006. This research was supported in part by the U.S. Air Force Offices of Scientific Research under grant AFOSR FA9550-04-1-0220.

http://www.siam.org/journals/siap/66-3/62143.html

†Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695-8205 (htbanks@unity.ncsu.edu, kito@ncsu.edu, gmkepler@ncsu.edu, jatoivan@ncsu.edu).

To investigate feasibility, we first formulate the problem as a differential game. For example, we assume the time-harmonic incident EM plane wave is impinging on the surface at $z = 0$ and we control the effective dielectric constant ϵ of the surface layer on top of the conducting material. The scattered field due to the interrogation can be evaluated based on the time-harmonic Maxwell equations [7]. In the case when the dielectric constant is homogeneous in the horizontal directions (planar geometry), the reflectance index $R = R(\vec{k}, \epsilon)$ is determined by Fresnel's law (see [15] and section 4). Thus, the problem of nullifying the scattered field can be cast as the minimization of the scattering EM wave in terms of $|R|^2$, i.e.,

$$\min_{\epsilon \in Q} \max_{\vec{k} \in K} |R|^2, \tag{1.1}$$

where Q is a set of admissible dielectric constants and K is a set of possible interrogation wave numbers. In order to determine the admissible set Q we must describe controllable mechanisms of the dielectric layer. Thus, the problem of nullifying the scattered field can be formulated as the min-max problem of minimizing the largest reflectance by interrogations over $\vec{k} \in K$, over all possible designs ($\epsilon \in Q$ in this case). This game theoretic formulation is used in many other design problems. It does not assume any information on the uncertainty of interrogations and thus it may lead to a conservative design. An alternative formulation can be given in a more robust form, i.e.,

$$\min_{\epsilon \in Q} \int_K |R|^2 \, d\mathcal{K}(\vec{k}), \tag{1.2}$$

where $\mathcal{K}$ is a probability distribution function on the wave numbers $\vec{k}$. This formulation then needs information about the distribution function $\mathcal{K}$ of the interrogating plane wave. As demonstrated in section 2, better knowledge of the distribution function greatly improves the performance of minimizing the scattered field.

After demonstrating attenuation capabilities, we turn to the general case of a nonplanar conducting medium with a controllable coating layer. The far field pattern $F(\theta)$ of the scattered field (see [7] and section 5) is then a function of the wave number $\vec{k}$ of the incident plane wave and the material properties (ϵ, μ) of the controllable layer, i.e.,

$$F(\theta) = U(\vec{k}, (\epsilon, \mu); \theta), \quad 0 \le \theta \le 2\pi. \tag{1.3}$$

One can select the performance index $J(\vec{k}, (\epsilon, \mu)) = \Phi(F)$ to perform specific alterations of the scattered field, which of course depend on the inverse techniques employed by the interrogator. Here $\Phi(F)$ is some performance index for the far field pattern $F(\theta)$. We investigate optimal radar cross sections (RCS) for one class of such problems.

We note that the existence of solutions to a general min-max problem is guaranteed under the condition that given $\epsilon \in Q$ the value function $V(\epsilon) = \sup_{\vec{k} \in K} J(\vec{k}, \epsilon)$ is lower semicontinuous, which is typically satisfied under very mild conditions (e.g., see [13]) when Q is compact. The saddle point property of a solution pair $(\vec{k}_0, \epsilon_0)$,

$$J(\vec{k}, \epsilon_0) \le J(\vec{k}_0, \epsilon_0) \le J(\vec{k}_0, \epsilon) \quad \text{for all } \epsilon \in Q, \ \vec{k} \in K, \tag{1.4}$$

holds locally if the Hessian of $J(\vec{k}, \epsilon)$ is hyperbolic at $(\vec{k}_0, \epsilon_0)$. The existence of solutions to the robust formulation then simply follows from the continuity of $J(\vec{k}, \epsilon_0)$ with respect to $(\vec{k}, \epsilon) \in K \times Q$.

A brief outline of our presentation here is as follows. In section 2 we consider the exact problem for the planar geometry, present representative numerical calculations for optimal design of the dielectric layer based on a robust formulation, and demonstrate the feasibility of this approach and the effectiveness of the design. In section 3 we discuss a controllable layer composed of ferromagnetic and ferroelectric materials as proposed by How and Vittoria [10, 11]. This composite model is designed so that a control mechanism for the material properties of the layers can be achieved in both a parametric and a dynamic manner. In section 4 we mathematically formulate the forward problem for the controllable composite layers, including the tensor permeability in the ferrite layer, by calculating the scattered field R as a function of the interrogating wave and the near surface composition. We consider the time-harmonic case with plane wave interrogations of the planar geometry (i.e., the composite layers are homogeneous in (x, y)). In this case we construct the plane wave solution and an analytic expression for the reflectance index R. Another important feature of our formulation is the possible identification of the interrogating wave in terms of its distribution $\mathcal{K}$. Since the plane wave calculation also yields surface currents as an explicit function of the incident interrogations, the surface current measurements can, in principle, be used to identify the distribution $\mathcal{K}$ of the interrogations $\vec{k}$. In sections 5–7 we present results for our formulation when applied to a nonplanar geometry by considering the NACA0012 airfoil [16]. In general (and in particular in this case) we do not have an analytic expression for the far field pattern F and thus we use a numerical computation of the scattered field. In this case, our numerical computations for the scattered field F are implemented using the finite element method (in section 6). Our numerical findings are presented for an optimal homogeneous coating layer.

2. Feasibility study. In this section we first demonstrate the feasibility of our approach. We consider an incident parallel polarized (TE_x mode) plane wave $\vec{H} = (H_x^{(i)}, 0, 0)e^{i\,\vec{k}\cdot\vec{x}}$ impinging on the interface of the first and second layers at $z = 0$, as depicted in Figure 1. The interface between the second and third layers is located at $z = -d$, where the third layer is a perfect conductor. We control the effective dielectric constant $\epsilon^{(2)}$ of the second layer coating the conducting material. By Fresnel's law (see [4, 15] and the discussion in section 4) we have that the reflection coefficient or reflectance index is given by

$$R = \frac{\dfrac{\epsilon^{(2)}k_z^{(1)} - \epsilon^{(1)}k_z^{(2)}}{\epsilon^{(2)}k_z^{(1)} + \epsilon^{(1)}k_z^{(2)}} + e^{-2ik_z^{(2)}d}}{1 + \dfrac{\epsilon^{(2)}k_z^{(1)} - \epsilon^{(1)}k_z^{(2)}}{\epsilon^{(2)}k_z^{(1)} + \epsilon^{(1)}k_z^{(2)}}\, e^{-2ik_z^{(2)}d}}, \tag{2.1}$$

where

$$\begin{aligned} k_z^{(1)} &= \frac{2\pi}{\lambda}\sqrt{1 - \sin^2\varphi_0}, \\ k_z^{(2)} &= \frac{2\pi}{\lambda}\sqrt{\epsilon^{(2)} - \epsilon^{(1)}\sin^2\varphi_0}. \end{aligned} \tag{2.2}$$

Here φ_0 is the incident angle with respect to the normal to the surface ($\tan\varphi_0 = \frac{k_y^{(1)}}{k_z^{(1)}}$, $k_x^{(1)} = 0$) and λ is the wavelength of the incident wave. We note from (2.1) that R depends on the ratio $\epsilon^{(2)}/\epsilon^{(1)}$ and hence without loss of generality we may normalize the parameters so that $\epsilon^{(1)} = 1$. The reflectance index R is a function of (λ, φ_0),

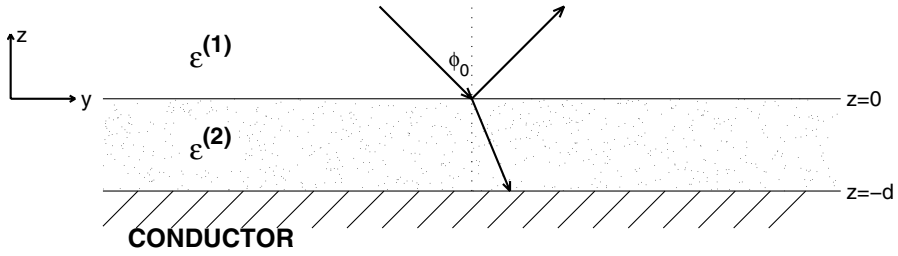


FIG. 1. *Schematic representation of the reflection of a plane wave incident with angle ϕ_0 on a planar three-layer stack. The top two layers are dielectric media. The third layer is a perfect conductor.*

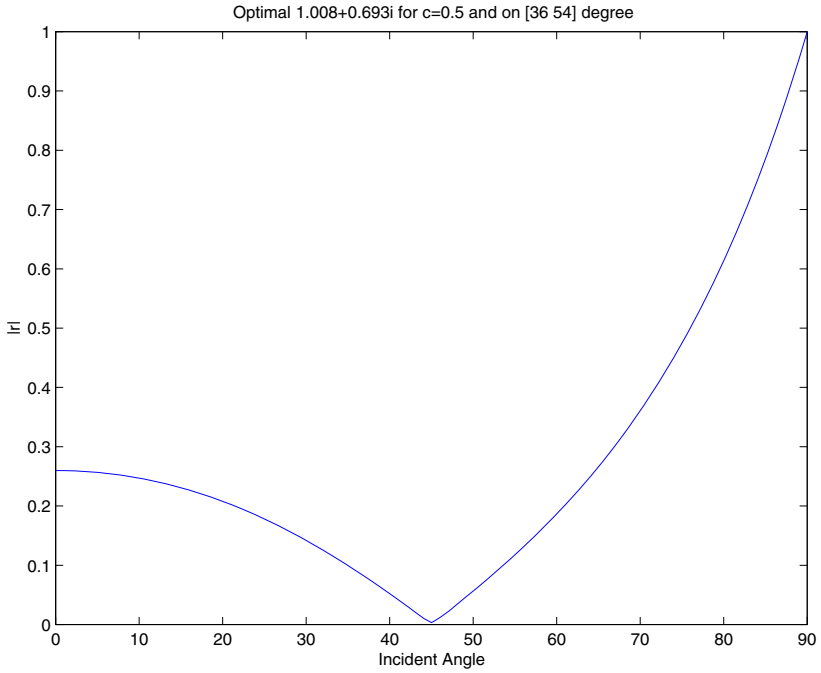


FIG. 2. *The reflected intensity $|R|$ as a function of the incident angle φ_0. The uncertainty interval is $[36, 54]$ degrees.*

the normalized dielectric constant $\epsilon^{(2)}$, and the thickness d of the surface layer. We assume that d is positive and fixed. We parameterize the incident wave in terms of (λ, φ_0).

In Figures 2 and 3 we depict the robustness of the optimal solution by plotting the reflectance intensity $|R|$ as a function of the incident angle and the normalized thickness/wavelength ratio, defined as $a = \frac{d}{\lambda}$. In Figure 2 we assume that the uncertainty in wave numbers is due only to uncertainty in the incident angle φ_0, which is uniformly distributed on the interval $[36, 54]$ degrees and graph the intensity $|R|$ corresponding to the optimal dielectric constant $\epsilon^{(2)} = 1.008 + .693i$ as a function of the incident angle. The integration of $|R|^2$ over φ_0 is performed using Simpson's rule. The intensity of the reflection is well attenuated over the uncertainty interval $[36, 54]$ degrees.

Next, we assume that there is uncertainty in both the incident angle φ_0 and the normalized thickness/wavelength ratio a, which are uniformly distributed on a rectangle $[36, 54] \times [0.3, 0.7]$. In Figure 3 we plot the intensity $|R|$ corresponding to the optimal dielectric constant $\epsilon^{(2)} = 1.4309 + 1.0724i$ for several sampled incident points in the frequency. A reasonable attenuation over the uncertainty box is obtained in this example. It is clear that the performance depends on the quality of the information on the distribution of the interrogating wave, as demonstrated by the better attenuation results in the first case (Figure 2) than in the second case (Figure 3).

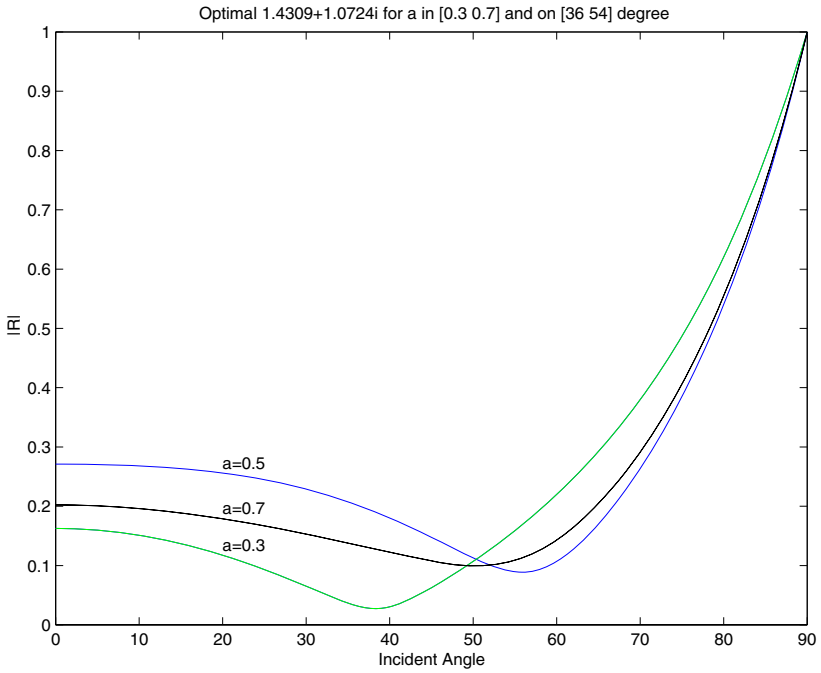


FIG. 3. *The reflected intensity* $|R|$ *as a function of the incident angle* φ_0 *and the normalized thickness/wavelength ratio* a. *The uncertainty box is* $[36, 54] \times [0.3, 0.7]$.

3. Controllable sublayers composed of ferromagnetic and ferroelectric materials. In this section we describe an experimental device that can be used to control the dielectric permittivity (and magnetic permeability) in a coating layer as discussed in the previous sections. In Figure 4 we present a schematic of the configuration of an active reflecting device proposed and investigated experimentally by How and Vittoria in [10, 11]. The reflector contains a ferrite layer and a ferroelectric layer as constituents. The permanent magnet provides a common magnetic bias so that the ferromagnetic resonance (FMR) condition can be readily achieved and thereby facilitate sensitive magnetic tuning by the local Helmholtz coils. The dielectric properties of the ferroelectric layer are controlled through the ground plane bias field. The purpose of this reflector design is to provide phase and impedance control of the composite layers so that nullification and alteration of the scattered wave can be achieved in the response to an incident interrogating EM wave. The integrated circuits are designed so that the tuning sensitivity of the device is enhanced. The key element of the device is that the material properties $\mu(H)$ and $\epsilon(E)$ of the composite layers are controllable in terms of the magnetic mean in the ferrite layer and the electric

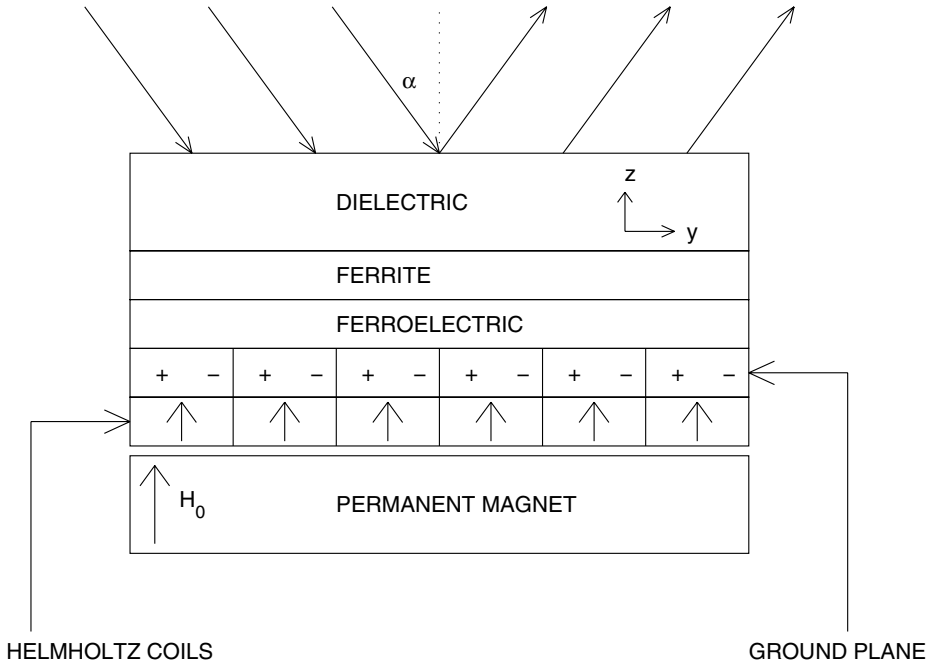


FIG. 4. *Composite sublayers composing an active reflector device.*

mean in the ferroelectric layer, and thus can support agile frequency attenuation.

The most important device characteristic of the ferrite in the investigation [10, 11] is that the magnetic permeability $\bar{\mu}$ is a tensor, so that, due to the gyromagnetic effect, EM propagation in the ferrite is anisotropic in the presence of a dc-bias magnetic field [12, 18]. For a ferrite magnetized in the y direction with damping and no demagnetization, the permeability tensor is given by [17]

(3.1)
$$\bar{\mu} = \begin{bmatrix} \mu & 0 & -i\kappa \\ 0 & \mu_0 & 0 \\ i\kappa & 0 & \mu \end{bmatrix}, \quad \text{where} \quad \mu = \mu_0 \left(1 + \frac{\bar{\omega}_0 \omega_m}{\bar{\omega}_0^2 - \omega^2}\right), \quad \kappa = \mu_0 \frac{\omega \omega_m}{\bar{\omega}_0^2 - \omega^2},$$
$$\omega_m = 4\pi\gamma M_z, \quad \bar{\omega}_0 = \omega_0 + i/\tau, \quad \omega_0 = \gamma H_0,$$

ω_0 is the precession frequency, H_0 is the impressed dc magnetic field, γ is the gyromagnetic ratio, M_z is the saturation magnetization, and τ is the relaxation time.

The ferrite device is most useful if it operates near the FMR frequency ω_0 so that the rapid change in magnetic permeability can be effectively utilized, either to obtain frequency-tuning capability or to remove the degeneracy between modes [10, 11].

4. Plane wave solution. We next discuss a plane wave solution as it interacts with a ferrite layer. Due to the tensor magnetic permeability $\bar{\mu}$, the electric and magnetic modes are coupled in the ferrite layer. In this section we present the detailed calculations for constructing the fundamental solution in the ferrite layer. First, the time-harmonic Maxwell equation (4.1) is reduced to a system (4.12) of the differential equations in the z (depth) direction and then the characteristic equation and the

form (4.16) of the fundamental solutions are established. A similar calculation can be carried out for the ferroelectric layer, but we shall not pursue that here.

The time-harmonic Maxwell equations are written as

$$\begin{aligned} \nabla \times H &= i\omega\epsilon\, E, \\ \nabla \times E &= -i\omega\bar{\mu}\, H, \\ \nabla \cdot (\bar{\mu} H) &= 0, \\ \nabla \cdot E &= 0. \end{aligned} \tag{4.1}$$

For the ferrite layer the permeability $\bar{\mu}$ is the tensor defined by (3.1), and thus the first three equations in (4.1) can be written as

$$\begin{pmatrix} \dfrac{\partial}{\partial y}H_z - \dfrac{\partial}{\partial z}H_y \\ \dfrac{\partial}{\partial z}H_x - \dfrac{\partial}{\partial x}H_z \\ \dfrac{\partial}{\partial x}H_y - \dfrac{\partial}{\partial y}H_x \end{pmatrix} = i\omega\epsilon \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix}, \tag{4.2}$$

$$\begin{pmatrix} \dfrac{\partial}{\partial y}E_z - \dfrac{\partial}{\partial z}E_y \\ \dfrac{\partial}{\partial z}E_x - \dfrac{\partial}{\partial x}E_z \\ \dfrac{\partial}{\partial x}E_y - \dfrac{\partial}{\partial y}E_x \end{pmatrix} = -i\omega \begin{pmatrix} \mu H_x - i\kappa H_z \\ \mu_0 H_y \\ i\kappa H_x + \mu H_z, \end{pmatrix}, \tag{4.3}$$

and

$$-\mu\left(\frac{\partial}{\partial x}H_x + \frac{\partial}{\partial z}H_z\right) = i\kappa\left(\frac{\partial}{\partial z}H_x - \frac{\partial}{\partial x}H_z\right) + \mu_0\,\frac{\partial}{\partial y}H_y. \tag{4.4}$$

Taking the cross product of (4.2) and using (4.3), we obtain (y-component)

$$i\omega\mu_0\, H_y = \frac{1}{i\omega\epsilon}\left(\frac{\partial^2}{\partial x^2}H_y + \frac{\partial^2}{\partial z^2}H_y - \frac{\partial}{\partial y}\left(\frac{\partial}{\partial x}H_x + \frac{\partial}{\partial z}H_z\right)\right). \tag{4.5}$$

Using the y-component of (4.2),

$$i\omega\epsilon\, E_y = \frac{\partial}{\partial z}H_x - \frac{\partial}{\partial x}H_z, \tag{4.6}$$

in (4.4) and substituting into (4.5), we have

$$\frac{\partial^2}{\partial x^2}H_y + \frac{\partial^2}{\partial z^2}H_y + \frac{\mu_0}{\mu}\,\frac{\partial^2}{\partial y^2}H_y + \omega^2\,\mu_0\epsilon\, H_y = \frac{\omega\kappa\epsilon}{\mu}\,\frac{\partial}{\partial y}E_y. \tag{4.7}$$

The components of (4.3) can be manipulated to yield the following equations:

$$\mu\left(\frac{\partial}{\partial y}E_z - \frac{\partial}{\partial z}E_y\right) + i\kappa\left(\frac{\partial}{\partial x}E_y - \frac{\partial}{\partial y}E_x\right) = -i\omega(\mu^2 - \kappa^2)H_x,$$

$$\frac{\partial}{\partial z}E_x - \frac{\partial}{\partial x}E_z = -i\omega\mu_0\, H_y, \tag{4.8}$$

$$-i\kappa\left(\frac{\partial}{\partial y}E_z - \frac{\partial}{\partial z}E_y\right) + \mu\left(\frac{\partial}{\partial x}E_y - \frac{\partial}{\partial y}E_x\right) = -i\omega(\mu^2 - \kappa^2)\, H_z.$$

Using the y-component of (4.2) and the first and last equations of (4.8), one can obtain

$$\frac{\partial}{\partial z}\left(\mu\left(\frac{\partial}{\partial y}E_z-\frac{\partial}{\partial z}E_y\right)+i\kappa\left(\frac{\partial}{\partial x}E_y-\frac{\partial}{\partial y}E_x\right)\right)$$
$$-\frac{\partial}{\partial x}\left(-i\kappa\left(\frac{\partial}{\partial y}E_z-\frac{\partial}{\partial z}E_y\right)+\mu\left(\frac{\partial}{\partial x}E_y-\frac{\partial}{\partial y}E_x\right)\right)=\omega^2\epsilon\left(\mu^2-\kappa^2\right)E_y.$$

This can be rearranged to obtain

$$-\mu\left(\frac{\partial^2}{\partial z^2}E_y+\frac{\partial^2}{\partial x^2}E_y\right)$$
$$+\frac{\partial}{\partial y}\left(\mu\left(\frac{\partial}{\partial z}E_z+\frac{\partial}{\partial x}E_x\right)-i\kappa\left(\frac{\partial}{\partial z}E_x-\frac{\partial}{\partial x}E_z\right)\right)=\omega^2\epsilon\left(\mu^2-\kappa^2\right)E_y.$$

Using this result with the y-component of (4.8) and the fourth equation of (4.1), we have

$$\frac{\partial^2}{\partial x^2}E_y+\frac{\partial^2}{\partial y^2}E_y+\frac{\partial^2}{\partial z^2}E_y+\frac{\omega^2\epsilon\left(\mu^2-\kappa^2\right)}{\mu}E_y=-\frac{\omega\mu_0\kappa}{\mu}\frac{\partial}{\partial y}H_y. \tag{4.9}$$

From (4.2)

$$i\omega\epsilon\begin{pmatrix}E_x\\E_z\end{pmatrix}=\begin{pmatrix}\dfrac{\partial}{\partial y}H_z-\dfrac{\partial}{\partial z}H_y\\[2ex]\dfrac{\partial}{\partial x}H_y-\dfrac{\partial}{\partial y}H_x\end{pmatrix}.$$

If we define

$$E^{\pm}=E_z\pm i\,E_x,\quad H^{\pm}=H_z\pm i\,H_x,$$

then it follows that

$$i\omega\epsilon\begin{pmatrix}E^+\\E^-\end{pmatrix}=\begin{pmatrix}-i\nabla^+H_y+i\dfrac{\partial}{\partial y}H^+\\[2ex]i\nabla^-H_y-i\dfrac{\partial}{\partial y}H^-\end{pmatrix}, \tag{4.10}$$

where

$$\nabla^{\pm}\phi=\frac{\partial}{\partial z}\phi\pm i\,\frac{\partial}{\partial x}\phi.$$

From (4.3)

$$-i\omega\begin{pmatrix}\mu\,H_x-i\kappa\,H_z\\\mu\,H_z+i\kappa\,H_x\end{pmatrix}=\begin{pmatrix}\dfrac{\partial}{\partial y}E_z-\dfrac{\partial}{\partial z}E_y\\[2ex]\dfrac{\partial}{\partial x}E_y-\dfrac{\partial}{\partial y}E_x\end{pmatrix},$$

and thus

$$-i\omega \begin{pmatrix} (\mu+\kappa)\, H^+ \\ (\mu-\kappa)\, H^- \end{pmatrix} = \begin{pmatrix} -i\nabla^+ E_y + i\dfrac{\partial}{\partial y} E^+ \\ i\nabla^- E_y - i\dfrac{\partial}{\partial y} E^- \end{pmatrix}. \tag{4.11}$$

From (4.10)–(4.11)

$$-\omega^2 \epsilon\,(\mu \pm \kappa)\, H^\pm - \frac{\partial^2}{\partial y^2} H^\pm \pm \omega\epsilon\, \nabla^\pm E_y + \frac{\partial}{\partial y} \nabla^\pm H_y = 0,$$

and we have

$$-\omega^2 \epsilon\,(\mu \pm \kappa)\, E^\pm - \frac{\partial^2}{\partial y^2} E^\pm \mp \omega(\mu \pm \kappa)\, \nabla^\pm H_y + \frac{\partial}{\partial y} \nabla^\pm E_y = 0.$$

Let $\hat{E}$ and $\hat{H}$ be the (partial) Fourier transform (in (x, y)) of E and H, i.e.,

$$\hat{E}(k_x, k_y, z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} E(x, y, z) e^{-ik_x\, x - ik_y\, y}\, dxdy.$$

Then (4.9) and (4.7) can be written, respectively, as

$$\begin{aligned} &\frac{\partial^2}{\partial z^2} \hat{E}_y - \left(k_x^2 + k_y^2 - \omega^2 \epsilon \frac{\mu^2 - \kappa^2}{\mu} \right) \hat{E}_y = -i\omega\mu_0 k_y \frac{\kappa}{\mu}\, \hat{H}_y, \\ &\frac{\partial^2}{\partial z^2} \hat{H}_y - \left(k_x^2 + \frac{\mu_0}{\mu} k_y^2 - \omega^2 \mu_0 \epsilon \right) \hat{H}_y = i\omega\epsilon k_y \frac{\kappa}{\mu}\, \hat{E}_y \end{aligned} \tag{4.12}$$

and (4.10)–(4.11) as

$$\begin{aligned} \hat{E}^\pm &= \frac{-ik_y\, \nabla^\pm \hat{E}_y \pm \omega(\mu \pm \kappa)\, \nabla^\pm \hat{H}_y}{k_y^2 - \omega^2 \epsilon(\mu \pm \kappa)}, \\ \hat{H}^\pm &= \frac{-ik_y\, \nabla^\pm \hat{H}_y \mp \omega\epsilon\, \nabla^\pm \hat{E}_y}{k_y^2 - \omega^2 \epsilon(\mu \pm \kappa)}, \end{aligned} \tag{4.13}$$

where

$$\nabla^\pm = \frac{\partial}{\partial z} \mp k_x.$$

Thus the electric and magnetic modes are coupled by (4.12). Next we find the fundamental solution within the ferrite layer. Let

$$\tilde{A} = k_x^2 + \frac{\mu_0}{\mu} k_y^2 - \omega^2 \mu_0 \epsilon, \quad \tilde{B} = k_x^2 + k_y^2 - \omega^2 \epsilon \frac{\mu^2 - \kappa^2}{\mu},$$

$$\tilde{C} = i\omega\epsilon k_y \frac{\kappa}{\mu}, \quad \tilde{D} = i\omega\mu_0 k_y \frac{\kappa}{\mu}.$$

From the second equation of (4.12)

$$\hat{E}_y = \frac{\dfrac{\partial^2}{\partial z^2} \hat{H}_y - \tilde{A}}{\tilde{C}}\, \hat{H}_y,$$

and substituting this into the first equation of (4.12), we obtain

$$\frac{\partial^4}{\partial z^4}\hat{H}_y - (\tilde{A}+\tilde{B})\frac{\partial^2}{\partial z^2}\hat{H}_y + (\tilde{A}\tilde{B}-\tilde{C}\tilde{D})\hat{H}_y = 0.$$

Thus, we obtain the characteristic equation for system (4.12)

$$k^4 + (\tilde{A}+\tilde{B})k^2 + \tilde{A}\tilde{B} - \tilde{C}\tilde{D} = 0$$

for the exponential solution

$$\hat{H}_y = e^{ik\,z}, \;\; \hat{E}_y = -\frac{k^2+\tilde{A}}{\tilde{C}}\,\hat{H}_y.$$

The characteristic equation has four roots, $\pm k_+$ and $\pm k_-$. In each case, $\hat{E}_y$ is proportional to $\hat{H}_y$ so we proceed to consider solutions of the form

$$\hat{E}_y = i\eta\,\hat{H}_y. \tag{4.14}$$

Then η must satisfy

$$k_y^2 - \omega^2\epsilon\frac{\mu^2-\kappa^2}{\mu} - \omega\mu_0 k_y\frac{\kappa}{\mu\eta} = \frac{\mu_0}{\mu}k_y^2 - \omega^2\mu_0\epsilon - \omega\epsilon k_y\frac{\kappa}{\mu}\eta. \tag{4.15}$$

Equation (4.15) has two roots, which we designate as η_+ and η_-. Then $\hat{H}_y$ must satisfy either of the equations

$$\frac{\partial^2}{\partial z^2}\hat{H}_y - \left(k_x^2 + \frac{\mu_0}{\mu}k_y^2 - \omega^2\mu_0\epsilon - \eta_\pm\,\omega\epsilon k_y\frac{\kappa}{\mu}\right)\hat{H}_y = 0.$$

Thus, $\hat{E}_y$ and $\hat{H}_y$ can be written in the general form

$$\begin{aligned}\hat{E}_y &= i\eta_+\,(A_1\,e^{ik_+z} + B_1\,e^{-ik_+z}) + i\eta_-\,(A_2\,e^{ik_-z} + B_2\,e^{-ik_-z}),\\ \hat{H}_y &= A_1\,e^{ik_+z} + B_1\,e^{-ik_+z} + A_2\,e^{ik_-z} + B_2\,e^{-ik_-z},\end{aligned} \tag{4.16}$$

$$k_\pm = \sqrt{\omega^2\mu_0\epsilon + \eta_\pm\omega\epsilon k_y\frac{\kappa}{\mu} - k_x^2 - \frac{\mu_0}{\mu}k_y^2}\,.$$

Formulae (4.12)–(4.16) are given in [12, 18] without detailed derivations. These results allow us to construct the plane wave solution in a ferrite layer.

In a general dielectric (including ambient) medium, since $\kappa = 0$ and $\mu = \mu_0$, the system (4.12) is decoupled. Thus the fundamental solutions (in the partial Fourier domain formulation) are given by

$$\vec{E} = Ae^{ik_z} + Be^{-ik_z}, \quad \vec{H} = Ce^{ik_z} + De^{-ik_z}$$

with (4.1) and

$$k_z = \sqrt{\omega^2\epsilon\mu - k_x^2 - k_y^2}.$$

For example, we have

$$(\bar{E}_x, 0, 0)e^{ik_z z}, \quad \left(0, -\frac{k_z}{\omega\mu}\bar{E}_x, \frac{k_y}{\omega\mu}\bar{E}_x\right)e^{ik_z z} \tag{4.17}$$

for a perpendicular polarized (TM_x mode) incident wave and

$$\left(0, \frac{k_z}{\omega\epsilon}\bar{H}_x, -\frac{k_y}{\omega\epsilon}\bar{H}_x\right)e^{ik_z z}, \quad (\bar{H}_x, 0, 0)e^{ik_z z} \tag{4.18}$$

for a parallel polarized (TE_x mode) incident wave, where $\bar{E}_x$ and $\bar{H}_x$ are constants, when $k_x = 0$ and k_y are fixed.

Now we consider the case when a ferrite layer with thickness d on a perfectly conducting medium is impinged upon by the parallel polarized incident wave (4.17) $\vec{H}^{(i)}(x, y, z) = (H_x^{(i)}, 0, 0)e^{i\vec{k}\cdot\vec{x}}$ (its partial Fourier transform is $(H_x^{(i)}, 0, 0)e^{ik_z z}$) with $k_x = 0$ and k_y fixed. The transmitted wave $(\vec{E}^{(t)}, \vec{H}^{(t)})$ defined by (4.16) in the ferrite layer is generally not TE_x mode alone and thus has a nontrivial $E_x^{(t)}$. Hence, for a given incident wave, the reflected wave $(\vec{E}^{(r)}, \vec{H}^{(r)})e^{-ik_z z}$ in the ambient layer is a linear combination of the two fundamental solutions of the form (4.17)–(4.18) (with $k_z = -k_z$). Given the TE_x incident field, the constant weights $(E_x^{(r)}, H_x^{(r)})$ for the reflected wave can be determined by the system of equations

$$\begin{cases} H_x^{(i)} + H_x^{(r)} = \hat{H}_x^{(t)}(0), \\ \dfrac{k_z}{\omega\mu}(0 + E_x^{(r)}) = \hat{H}_y^{(t)}(0), \\ 0 + E_x^{(r)} = \hat{E}_x^{(t)}(0), \\ \dfrac{k_z}{\omega\epsilon}(H_x^{(i)} - H_x^{(r)}) = \hat{E}_y^{(t)}(0), \\ \hat{E}_x^{(t)}(-d) = \hat{E}_y^{(t)}(-d) = 0 \end{cases} \tag{4.19}$$

for $(A_1, B_1, A_2, B_2, E_x^{(r)}, H_x^{(r)})$ and the transmitted wave $(\vec{E}^{(t)}, \vec{H}^{(t)})$ determined by (4.13) and (4.16). The first two equations impose the continuity of H components, the next two equations impose the continuity of E components at $z = 0$ (the interface between the ambient and the ferrite layer), and the last equation enforces the perfectly conducting boundary conditions at $z = -d$. Moreover, the induced surface current $\vec{J}$ due to the incident wave is given by

$$\vec{J} = (\hat{H}_y(-d), -\hat{H}_x(-d), 0) = \vec{n} \times \vec{\hat{H}}.$$

This construction procedure can be readily extended to the case of the composite of sublayers.

4.1. Dielectric case (Fresnel's law). In this section we consider the dielectric layer ($\mu = \mu_0$, $\kappa = 0$) with thickness d on the perfectly conducting medium and show that (4.19) reduces to the usual Fresnel's law for the parallel polarized (TE_x) incident

wave. In this case we have $E_x^{(r)} = E_x^{(t)} = H_y^{(r)} = H_y^{(t)} = 0$ and

$$k_z^{(1)} = \sqrt{\omega^2 \mu_0 \epsilon^{(1)} - k_y^2},$$

$$k_+ = k_- = k_z^{(2)} = \sqrt{\omega^2 \mu_0 \epsilon^{(2)} - k_y^2}.$$

In the dielectric layer we have (in the partial Fourier domain)

$$\hat{H}_x^{(t)} = H_+^{(t)} e^{ik_z^{(2)} z} + H_-^{(t)} e^{-ik_z^{(2)} z},$$

$$\hat{E}_y^{(t)} = \frac{k_z^{(2)}}{\omega \epsilon^{(2)}} (H_+^{(t)} e^{ik_z^{(2)} z} - H_-^{(t)} e^{-ik_z^{(2)} z}),$$

where $H_\pm^{(t)}$ are constants. Thus, (4.19) becomes

$$\begin{cases} H_x^{(i)} + H_x^{(r)} = H_+^{(t)} + H_-^{(t)} \dfrac{k_z^{(1)}}{\omega \epsilon^{(1)}} (H_x^{(i)} - H_x^{(r)}) = \dfrac{k_z^{(2)}}{\omega \epsilon^{(2)}} H_+^{(t)} - \dfrac{k_z^{(2)}}{\omega \epsilon^{(2)}} H_-^{(t)}, \\ e^{-ik_z^{(2)} d} H_+^{(t)} - e^{ik_z^{(2)} d} H_-^{(t)} = 0. \end{cases}$$

Hence we obtain Fresnel's law

$$H_x^{(r)} = \frac{\dfrac{\epsilon^{(2)} k_z^{(1)} - \epsilon^{(1)} k_z^{(2)}}{\epsilon^{(2)} k_z^{(1)} + \epsilon^{(1)} k_z^{(2)}} + e^{-2ik_z^{(2)} d}}{1 + \dfrac{\epsilon^{(2)} k_z^{(1)} - \epsilon^{(1)} k_z^{(2)}}{\epsilon^{(2)} k_z^{(1)} + \epsilon^{(1)} k_z^{(2)}} e^{-2ik_z^{(2)} d}} H_x^{(i)}.$$

4.2. Numerical tests. In this section we demonstrate the feasibility of using the controllable property of the ferrite layer to attenuate reflections. We select $\epsilon_0 = 1$, $\mu_0 = 10$, and $d = .5$ (normalized). The magnetic permeability $\bar{\mu}$ of the ferrite layer is parameterized by

$$\mu = \mu_0 \left(0.3 + \frac{\beta}{100}\right), \quad \kappa = \mu_0 \left(-0.01 + \frac{\beta}{100}\right) \quad \text{for } 1 \le \beta \le 100,$$

where a parameter β plays a role of tuning the frequency ω_m in (3.1). In Figure 5 we depict $|(E_x^{(r)}, H_x^{(r)})|$ as a function of β for the three different incident angles $\phi_0 = 40°, 45°, 50°$. This figure establishes the attenuation capability of the ferrite layer that can be achieved by tuning the permeability. To reveal the phase dependence, in Figure 6 we graph the real and imaginary parts of $H_x^{(r)}$ as a function of the frequency for the incident angle $\phi = 45°$.

5. Optimization of material parameters of a coated airfoil. Having discussed the feasibility of tuning dielectric and magnetic properties of a coating on a perfect conductor in the previous sections, we turn in the next several sections to field calculations for a coated airfoil and demonstrate our ability to significantly affect the RCS by appropriate manipulation of the parameters ϵ and μ in the coating.

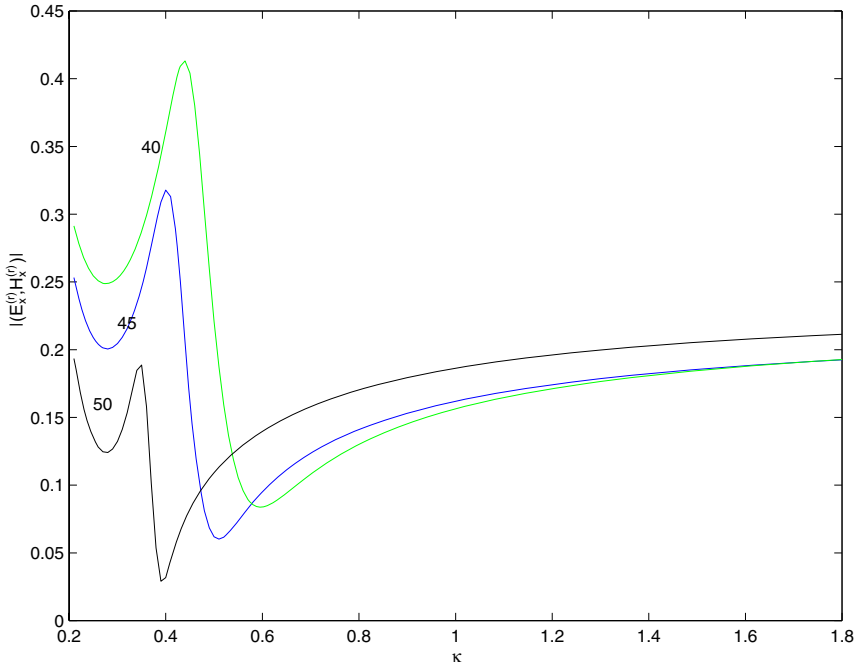


FIG. 5. *Attenuation capability.*

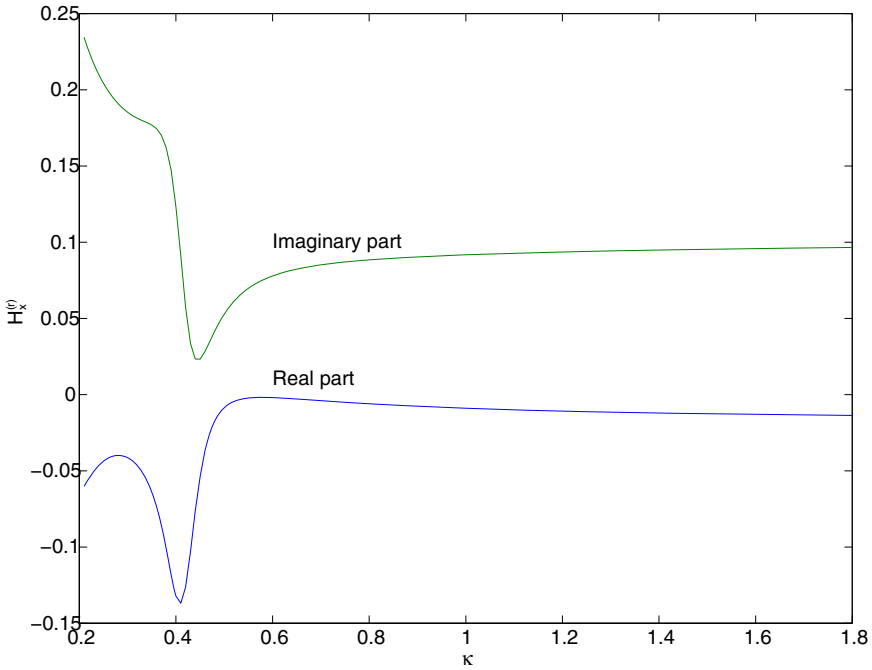


FIG. 6. *Phase shift property.*

5.1. Time-harmonic Maxwell equation for the transverse magnetic mode. We consider the scattering of a perfectly conducting airfoil coated by a layer of constant thickness. The interrogating electromagnetic incident wave is assumed to be a time-harmonic and transverse magnetic (more precisely, TM_x) mode. Thus, the time-harmonic electric and magnetic fields have the form

$$E = \begin{pmatrix} E_x \\ 0 \\ 0 \end{pmatrix} \qquad \text{and} \qquad H = -\frac{i}{\omega\mu} \begin{pmatrix} 0 \\ \frac{\partial}{\partial z} E_x \\ -\frac{\partial}{\partial y} E_x \end{pmatrix}, \tag{5.1}$$

where E_x is a function of y and z. We denote the airfoil by Ω and the coating layer by Ω_1; see Figure 7.

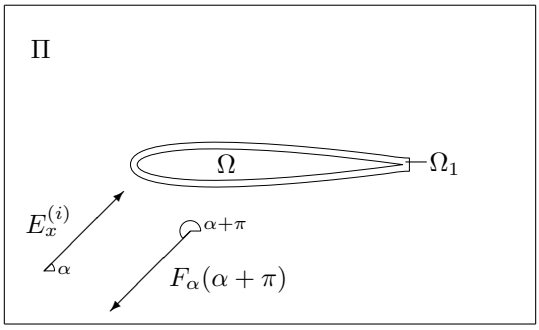


FIG. 7. *The computational domain, an interrogating wave $E_x^{(i)}$ with an interrogation angle α, and a far field pattern $F_\alpha(\alpha+\pi)$ of a scattered field $E_x^{(r)}$ to the direction $\alpha+\pi$.*

We decompose the total electric field E_x into two fields $E_x^{(i)}$ and $E_x^{(r)}$; that is, $E_x = E_x^{(r)} + E_x^{(i)}$. Furthermore, $E_x^{(i)}$ is chosen to coincide with the free space interrogating plane wave everywhere, i.e., in $\mathbb{R}^2 \backslash \bar{\Omega}$. With an interrogation angle $\alpha = (\pi/2 - \phi_0)$, $E_x^{(i)}$ is given by $E_x^{(i)}(y,z) = e^{i(k_y y + k_z z)}$, where $k_y = k\cos\alpha$, $k_z = k\sin\alpha$, $k = 2\pi/\lambda = \omega/c_0$ is the wave number, and λ is the wavelength. Thus, $E_x^{(i)}$ is the incident field outside the coated airfoil $\bar{\Omega} \cup \bar{\Omega}_1$, while in the coating layer Ω_1 the field $E_x^{(i)}$ is *nonphysical*. From this it follows that $E_x^{(r)}$ is the scattered field outside the coated airfoil, while neither of the fields $E_x^{(i)}$ and $E_x^{(r)}$ by itself has a physical meaning in the coating layer. That is, in Ω_1 they are simply a convenient computational decomposition of the total field and do not represent the individual coating-modified incident fields and reflected fields, respectively.

By eliminating the magnetic field from the Maxwell equation and substituting the time-harmonic electric field E of the form given in (5.1) into the resulting equation, we obtain the following Helmholtz equation:

$$\begin{aligned} \nabla \cdot \left(\frac{1}{\mu} \nabla E_x^{(r)} \right) + \epsilon\omega^2 E_x^{(r)} &= -\nabla \cdot \left(\frac{1}{\mu} \nabla E_x^{(i)} \right) - \epsilon\omega^2 E_x^{(i)} && \text{in } \mathbb{R}^2 \setminus \bar{\Omega}, \\ E_x = E_x^{(r)} + E_x^{(i)} &= 0 && \text{on } \partial\Omega, \\ \left[\frac{1}{\mu} \frac{\partial E_x}{\partial n} \right] = [E_x] &= 0 && \text{on } \partial\Omega_1 \setminus \partial\Omega, \end{aligned} \tag{5.2}$$

$$\lim_{r\to\infty} \sqrt{r}\left(\frac{\partial E_x^{(r)}}{\partial r} - ikE_x^{(r)}\right) = 0,$$

where $[\,\cdot\,]$ denotes the jump and n is a normal direction of the surface $\partial\Omega_1 \setminus \partial\Omega$. The far field behavior of the scattered field described by the Maxwell equation satisfies the Silver–Müller radiation condition [7]. For the time-harmonic TM_x mode this condition reduces to be the Sommerfeld radiation condition given by the limit in (5.2). The material permittivity ϵ and permeability μ are piecewise constant functions defined by

$$\epsilon(y,z) = \begin{cases} \epsilon_r\epsilon_0, & (y,z) \in \bar{\Omega}_1, \\ \epsilon_0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \mu(y,z) = \begin{cases} \mu_r\mu_0, & (y,z) \in \bar{\Omega}_1, \\ \mu_0 & \text{otherwise.} \end{cases}$$

5.2. Far field pattern. The far field pattern $F : [0, 2\pi] \to \mathbb{C}$ describes the intensity and phase of the scattered field $E_x^{(r)}$ far away from the scatterer [6, p. 340]. It can be defined as

$$F_\alpha(\theta) = \lim_{r\to\infty} \left(\sqrt{8\pi kr}\; e^{-i(kr+\pi/4)}\; E_x^{(r)}(r\cos\theta, r\sin\theta)\right), \tag{5.3}$$

where we have added the subscript α to denote the interrogation angle.

5.3. Backscatter reduction. In this optimization problem, we want to find constant material parameters ϵ_r and μ_r so that the intensity of the backscattered wave is minimized over a given sector $[\alpha_0, \alpha_1]$. The objective function is the integral

$$J(\epsilon_r, \mu_r) = \int_{\alpha_0}^{\alpha_1} |F_\alpha(\alpha + \pi)|^2 \, d\alpha. \tag{5.4}$$

The minimization problem is given by

$$\min_{(\epsilon_r,\mu_r)\in Q} J(\epsilon_r, \mu_r), \tag{5.5}$$

where Q is the set of admissible material parameters. In our considerations here, the set Q is chosen so that each material parameter belongs to a given interval of feasible values.

This objective function corresponds to a situation where the same radar is illuminating and detecting the scattered wave. The interrogation angle of the interrogating wave varies within the interval $[\alpha_0, \alpha_1]$ and the formulation (5.4) corresponds to an assumption of a uniform distribution on possible angles of interrogation.

We use the NAG Fortran Library's [19] E04UCF implementation of a sequential quadratic programming (SQP) method. This is a gradient-based optimization method which approximates the gradient using finite differences.

6. Approximation.

6.1. Truncation of domain and variational formulation. For the discretization of (5.2), we restrict the problem to a rectangular domain Π and impose a second-order absorbing boundary condition [1] on the artificial boundary $\partial\Pi$ to approximate the Sommerfeld radiation condition. Now the scattered field $E_x^{(r)}$ satisfies the follow-

ing equations:

$$
\begin{aligned}
\nabla\cdot\left(\frac{1}{\mu}\nabla E_x^{(r)}\right)+\epsilon\omega^2 E_x^{(r)} &= -\nabla\cdot\left(\frac{1}{\mu}\nabla E_x^{(i)}\right)-\epsilon\omega^2 E_x^{(i)} && \text{in } \Pi\setminus\bar{\Omega},\\
E_x^{(r)} &= -E_x^{(i)} && \text{on } \partial\Omega,\\
\left[\frac{1}{\mu}\frac{\partial E_x}{\partial n}\right] &= [E_x]=0 && \text{on } \partial\Omega_1\setminus\partial\Omega,\\
\frac{\partial E_x^{(r)}}{\partial n}-ikE_x^{(r)}-\frac{i}{2k}\frac{\partial^2 E_x^{(r)}}{\partial s^2} &= 0 && \text{on } \partial\Pi,\\
\frac{\partial E_x^{(r)}}{\partial s}-ik\frac{3}{2}E_x^{(r)} &= 0 && \text{at } C,
\end{aligned}
\tag{6.1}
$$

where n and s denote the normal and tangential directions of the boundary $\partial\Pi$, respectively, and C is the set of the corner points of Π.

The variational formulation of (6.1) is as follows:
Find $E\in\{v\in H^1(\Pi\setminus\bar{\Omega})\mid v|_{\partial\Pi}\in H^1(\partial\Pi),\ v=-E_x^{(i)} \text{ on } \partial\Omega\}$ such that

$$
\begin{aligned}
&\int_{\Pi\setminus\bar{\Omega}}\left(\frac{1}{\mu}\nabla E_x^{(r)}\cdot\nabla v-\epsilon\omega^2 E_x^{(r)}\,v\right)d\xi+\frac{i}{k\mu_0}\int_{\partial\Pi}\left(\frac{1}{2}\frac{\partial E_x^{(r)}}{\partial s}\frac{\partial v}{\partial s}-k^2E_x^{(r)}\,v\right)d\sigma\\
&+\frac{3}{4\mu_0}\sum_{(y,z)\in C}E_x^{(r)}(y,z)v(y,z)=-\int_{\Pi\setminus\bar{\Omega}}\left(\frac{1}{\mu}\nabla E_x^{(i)}\cdot\nabla v-\epsilon\omega^2E_x^{(i)}\,v\right)d\xi
\end{aligned}
\tag{6.2}
$$

for all $v\in\{v\in H^1(\Pi\setminus\bar{\Omega})\mid v|_{\partial\Pi}\in H^1(\partial\Pi),\ v=0 \text{ on } \partial\Omega\}$.

6.2. Finite element approximation. A finite element approximation as developed in [8, 9] was implemented using linear elements. The techniques in [8] and [9] are two-dimensional finite element methods for scattered electromagnetic solutions from noncoated and coated objects, respectively. We employ the same methods here. The mesh is constructed from two uniform triangular meshes. The finer mesh is for the coating layer, and the coarser is for the exterior domain outside the coating. The mesh step sizes are chosen in such a way that the number of nodes per wavelength is approximately the same in the air and in the coating. The finer mesh is locally fitted to the surfaces of the obstacle and the coating layer. The local fitting is done using the algorithm [5] with slight modifications. Between the meshes, there is a layer that fits the meshes together in a conforming way. An example of a part of a mesh is shown in Figure 8.

After the discretization of the variational formulation (6.2), we obtain a system of linear equations

$$Ax=b, \tag{6.3}$$

where the matrix A is a symmetric non-Hermitian complex matrix. The vector x contains the nodal values of the scattered field $E_x^{(r)}$, and the vector b corresponds to the right-hand terms in (6.2). The resulting systems of linear equations are solved using an iterative method which combines fictitious domain and domain decomposition methods [9].

The far field pattern $F_\alpha(\alpha+\pi)$ in (5.3) is computed as a surface integral [7] of the computed near field $E_x^{(r)}$ and its flux. Our particular implementation of the

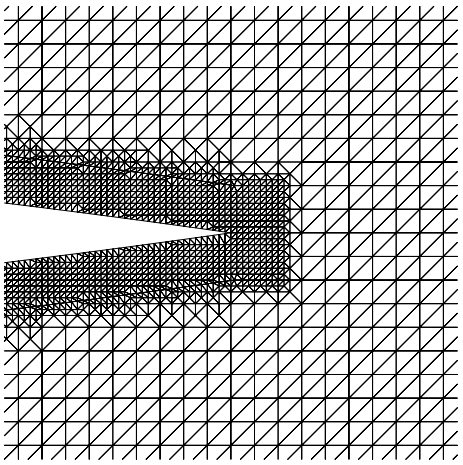


FIG. 8. *A magnified view of the mesh for a coated NACA*0012 *airfoil.*

computations is described in [8]. We need to evaluate the far field pattern $F_\alpha(\alpha+\pi)$ in (5.3) using the computed near field $E_x^{(r)}$. Our particular implementation of the computations is described in [8]. The basic idea of this procedure is the following. Let $\tilde{E}_x^{(r)}$ be the harmonic extension of $E_x^{(r)}$ from $\Pi \setminus (\bar{\Omega} \cup \Omega_1)$ to $\mathbb{R}^2$. We obtain this extension as a byproduct of our solution procedure. Then at the point η outside $\bar{\Omega} \cup \Omega_1$ the scattered field $\tilde{E}_x^{(r)}$ is given by

$$\tilde{E}_x^{(r)}(\eta) = \int_\Pi \left(\Delta \tilde{E}_x^{(r)}(\xi) + k^2 \tilde{E}_x^{(r)}(\xi) \right) \Phi(\eta, \xi)\, d\xi, \tag{6.4}$$

where $\Phi(\eta,\xi)$ is the fundamental solution of the homogeneous Helmholtz equation in $\mathbb{R}^2$ given by the Hankel function $\Phi(\eta,\xi) = \frac{i}{4} H_0^{(1)}(k|\eta-\xi|)$. By using Green's formula one can show that (6.4) is equivalent to

$$\tilde{E}_x^{(r)}(\eta) = \int_{\partial\Pi} \left(E_x^{(r)}(\sigma) \frac{\partial \Phi(\eta,\sigma)}{\partial n} - \frac{\partial E_x^{(r)}(\sigma)}{\partial n} \Phi(\eta,\sigma) \right) d\sigma,$$

which is a more traditional expression for $\tilde{E}_x^{(r)}(\eta)$ [7]. The far field pattern F_α is obtained by first discretizing the Helmholtz operator and integral in (6.4) and then taking the limit in (5.3). At the discrete level the previous procedure reduces to the evaluation of a sum of exponential functions. This can be performed easily and quickly.

We have compared numerical results computed using the proposed method and implementation with test cases presented at a workshop in Oxford (see [20]). The two-dimensional test cases did not include a coated NACA airfoil, but they did include a similar coated ogive (the intersection of two nonconcentric disks). We computed numerical results for 10 test cases. All of our numerical results and especially the radar cross sections were in very good agreement with the majority of the results presented in the workshop for each test case. Because of this, we expect our results to be accurate also for the coated NACA airfoils.

7. Numerical experiments. In our experiments, we minimize the backscatter by a coated NACA0012 airfoil. The length of the airfoil is one unit without coating,

TABLE 1
The results of material parameter optimizations for the case $\lambda = 1/4$.

	Real part			Imaginary part			
	min	opt	max	min	opt	max	$J(\epsilon_r, \mu_r)$
ϵ_r	1	1	1	0	0	0	
μ_r	1	1	1	0	0	0	324.340
ϵ_r	1	3.92	10	0	0	0	
μ_r	1	1	10	0	0	0	318.738
ϵ_r	1	6.52	10	0	3.14	10	
μ_r	1	1	1	0	0	0	2.918
ϵ_r	1	7.75	10	0	0.80	10	
μ_r	1	5.41	10	0	2.51	10	0.083

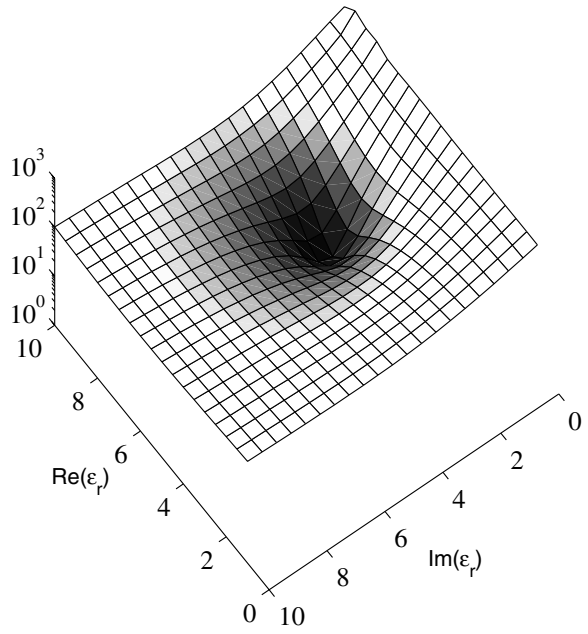


FIG. 9. *The objective function $J(\epsilon_r, \mu_r)$ for complex-valued ϵ_r and $\mu_r = 1$ for the case $\lambda = 1/4$.*

and the trailing edge of the perfectly conducting material is at the origin. We minimize the backscatter for the interrogation angles in the sector $[\alpha_0, \alpha_1] = [0, \pi]$. We considered two wavelengths, $\lambda = 1/4$ and $\lambda = 1/10$. The thickness of the coating is taken as $\lambda/10$ to provide a comparison of coating layers of different thicknesses possessing similar material parameters.

We consider first the lower frequency experiments with a four-wavelength-long airfoil. The computational domain is $[-1.5, 0.5] \times [-0.6, 0.6]$. Our discretization has 20 nodes per wavelength in the ambient medium, leading to a triangulation with 22157 nodes and 43158 elements. A magnified view of the mesh at the trailing edge is shown in Figure 8. We perform several optimizations with different box constraints for the real and imaginary parts of ϵ_r and μ_r. The results of these optimizations are given in Table 1. A surface plot of the objective function is shown in Figure 9. The radar

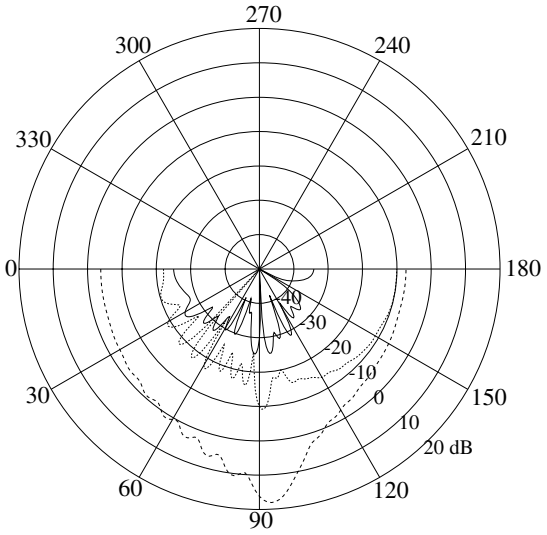


FIG. 10. *The RCS for optimized complex-valued material parameters ($\epsilon_r = 7.75 + 0.80i$; $\mu_r = 5.41 + 2.51i$; solid line), optimized complex-valued permittivity ($\epsilon_r = 6.52 + 3.14i$; $\mu_r = 1$; dotted line), and for no coating ($\epsilon_r = 1$; $\mu_r = 1$; dashed line) for the case $\lambda = 1/4$.*

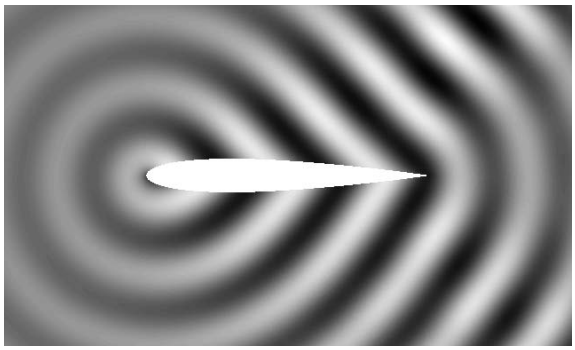


FIG. 11. *The reflected field $E_x^{(r)}$ for no coating ($\epsilon_r = 1$; $\mu_r = 1$) for the case $\lambda = 1/4$ and angle of interrogation $\alpha = \pi/4$.*

cross sections defined by

$$\mathrm{RCS}(\alpha) = 10 \log_{10} \left(\frac{1}{8\pi} |F_\alpha(\alpha + \pi)|^2 \right)$$

are shown for two optimized materials in Figure 10. Corresponding reflected field intensities for comparison between the no coating layer case and the optimized complex-valued parameters case of Figure 10 are depicted in Figures 11 and 12 for an angle of interrogation $\alpha = \pi/4$.

The computational domain for the higher frequency experiments with a 10-wavelength-long airfoil is $[-1.3, 0.3] \times [-0.4, 0.4]$. Again our discretization has 20 nodes per wavelength in the ambient medium, leading to a triangulation with 65551 nodes and 128530 elements. The results of optimizations with different box constraints are given

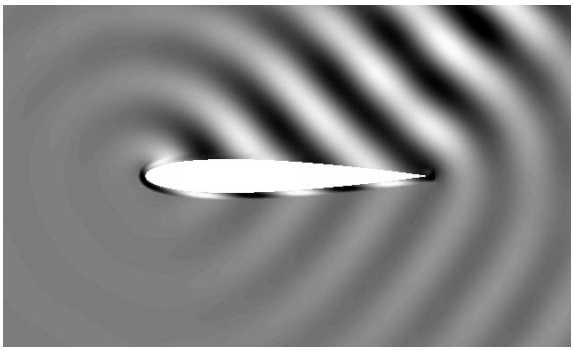


FIG. 12. *The reflected field* $E_x^{(r)}$ *for optimized complex-valued material parameters* ($\epsilon_r = 7.75 + 0.80i$; $\mu_r = 5.41 + 2.51i$) *for the case* $\lambda = 1/4$ *and angle of interrogation* $\alpha = \pi/4$.

TABLE 2
The results of material parameter optimizations for the case $\lambda = 1/10$.

	Real part			Imaginary part			
	min	opt	max	min	opt	max	$J(\epsilon_r, \mu_r)$
ϵ_r	1	1	1	0	0	0	
μ_r	1	1	1	0	0	0	806.036
ϵ_r	1	3.84	10	0	0	0	
μ_r	1	1	10	0	0	0	802.695
ϵ_r	1	6.58	10	0	3.14	10	
μ_r	1	1	1	0	0	0	2.901
ϵ_r	1	4.54	10	0	2.99	10	
μ_r	1	4.99	10	0	2.92	10	0.234

in Table 2. Radar cross sections for two optimized materials are shown in Figure 13. Corresponding reflected field intensities for comparison between the no coating layer case and the optimized complex-valued parameters case of Figure 13 are depicted in Figures 14 and 15 for an angle of interrogation $\alpha = \pi/4$.

8. Summary and conclusions. In summary, the efforts reported on in this paper are an important first step in developing an attenuation or anti-interrogation technology. We first considered the question of the feasibility of reduction in reflected electromagnetic waves from a planar layered coating on a perfect conductor. We demonstrated that even under uncertainty of the interrogating wavelengths (frequencies), one can achieve reduction of the reflection coefficient through optimizing the dielectric permittivity in a coating layer. We then considered a Maxwell equation-based formulation for a composite ferromagnetic-ferroelectric device built and experimentally tested by How and Vittoria. We derived the pertinent reflection field equations for time-harmonic interrogating TM_x mode plane waves and showed that substantial control of reflected waves (both magnitude and phase) can be obtained by tuning the magnetic permeability of a ferrite layer. We next turned to a nonplanar geometry, in this case a two-dimensional airfoil, with a coating layer wherein both the dielectric permittivity ϵ and the magnetic permeability μ can be optimized. Allowing a uniform uncertainty on the interrogating signal angles, we use computational methods from two-dimensional scattering theory (the Helmholtz equation with Sommerfeld far field radiation conditions) to verify that significant reduction in the far field reflection can be obtained by an optimal choice of ϵ and μ.

In ongoing efforts [3], we are continuing the investigations begun in this paper in

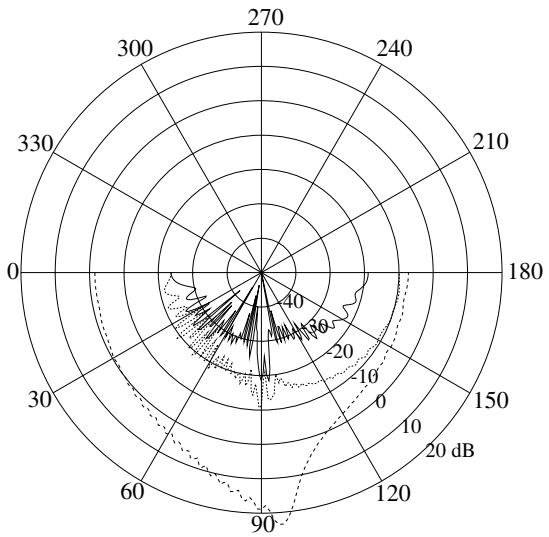


FIG. 13. *The RCS for optimized complex-valued material parameters* ($\epsilon_r = 4.54 + 2.99i$; $\mu_r = 4.99 + 2.92$*i; solid line), optimized complex-valued permittivity* ($\epsilon_r = 6.58 + 3.14i$; $\mu_r = 1$*; dotted line), and for no coating* ($\epsilon_r = 1$; $\mu_r = 1$*; dashed line) for the case* $\lambda = 1/10$.

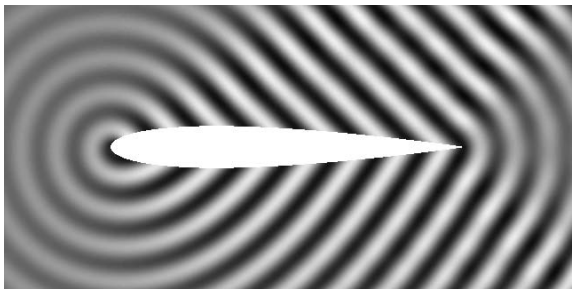


FIG. 14. *The reflected field* $E_x^{(r)}$ *for no coating* ($\epsilon_r = 1$; $\mu_r = 1$*) for the case* $\lambda = 1/10$ *and angle of interrogation* $\alpha = \pi/4$.

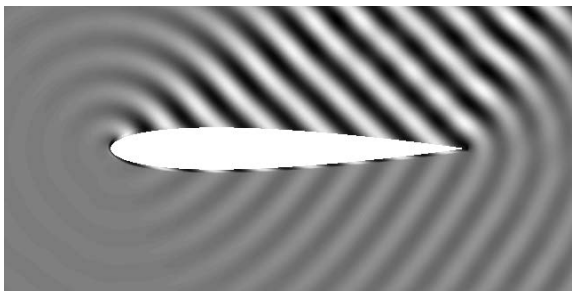


FIG. 15. *The reflected field* $E_x^{(r)}$ *for optimized complex-valued material parameters* ($\epsilon_r = 4.54 + 2.99i$; $\mu_r = 4.99 + 2.92i$*) for the case* $\lambda = 1/10$ *and angle of interrogation* $\alpha = \pi/4$.

several directions. One includes allowing optimal design for the coating over ranges (distributions) of interrogating wavelengths and angles of incidence. We have also begun consideration of countermeasures for the anti-interrogation ideas outlined in this paper. These problems result in classical *static zero-sum two-player games* for evader and interrogator where each player must consider uncertainty in knowledge of the opposing player's capabilities. The resulting min-max games must be played over spaces of probability measures (see [3] for details of these interesting but challenging problems and our efforts with them).

All of the investigations discussed in this paper were pursued under an *active design* scenario and did not allow for online adaptivity of the coating layers. Future investigations of great interest include the feasibility of combining the formulations in this paper with real time sensing and adaptive (feedback) control of coatings such as those described above to develop an *active control* attenuation capability.

Acknowledgments. The authors would like to thank Dr. Richard Albanese of the AFRL, Brooks AFB, San Antonio, TX, for numerous stimulating and valuable comments and suggestions on this work. They are also grateful to Dr. Hoton How for several helpful conversations.

REFERENCES

[1] A. Bamberger, P. Joly, and J. E. Roberts, *Second-order absorbing boundary conditions for the wave equation: A solution for the corner problem*, SIAM J. Numer. Anal., 27 (1990), pp. 323–352.

[2] H. T. Banks, M. W. Buksas, and T. Lin, *Electromagentic Material Interrogation Using Conductive Interfaces and Acoustic Wavefronts*, Frontiers Appl. Math. 21, SIAM, Philadelphia, 2000.

[3] H. T. Banks, K. Ito, and J. A. Toivanen, *Determination of Interrogating Frequencies to Maximize Electromagnetic Backscatter from Objects with Material Coatings*, Technical report CRSC-TR05-30, Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, 2005.

[4] S. Beeler, H. T. Tran, and N. Dietz, *Representation of GaP formation by a reduced order surface kinetics model using p-polarized reflectance measurements*, J. Appl. Phys., 86 (1999), pp. 674–682.

[5] C. Börgers, *A triangulation algorithm for fast elliptic solvers based on domain imbedding*, SIAM J. Numer. Anal., 27 (1990), pp. 1187–1196.

[6] D. Colton, *The inverse scattering problem for time-harmonic acoustic waves*, SIAM Rev., 26 (1984), pp. 323–350.

[7] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer-Verlag, New York, 1998.

[8] E. Heikkola, Yu. A. Kuznetsov, P. Neittaanmäki, and J. Toivanen, *Fictitious domain methods for the numerical solution of two-dimensional scattering problems*, J. Comput. Phys., 145 (1998), pp. 89–109.

[9] E. Heikkola, T. Rossi, and J. Toivanen, *A domain decomposition technique for two-dimensional scattering problems with coated obstacles*, in Innovative Tools for Scientific Computation in Aeronautical Engineering, J. Periaux, P. Joly, O. Pironneau, and E. Onate, eds., CIMNE, Barcelona, Spain, 2001, pp. 183–194.

[10] H. How and C. Vittoria, *Implementation of Microwave Active Nulling*, private communication.

[11] H. How and C. Vittoria, *Microwave impedance control over a ferroelectric boundary layer*, IEEE Trans. Microwave Theory Tech., 52 (2004), pp. 2177–2182.

[12] H. How, X. Zuo, E. Hokanson, L. Kempel, and C. Vittoria, *Calculated and measured characteristics of a microstrip fabricated on a Y-type hexaferrite substrate*, IEEE Trans. Microwave Theory Tech., 50 (2002), pp. 1280–1288.

[13] K. Ito and K. Kunisch, *Sensitivity analysis of solutions to optimization problems in Hilbert spaces with applications to optimal control and estimation*, J. Differential Equations, 99 (1992), pp. 1–40.

[14] K. Ito and F. Reitich, *A high-order perturbation approach to profile reconstruction.* I. *Perfectly conducting gratings*, Inverse Problems, 15 (1999), pp. 1067–1085.

[15] J. D. Jackson, *Classical Electrodynamics*, Wiley & Sons, New York, 1975.
[16] C. L. Ladson, C. W. Brooks, Jr., A. S. Hill, and D. W. Sproles, *Computer Program to Obtain Ordinates for NACA Airfoils*, NASA Technical Memorandum 4741, NASA, Langley Research Center, Hampton, VA, 1996.
[17] B. Lax and K. J. Button, *Microwave Ferrites and Ferromagnetics*, McGraw-Hill, New York, 1962.
[18] E. L. B. El-Sharawy and R. W. Jackson, *Coplanar waveguide and slot line on magnetic substrates: Analysis and experiment*, IEEE Trans. Microwave Theory Tech., 36 (1988), pp. 1071–1079.
[19] *The NAG Fortran Library Manual: Mark* 20, The Numerical Algorithms Group, Oxford, UK, 2002.
[20] *Technical Description of Workshop on Approximations and Numerical Methods for the Solution of the Maxwell Equations*, IMA (UK), GAMNI/SMAI (France), and Oxford University Computing Laboratory (UK), Oxford, UK, 1995.

SIAM J. APPL. MATH.
Vol. 66, No. 3, pp. 1050–1079

PREDATOR-PREY INTERACTIONS WITH DELAYS DUE TO JUVENILE MATURATION*

KENNETH L. COOKE†, RICHARD H. ELDERKIN†, AND WENZHANG HUANG‡

Abstract. This paper focuses on predator-prey models with juvenile/mature class structure for each of the predator and prey populations in turn, further classified by whether juvenile or mature individuals are active with respect to the predation process. These models include quite general prey recruitment at every stage of analysis, with mass action predation, linear predator mortality as well as delays in the dynamics due to maturation. As a base for comparison we briefly establish that the similar model without delays cannot support sustained oscillation, but it does have predator extinction or global approach to predator-prey coexistence depending on whether the ratio α of per predator predation at prey carrying capacity to the predator death rate is less than or greater than one.

Our first model shows the effect of introducing an invulnerable juvenile prey class, appropriate, e.g., for some host-parasite interactions. In contrast our second model shows the effect of limiting predation to a prey juvenile class. Finally, in a third model we consider an inactive juvenile predator class, which would be appropriate for many traditional situations in which the generation time for the predator is significantly larger than that of the prey. In all cases the introduction of a juvenile class results in a system of three delay-differential equations from which the two equations for the mature class and the nonstructured class can be decoupled. We obtain some global stability results and identify a parameter α, similar to the α of the unlagged model, which determines whether or not the predator is driven to extinction. With $\alpha > 1$, and considering the maturation age of the juvenile class as a bifurcation parameter, we obtain Hopf bifurcations in our second and third models, while in the case of juvenile prey (in the first model) the unique coexistence equilibrium remains stable for all positive delays. Although the delay is "physically present" in all three models, we obtain scaled, nondimensional replacement models with that physical presence scaled out. After analyzing the scaled equations we show that all our results hold for the original models.

We pursue the bifurcation in the inactive juvenile predator model with numerical simulations. Strikingly similar results over a variety of birth functions are observed. Increases of the maturation delay first produce Hopf bifurcation from steady state to periodic behavior. Even further increase in the delay produces instabilities of the bifurcating periodic solutions with corresponding interesting geometry in a two-dimensional plot of period vs. delay.

Key words. predator-prey, host-parasite, age structure, maturation delay, delay differential equations, bifurcation

AMS subject classifications. 92D25, 34K60, 34K18

DOI. 10.1137/05063135

1. Introduction.

It is well known that some predators may preferentially attack prey of certain ages or developmental stages. Likewise, the predators themselves may be distinguished in some cases between inactive juveniles and active adults. These situations may be modeled mathematically by dividing the populations into age classes. Our models of this are closely related to those of Hastings ([16], 1983), Murdoch et al. ([22], 1987), the very general ones of Nunney ([25],[26],[27]), and recent work of Gourley and Kuang ([13], 2004). In a model comparable to one of ours (5), Hastings uses a linear mature prey birth rate and a general predation functional

*Received by the editors May 11, 2005; accepted for publication (in revised form) August 31, 2005; published electronically March 10, 2006.
http://www.siam.org/journals/siap/66-3/63135.html
†Department of Mathematics, Pomona College, Claremont, CA 91711 (klc04747@pomona.edu, relderkin@pomona.edu).
‡Department of Mathematical Sciences, University of Alabama in Huntsville, Huntsville, AL 35899 (huang@ultra.math.uah.edu). This research was supported in part by NSF grant DMS-0204676.

response to show the possibility of switches between stability and instability of the positive equilibrium as the time delay changes. In contrast we show that this is not possible for our model (5) with its nondecreasing birth rate $NB(N)$. It is known [8] that if $B(N)$ does not satisfy this hypothesis, more complicated dynamic behavior may occur even in a single species population model. In order to analyze models with nonlinear delayed recruitment, we limit our attention to a mass action (i.e., Holling Type I) predation response.

Nunney studied prey-predator systems which are similar to ours, but without the "physical presence" of the delay (that is, occurring outside the arguments of the time-varying states). Assuming the existence of a unique positive equilibrium, he derives a condition for its absolute stability (i.e., linear stability for all delays R). For our "host-parasite" model (section 3) this condition is satisfied; however, in addition we identify a parameter, α, which governs stability of the equilibrium with extinction of the predator. Nunney also considers a similar general system, but in which the delay is due to predator maturation, including our scaled system (50) as special case. Again he assumes existence of a positive equilibrium, and derives a condition for its absolute stability. However, in contrast, with our more specific functional forms we derive conditions on parameters which guarantee Hopf bifurcation to periodic solutions, building on the analysis of Cooke and van den Driessche [7].

Gourley and Kuang consider a model with inactive juvenile predators which is similar to ours in section 5, but with logistic prey recruitment, which we generalize. As in our case, they are faced with the physical presence of the delay in the equations, which causes them (and others) significant difficulty. We exhibit a scaling of the states and parameters which avoids those difficulties (e.g., (8)), yet causes no loss of detail or generality (section 7). Although their predation functional response is initially more general than ours, they specialize, as do we, to the mass-action response before deriving any results. Using our scaling, we are able to analytically produce results, while they relied on computational assistance for theirs. We also implement three examples of birth functions (affine, concave up, and concave down) and conduct numerical experimentation. Over parameter ranges of biological interest the numerical results are very similar across the three function types. However, by extending the maturation delay even further we first find interesting geometries of the bifurcated periodic solutions and at even greater delays find an apparent relation between onset of instabilities and a curious lack of monotonicity in a bifurcation diagram.

We intend our results to be a coherent study across a selection of models that deserve to be considered in the context of each other. We present theoretical results that can be compared across the variation of juvenile/adult roles, both active and inactive, and for both predator and prey. In studying these models, we seek to obtain as much information as we can about how the dynamics of the systems depend on the multiple parameters in the equations, such as attack and mortality rates and the maturation delay. In general, we are looking for conditions that ensure stable equilibrium or bifurcation phenomena. Basic questions are the following: (1) to what extent does the inclusion of the natural mortality parameters alter the qualitative or quantitative behavior of the systems as the maturation delay is varied? (2) is maturation delay more destabilizing in the prey or in the predator? (3) what kinds of destabilizations other than Hopf bifurcations occur as the maturation delay increases? We give partial answers to each of these questions and point the way to further investigation in each case. In general, we find that there is a considerable difference in dynamics depending on whether the prey or predator has a differently behaving

juvenile class, and on whether the adults or the juveniles are active with respect to the predation process. A moral is that any theoretical or practical study of these situations should carefully take account of the potential for such differences.

We begin the analysis of each of our models by scaling the equations to obtain a dimensionless system in which the delay is no longer physically present. This has the usual advantages of reducing the number of parameters and making the mathematical analysis somewhat simpler and more transparent. Implicit in this strategy is the expectation that the dynamics of the original and scaled systems will be "equivalent" or that the dynamics of one system will be "mirrored" by the dynamics of the other system. Since some of the coefficient parameters of the original system, and some of the transformations in the scaling, depend on the delay, and since we are probing how stability or destabilization depends on the delay, we have included a careful discussion of these and other aspects in section 7.

The structure of the paper is as follows. In section 2, as a point of departure for what is to come in subsequent sections, we consider a simple predator-prey model that does not include any age structure or resulting delay in recruitment. It consists of two ordinary differential equations in which there is a general prey recruitment function, mass-action predation, and linear mortality in both prey and predator. A simple and natural condition suffices to determine global asymptotic stability of equilibrium; there are no positive periodic solutions.

In section 3 we propose and analyze a model in which the prey has a juvenile class (in addition to an interactive adult class) that is invulnerable to predation while the predator is considered as a single class. For example, this might apply to the sheltered existence of human infants while breast-feeding in a world where parasitism is widespread. The results for our equivalent nondimensionalized system may be stated in terms of a single scaled parameter, α, which has a natural interpretation similar to that of R_0 in many population (especially epidemiological) models. In this case, the system is always dissipative, and when $\alpha > 1$ there is a unique equilibrium with both populations present, which must be locally asymptotically stable, independent of the delay. There is no bifurcation to periodic behavior as the delay increases. Later, in section 7, we discuss what this tells us about the original, nonscaled system.

In section 4 we present a model similar to the one in the preceding section in that there are juvenile and adult prey along with a single class of predators. However, we now assume that survival of juvenile prey is reduced in proportion to the mean population size of predators. Under this hypothesis, our system is now one with "distributed delay." In contrast with the previous case, it turns out that the positive equilibrium is stable for small maturation delays, but unstable with bifurcation to periodic behavior for large ones.

In section 5 we present a model in which there is a single prey population, but a predator with adult and juvenile classes in which the latter do not attack the prey. We show in this case that bifurcation is possible with periodic solutions emerging for large values of the delay under an additional condition on parameters. Here we calculate bifurcation diagrams for examples of the original and the scaled systems, and later in section 7, we discuss more general relationships.

In section 6 we present results of numerical studies of the bifurcations established in section 5. Before chosing parameter values, we briefly provide interpretations for the more important ones. Then we numerically compare a variety of birth functions in which one is concave down, another concave up, and yet another is affine (over the population range of interest). We arrange that all three resulting models have the same

interior equilibrium populations and the same populations that are analogous to a carrying capacity (in ODE models), and have the same birth rates at those population levels. We compare the three over a common range of parameters, finding strong similarities. In the case of the affine birth function we extend the Hopf bifurcation branch with increasing delay until a Floquet multiplier leaves the unit disk in the complex plane, signifying onset of instability of the bifurcating periodic solutions. Observing this in a two-dimensional plot of the period of the bifurcating solutions vs. the delay, we see this onset of instability simultaneously as the delay changes from increase to decrease along a backwards S shape just after the period has begun to decrease. At the other end of the backwards S where the delay begins increasing again, another multiplier leaves the unit disk and the period moves toward increase.

Section 7 addresses the correspondence between the nature of the original models and our analysis of their nondimensional replacements. This analysis is especially motivated by the appearance of the maturation delay physically in the predation coefficients as well as in the populations in the original models, but only within the arguments of the scaled populations in the scaled equations. We show in a precise sense that no bifurcation structure is lost as a result of these scalings.

2. A prototypical model. In order to provide a basis for comparison of our primary results on models with delays due to maturation, we first establish the basic properties exhibited by our model without age structure or corresponding delay. The basic ODE model for a prey population N and a predator population P is

$$\frac{dN}{dT} = NB(N) - aNP - dN, \tag{1a}$$

$$\frac{dP}{dT} = cNP - d_P P \tag{1b}$$

in which we assume that the per capita prey birth rate B satisfies

$$B(N) \geq 0 \quad \text{and} \quad B'(N) < 0 \qquad \text{for } N \geq 0, \tag{2a}$$

$$B(0) > d > B(\infty). \tag{2b}$$

The assumptions (2a) and (2b) are satisfied, for example, in each of the forms $B(N) = p/(q+N)$ and $B(N) = \exp(-pN)$ where p and q are positive constants (for appropriate d). However see section 3 where the latter form will not work. Furthermore, although (2a) is not satisfied by $B(N) = p - qN$, we adapt this form later to provide a viable example.

It is obvious that (1) always has a trivial equilibrium $(N, P) = (0, 0)$ which is an unstable saddle point. Another equilibrium is $(N_0, 0)$, where $N_0 > 0$ is the unique solution of $B(N) = d$, existing by (2). Some simplification of notation can be achieved by scaling N, P, T and various coefficients. In fact, if we set

$$x = N/N_0, \qquad t = d_P T, \qquad \alpha = cN_0/d_P,$$
$$y = aP/cN_0, \qquad b(x) = B(xN_0)/d_P, \qquad \gamma = d/d_P,$$

then the system (1) takes the (nondimensional) form

$$\frac{dx}{dt} = xb(x) - \alpha xy - \gamma x, \tag{3a}$$

$$\frac{dy}{dt} = \alpha xy - y. \tag{3b}$$

Notice that in this scaling the transfer from x to y by the scaled predation is perfectly efficient! The properties assumed in (2) take the simpler form

$$b(x) \geq 0 \quad \text{and} \quad b'(x) < 0 \qquad \text{for } x \geq 0, \tag{4a}$$

$$b(0) > \gamma > b(\infty) \quad \text{and} \quad b(1) = \gamma \tag{4b}$$

and the equilibria are now at $(0,0)$ and $(1,0)$.

LEMMA 1. *The system* (3) *is dissipative, that is, there is a compact set* Ω *(in this case of the form* $\{(x,y) : x, y \geq 0,\ x + y \leq m\}$ *for some* m*) such that each solution in the first quadrant of the* (x,y) *plane has a* T_0 *such that when* $t \geq T_0$, $(x(t), y(t)) \in \Omega$.

Proof. With $V(x,y) = x + y$ we have

$$\dot{V} \stackrel{def}{=} \frac{d}{dt} V(x(t), y(t)) = xb(x) - \gamma x - y.$$

When $x > 1$ this is negative, and when $x \leq 1$ it is negative for sufficiently large y. Standard Lyapunov function considerations complete the argument. ∎

THEOREM 2. *If* $\alpha \leq 1$*, then all positive solutions of* (3) *converge to the equilibrium* $(1,0)$ *as time* $t \to \infty$.

Proof. The first quadrant of the (x,y) plane is invariant and it follows from the first lemma that every positive solution is bounded. So Poincaré-Bendixson considerations obtain. There is no positive equilibrium; $(0,0)$ is a saddle point and its stable manifold lies on the y-axis; and $(1,0)$ is linearly stable. It follows then that $(x(t), y(t)) \to (1,0)$ as $t \to \infty$. ∎

It is easily seen that our hypotheses on the birth function $B(x)$ imply the following lemma.

LEMMA 3. *The system* (3) *has a positive equilibrium* (x^*, y^*) *if and only if* $\alpha > 1$*, in which case it is unique.*

For the remainder of this section we study the case when $\alpha > 1$. Since $b'(x^*) < 0$, it is easy to see that all eigenvalues of the Jacobian matrix J at (x^*, y^*) have negative real part, and hence (x^*, y^*) is locally asymptotically stable. However, we can establish the stronger result.

THEOREM 4. *If* $\alpha > 1$*, then* (x^*, y^*) *is globally asymptotically stable.*

Proof. We first establish that any positive nonconstant periodic solution of the system (3) must be asymptotically stable. Suppose that $(\tilde{x}(t), \tilde{y}(t))$ is a positive periodic solution and let $X' = A(t)X$ be the linearization of (3) around it. Then a straightforward calculation gives that

$$A(t) = \begin{bmatrix} b(\tilde{x}(t)) + \tilde{x}(t)b'(\tilde{x}(t)) - \alpha\tilde{y}(t) - \gamma & -\alpha\tilde{x}(t) \\ \alpha\tilde{y}(t) & \alpha\tilde{x}(t) - 1 \end{bmatrix}.$$

Let T be a period of this periodic solution. To establish its asymptotic stability it suffices to show that

$$\int_0^T \operatorname{tr} A(t)dt < 0,$$

but in fact

$$\begin{aligned} \int_0^T \operatorname{tr} A(t)dt &= \int_{\tilde{x}(0)}^{\tilde{x}(T)} \frac{dx}{x} + \int_0^T \tilde{x}(t)b'(\tilde{x}(t))\,dt + \int_{\tilde{y}(0)}^{\tilde{y}(T)} \frac{dy}{y} \\ &= 0 + \int_0^T \tilde{x}(t)b'(\tilde{x}(t))\,dt + 0 < 0 \end{aligned}$$

since $\tilde{x}(t)b'(\tilde{x}(t)) < 0$ by (4a).

Because of this we can now show that the system (3) has no positive periodic solution. Indeed, suppose there were a positive periodic orbit Γ. The α-limits of each point enclosed by Γ must be nonempty and also enclosed by Γ. Since both the positive equilibrium (x^*, y^*) and Γ are locally asymptotically stable, neither can be contained in an α-limit of a point enclosed by Γ. But there are no other equilibria, and any positive periodic solution must be locally asymptotically stable, so there can be no α-limits of points other than (x^*, y^*) (the α-limit of itself) enclosed by Γ. This is a contradiction.

Finally, it follows from Lemma 1 that every positive solution is bounded for $t \geq 0$, and therefore must have a nonempty ω-limit. By the preceding argument and Poincaré-Bendixson theory every such limit must contain at least one of the equilibria. However, the two boundary equilibria $(0, 0)$ and $(1, 0)$ cannot be contained in the necessarily bounded ω-limits of any other points. Hence (x^*, y^*) must be in the ω-limit of every interior solution, and since it is locally asymptotically stable it must be the entire limit set. □

3. Invulnerable juvenile prey: A host-parasite situation. We now consider the effect of taking into account a juvenile class of prey, which we assume to be invulnerable to predation from birth until an age large enough to warrant inclusion of the class in our model. Taking the model (1) as our starting point, we assume that the juvenile class, consisting of those prey from ages 0 to R, is subject to a constant mortality rate d_1 and so is given by

$$J(T) = \int_{T-R}^{T} N(s)\, B(N(s)) e^{-d_1(T-s)}\, ds,$$

where $N(t)$ is the population of adults of the prey species, and with elementary calculations we find its derivative is

$$J'(T) = N(T)\, B(N(T)) - N(T-R)\, B(N(T-R)) e^{-d_1 R} - d_1 J(T). \tag{5a}$$

Interpreting the three terms of the last expression, we find that the first is the current rate of juvenile births, while the second is the current rate of maturation of surviving juveniles to adulthood, and the third is current juvenile mortality. These considerations motivate the alteration of our original model to the form

$$N'(T) = N(T-R)B(N(T-R))e^{-d_1 R} - aN(T)P(T) - dN(T), \tag{5b}$$

$$P'(T) = cN(T)P(T) - d_P P(T). \tag{5c}$$

We suppose that (2a) continues to hold, as well as the obvious generalization of (2b),

$$B(0)e^{-d_1 R} > d > B(\infty)e^{-d_1 R}. \tag{6}$$

In this situation we define N_0 by the condition

$$B(N_0)\, e^{-d_1 R} = d.$$

In addition, for $N \geq 0$ we assume that

$$NB'(N) + B(N) \geq 0 \tag{7}$$

or equivalently that the "per capita" recruitment rate, $B(N)$, is kept from decreasing too fast in the sense that $B'(N) \geq -B(N)/N$. In contrast with the ODE case, by this assumption we no longer admit the case of $B(N) = \exp(-pN)$. See Cooke et al. [8], where it is shown in a model of a single population that dynamic behavior can be much more complicated when $B(N)$ has the exponential form. Since we want to concentrate on the effects of maturation delays in prey or predator, we have chosen to work with the simpler form. We also wish to point out that many studies in the literature deal with models in which the term aNP is replaced by $Pf(N)$, where f is called the functional response. We have retained the simpler form aNP in order to isolate the effects of delayed recruitment.

Since (5b, 5c) can be decoupled and solved independently from (5a), we can again restrict our attention to the differential equations (now with delays) for (N, P). We will again find that our analysis is facilitated by a scaling of the variables. If

$$\begin{array}{ccc} x = N/N_0, & t = d_P T, & \alpha = cN_0/d_P, \\ y = aP/cN_0, & r = d_P R, & \gamma = d/d_P, \\ & b(x) = B(xN_0)\, e^{-d_1 R}/d_P, & \end{array} \tag{8}$$

then the system (5) takes the (nondimensional) form

$$\frac{dx}{dt} = x(t-r)\, b(x(t-r)) - \alpha x(t)\, y(t) - \gamma x(t), \tag{9a}$$

$$\frac{dy}{dt} = \alpha x(t)\, y(t) - y(t). \tag{9b}$$

Notice that in this scaling the transfer from x to y by the scaled predation is again perfectly efficient, and the mortality factor, with its physical presence of the delay in $e^{-d_1 R}$, is scaled out. We will show in section 7 that there is no loss in generality for, e.g., bifurcation as R increases, resulting from this scaling. The properties assumed in (2a), (6), and (7) take the simpler form

$$b(x) \geq 0 \quad \text{and} \quad b'(x) < 0 \qquad \text{for } x \geq 0, \tag{10a}$$

$$b(0) > \gamma > b(\infty) \quad \text{and} \quad b(1) = \gamma, \tag{10b}$$

$$b(x) + xb'(x) \geq 0, \tag{10c}$$

and the boundary equilibria are now at $(0,0)$ and $(1,0)$.

LEMMA 5. *The system* (9) *is dissipative, that is, there is a compact set* Ω *(in this case of the form* $\{(x,y) : x, y \geq 0,\ x + y \leq m\}$*) such that each solution in the first quadrant of the* (x,y) *plane has a* T_0 *such that when* $t \geq T_0$, $(x(t), y(t)) \in \Omega$. *Moreover,* $\limsup_{t\to\infty} x(t) \leq 1$.

Proof. Since $b(0) > 0$ it is easy to check that the first quadrant is forwardly invariant under solutions of (9). With $V(x,y) = x + y$ we have

$$\dot{V} \overset{def}{=} \frac{d}{dt} V(x(t), y(t)) = x(t-r)\, b(x(t-r)) - \gamma x(t) - y(t).$$

To set up the analysis of $\dot{V}$, we first establish the boundedness of $x(t)$ for every positive solution.

Let $\bar{x} = \limsup_{t\to\infty} x(t)$. Suppose, contrary to our claim, that $\bar{x} > 1$. Since the case in which $\bar{x}$ is infinite is similar to the finite case, but slightly less complicated, we will consider only the finite case . Then we can find a sequence $\{t_n\}$ with $t_n \to \infty$

as $n \to \infty$ such that $x(t_n) \to \bar{x}, x'(t_n) \to 0$, and $\limsup_{n\to\infty} x(t_n - r) \le \bar{x}$. Then we have

$$\begin{aligned}\lim_{n\to\infty} x'(t_n) &\le \lim_{n\to\infty} \sup\left(x(t_n - r)\, b(x(t_n - r)) - \gamma x(t_n)\right)\\ &= \lim_{n\to\infty} \sup\left(x(t_n - r)\, b(x(t_n - r))\right) - \gamma\bar{x}\\ &\le \left(b(\bar{x}) - \gamma\right)\bar{x} < 0,\end{aligned}$$

contradicting our choice of $\{t_n\}$ so that $x'(t_n) \to 0$. Because of this every $x(t)$ has $\limsup_{t\to\infty} x(t) \le 1$.

Finally we want to show that $V(x(t), y(t))$ is decreasing whenever it has sufficiently large values. To this end, let $\tilde{x} > 1$ and then choose $\tilde{y}$ such that $xb(x) - y < 0$ when $x \le \tilde{x}$ and $y \ge \tilde{y}$. Now when $V(x(t), y(t)) > \tilde{x} + \tilde{y}$ for large t, we must have $x(t), x(t-r) \le \tilde{x}$ (by the previous paragraph) and hence $y(t) > \tilde{y}$ so that $\dot{V} < x(t-r)\, b(x(t-r)) - y(t) < \tilde{x} b(\tilde{x}) - \tilde{y} < 0$. Thus for any solution after sufficiently long time, we have $\dot{V} < 0$ whenever $V(x(t), y(t)) > \tilde{x} + \tilde{y}$. Standard Lyapunov function considerations complete the argument. □

COROLLARY 6. *Suppose* $\alpha < 1$. *Then all positive solutions of* (9) *converge to* $(1, 0)$ *as* $t \to \infty$.

Proof. First we establish that $y(t) \to 0$ as $t \to \infty$. Let $\beta > 0$ such that $\alpha < \beta < 1$, and let $(x(t), y(t))$ be a positive solution of (9). By Lemma 5 there is a sufficiently large t^* such that $\alpha x(t) \le \beta$ for all $t \ge t^*$. This implies that

$$y'(t) \le -(1 - \beta) y(t), \quad t \ge t^*.$$

Therefore for $t \ge t^*$ we have

$$0 \le y(t) \le y(t^*) e^{-(1-\beta)(t-t^*)} \to 0 \quad \text{as } t \to \infty.$$

It follows that $y(t) \to 0$ as $t \to \infty$.

Next we claim that

$$M = \lim_{t\to\infty} \inf x(t) > 0.$$

Suppose to the contrary that $M = 0$. Then since $x(t) > 0$ for all $t \ge 0$, one is able to choose a sequence $\{t_n\}$ having the properties that

$$\begin{aligned} x(t_n - r) &\ge x(t_n),\\ x'(t_n) &\le 0 \end{aligned} \tag{11}$$

for all $n \ge 1$ and

$$\lim_{n\to\infty} x(t_n) = \lim_{n\to\infty} x'(t_n) = 0.$$

Since $\lim_{n\to\infty} x(t_n) = 0$, for which there is a J such that

$$x(t_n) < \frac{1}{2}, \quad n \ge J, \tag{12}$$

and since $b(x)$ is strictly decreasing, there is an $\varepsilon > 0$ such that

$$b(x(t_n)) \ge b(1) + \varepsilon = \gamma + \varepsilon, \quad n \ge J. \tag{13}$$

Moreover since $y(t) \to 0$ as $t \to \infty$, there exists $J_1 \geq J$ such that

$$\alpha y(t_n) < \varepsilon, \quad n \geq J_1. \tag{14}$$

Then for $n \geq J_1$ we have from (11), the increasing of $xb(x)$ and (11), and (12–14) that

$$\begin{aligned} 0 &\geq x'(t_n) \\ &= x(t_n - r)b(x(t_n - r)) - [\alpha y(t_n) + \gamma]\, x(t_n) \\ &\geq x(t_n)\, b(x(t_n)) - [\alpha y(t_n) + \gamma]\, x(t_n) \\ &> x(t_n)(\gamma + \varepsilon) - [\alpha y(t_n) + \gamma]\, x(t_n) \\ &> 0. \end{aligned}$$

This contradiction establishes that $M = \lim_{t\to\infty} \inf x(t) > 0$.

Finally we can show that $M = 1$. Let $t_n \to \infty$ in such a way that $x(t_n) \to M$ as $n \to \infty$ and

$$\lim_{n\to\infty} x'(t_n) = 0. \tag{15}$$

Since $\{x(t_n - r)\}$ is a bounded sequence, without loss of generality (or by choosing a subsequence) we can suppose that $x(t_n - r)$ converges,

$$\lim_{n\to\infty} x(t_n - r) = M_1. \tag{16}$$

By Lemma 5 and the definition of M it is obvious that

$$M \leq M_1 \leq 1.$$

Therefore (15, 16) yield

$$\begin{aligned} 0 &= \lim_{n\to\infty} x'(t_n) \\ &= \lim_{n\to\infty} x(t_n - r)b(x(t_n - r)) - \gamma \lim_{n\to\infty} x(t_n) \\ &= M_1 b(M_1) - \gamma M. \end{aligned}$$

We now show that this equality forces $M = M_1 = 1$. For if $M < 1$, then either $M < M_1$ or $M = M_1 < 1$. If $M < M_1$, then we have

$$M_1 b(M_1) \geq M_1 b(1) = M_1 \gamma > \gamma M,$$

a contradiction of the established equality. On the other hand, if $M = M_1 < 1$, then

$$M_1 b(M_1) > M_1 b(1) = M_1 \gamma = \gamma M,$$

again contradicting the equality.

Hence $M = \liminf_{t\to\infty} x(t) = 1$, which, in conjunction with Lemma 5 gives $x(t) \to 1$. □

Now let us suppose $\alpha > 1$ (i.e., $cN_0 > d_P$). Then the system (9) has a unique positive equilibrium (x^*, y^*). We proceed to study the local stability of this positive equilibrium.

First, following a straightforward calculation one is able to verify that the linearization of (9) at the positive equilibrium takes the form

$$u'(t) = [x^* b'(x^*) + b(x^*)]\, u(t-r) - (\alpha y^* + \gamma) u(t) - \alpha x^* v(t), \tag{17a}$$

$$v'(t) = \alpha y^* u(t). \tag{17b}$$

THEOREM 7. *The interior equilibrium (x^*, y^*) (which exists if and only if $\alpha > 1$) is locally asymptotically stable for all $r \geq 0$.*

Proof. Let $K = [x^* b'(x^*) + b(x^*)]$. The characteristic equation of (17) is given by

$$\begin{aligned}\Delta(\lambda, r) &= \det \begin{bmatrix} Ke^{-\lambda r} - \lambda - (\alpha y^* + \gamma) & -\alpha x^* \\ \alpha y^* & -\lambda \end{bmatrix} \\ &= \lambda^2 - \lambda K e^{-\lambda r} + \lambda(\alpha y^* + \gamma) + \alpha^2 x^* y^*.\end{aligned}$$

Note that $b'(x^*) < 0 \quad \text{and} \quad b(x^*) = \alpha y^* + \gamma$. Hence from the assumption (10c) we have

$$\alpha y^* + \gamma > |K|. \tag{18}$$

We observe that

$$\Delta(\lambda, 0) = \lambda^2 + (\alpha y^* + \gamma - K)\lambda + \alpha^2 x^* y^*.$$

It follows from (18) that all coefficients of $\Delta(\lambda, 0)$ are positive, and hence all zeros of $\Delta(\lambda, 0)$ have the negative real parts. We claim that for each fixed $r > 0$, all zeros of $\Delta(\lambda, r)$ are located in the left half complex plane. If this is not true, then there must be a $r > 0$ and $v \in \Re$ such that $\Delta(iv, r) = 0$, or equivalently,

$$\Delta(iv, r) = -v^2 - ivKe^{-i\nu r} + iv(\alpha y^* + \gamma) + \alpha^2 x^* y^* = 0.$$

That is,

$$-v^2 + \alpha^2 x^* y^* + iv(\alpha y^* + \gamma) = ivKe^{-ivr}.$$

This yields that

$$\left|-v^2 + \alpha^2 x^* y^* + iv(\alpha y^* + \gamma)\right|^2 = |ivKe^{-ivr}|^2.$$

Hence we obtain that

$$(\alpha^2 x^* y^* - v^2)^2 + (\alpha y^* + \gamma)^2 v^2 = K^2 v^2.$$

However, since $(\alpha y^* + \gamma)^2 > K^2$, the above equality can never hold for $v \in \Re$. This contradiction establishes our result. ∎

4. Invulnerable mature prey. Next we consider a case similar to that in the previous section in distinguishing between juvenile and adult prey, but opposite from it in assuming that predation affects juvenile prey but not mature prey. As before, we begin with the idea that the class of juveniles at time T consists of all those prey surviving from birth in the time interval $[T - R, T]$ and write

$$J(T) = \int_{T-R}^{T} N(s)\, B(N(s)) e^{-M}\, ds,$$

where N and B respectively represent the mature prey class and its per capita birth rate and M represents mortality effects. In addition to the "natural" mortality given by a constant per capita rate d_1 and represented in the previous section by $-d_1(T-s)$, we now include mortality by predation that is jointly proportional to the length of time $(T-s)$ such predation can occur and the average size of the predator class (denoted by P as before) over the interval $[s,T]$, so that

$$M = d_1(T-s) + d_2(T-s)\left(\frac{1}{T-s}\int_s^T P(u)\,du\right) = \int_s^T [d_1 + d_2 P(u)]\,du$$

and hence

$$J(T) = \int_{T-R}^{T} N(s)\,B(N(s))e^{-\int_s^T [d_1+d_2P(u)]du}\,ds.$$

Then from elementary calculations we find that

$$\begin{aligned}(19)\qquad \frac{dJ}{dT}(T) &= N(T)\,B(N(T)) - N(T-R)\,B(N(T-R))\,e^{-\int_{T-R}^{T}[d_1+d_2P(u)]du} \\ &\quad - [d_1 + d_2P(T)]\,J(T).\end{aligned}$$

The three terms comprising dJ/dT have nice interpretations: the first as current births, the second as loss to maturation of those who survived from birth (at time $T-R$) to the present, and the third as current loss due to the combination of constant per capita mortality and "mass-action" predation.

Assuming no predation directly on mature prey we arrive at the system of equations for (J,N,P) (with notation similar to that in the previous section)

$$\begin{aligned}\frac{dN}{dT}(T) &= N(T-R)\,B(N(T-R))\,F\left(\int_{T-R}^{T} P(u)\,du\right) - dN(T), \\ \frac{dP}{dT}(T) &= kP(T)\,J(T) \quad - d_P P(T),\end{aligned}$$

wherein we assume about F only the conditions

$$1 \geq F(0) > 0, \quad F(\infty) = 0, \quad F'(Z) < 0, \quad Z > 0$$

as motivated by our considerations in (19) where $F(Z) = e^{-d_1R - d_2Z}$. Notice that if the juvenile prey $J(T)$ were a constant proportional part of the mature prey population then the last two equations would decouple from the first and could be solved independently. Thus, with the goal of facilitating comparison with our two other models, we make the a priori assumption that $J(T) = CN(T)$. Thus, we consider the system

$$\begin{aligned}(20)\qquad \frac{dN}{dT}(T) &= N(T-R)\,B(N(T-R))\,F\left(\int_{T-R}^{T} P(u)\,du\right) - dN(T), \\ \frac{dP}{dT}(T) &= cP(T)\,N(T) \quad - d_P P(T).\end{aligned}$$

In order to achieve a satisfactory scaling of the model we desire an N_0 that functions like a prey carrying capacity in the absence of predators. As before we suppose

that the conditions (2a) and (7) on B continue to hold, as well as the modification of (2b) given by

$$B(0)F(0) > d > B(\infty)F(0). \tag{21}$$

In this situation, N_0 is uniquely determined by the condition

$$B(N_0)F(0) = d,$$

providing an equilibrium at $(N_0, 0)$. Under the scaling

$$\begin{aligned} x &= N/N_0, \quad t = d_P T, \quad \alpha = cN_0/d_P, \\ y &= P/N_0, \quad r = d_P R, \quad \gamma = d/d_P, \\ b(x) &= B(xN_0)/d_P, \qquad f(z) = F\left(\frac{N_0}{d_P} z\right) \end{aligned}$$

we find that (20) becomes the dimensionless system

$$\begin{aligned} \frac{dx}{dt}(t) &= x(t-r)b(x(t-r))f\left(\int_{-r}^{0} y(t+s)ds\right) - \gamma x(t), \\ \frac{dy}{dt}(t) &= \alpha x(t)y(t) - y(t). \end{aligned} \tag{22}$$

Notice that time has again been scaled by the reciprocal of the predator death rate and that $b(\cdot)$ and $f(\cdot)$ satisfy conditions similar to (10) in section 3:

$$b(x) \geq 0 \quad \text{and} \quad b'(x) < 0 \qquad \text{for } x \geq 0, \tag{23}$$

$$b(0)f(0) > \gamma > b(\infty)f(0) \quad \text{and} \quad b(1)f(0) = \gamma, \tag{24}$$

$$b(x) + xb'(x) \geq 0, \tag{25}$$

$$1 \geq f(0) > 0, \quad f(\infty) = 0, \quad \frac{df}{dz} < 0, \quad z > 0. \tag{26}$$

Just as in the systems (3) and (9), this system has no strictly positive equilibrium if $\alpha < 1$ and has a unique positive equilibrium if $\alpha > 1$.

THEOREM 8. *If $\alpha < 1$, then all positive solutions of (22) converge to the equilibrium $(1, 0)$.*

Proof. Following essentially the same argument as in the second paragraph of the proof of Lemma 5, one establishes that $\limsup_{t\to\infty} x(t) \leq 1$. Given a positive solution $(x(t), y(t))$ we can find a T such that for $t \geq T$, $\alpha x(t) - 1$ is negative and bounded away from zero. Since

$$y'(t) \leq (\alpha x(t) - 1)\, y(t),$$

we find that $y(t) \to 0$ as claimed. The rest of the argument is similar to that given for proof of Corollary 6. ∎

For the rest of this section we consider the remaining case in which $\alpha > 1$. In this case an equilibrium (x^*, y^*) must satisfy $x^* = 1/\alpha < 1$ so that

$$b(x^*)f(z^*) - \gamma = 0$$

has a unique solution $z^* = ry^* > 0$. Thus for each fixed $r > 0$, (22) has the unique positive equilibrium $(x^*, y^*) = (1/\alpha, z^*/r)$. We will show that an increase of the time

delay r will destabilize the stability of this positive equilibrium and cause a Hopf bifurcation.

First a direct computation yields that the linearization of (22) around the positive equilibrium is given by

$$\begin{aligned} u'(t) &= au(t-r) - \gamma u(t) - b\int_{-r}^{0} v(t+s)ds, \\ v'(t) &= \alpha y^* u(t), \end{aligned}$$

where

$$a = [b(x^*) + x^* b'(x^*)]\, f(z^*), \tag{27a}$$

$$b = -x^* b(x^*) \frac{df}{dz}(z^*) > 0. \tag{27b}$$

Therefore the characteristic equation is given by

$$\begin{aligned} \Delta(\lambda, r) &= \det \begin{bmatrix} ae^{-\lambda r} - \gamma - \lambda & -b\int_{-r}^{0} e^{\lambda s} ds \\ \alpha y^* & -\lambda \end{bmatrix} \\ &= \lambda^2 + \gamma\lambda - a\lambda e^{-\lambda r} + \frac{\beta(1 - e^{-\lambda r})}{r\lambda} \qquad (28) \\ &= 0, \end{aligned}$$

where $\beta = \alpha b z^*$.

To study the location of eigenvalues of the characteristic equation (28), an important first step is to investigate the existence of eigenvalues that lie on the imaginary axis of the complex plane and the direction in which these eigenvalues cross that axis as the delay increases. We note that if $\Delta(iv, r) = 0$, then $\Delta(-iv, r) = 0$ so that we shall only search for eigenvalues iv with $v > 0$. Letting

$$h(\lambda, r) = r\lambda^3 + \gamma r\lambda^2 - ar\lambda^2 e^{-\lambda r} + \beta(1 - e^{-\lambda r}), \tag{29}$$

it is clear that $\Delta(\lambda, r) = 0$ if and only if $h(\lambda, r) = 0$ for $\lambda \neq 0$ and $r > 0$.

LEMMA 9. *There are infinitely many positive pairs of (iv, r) with $r > 0$ and $v = v(r) > 0$ such that $\Delta(iv, r) = 0$. However, there is an interval $0 < r < r_1$ such that $\Delta(\lambda, r)$ has no purely imaginary zeros.*

Proof. Note that with $r > 0$, $\gamma = b(x^*)f(z^*)$, the assumptions on b and f imply that

$$\gamma > |a|, \quad \beta > 0.$$

Notice that in $\Delta(\lambda, r)$, neither a nor β depends on r. Thus

$$\Delta(\lambda, 0) = \lambda^2 + (\gamma - a)\lambda + \beta \tag{30}$$

and the interval $(0, r_1)$ on which there are no purely imaginary characteristic zeros follows immediately from continuity.

Considering $\lambda = iv$, if

$$h(iv, r) = 0 \tag{31}$$

for some $v > 0$ and $r > 0$, then we have

$$\frac{-irv^3 + \beta - \gamma r v^2}{(\beta - arv^2)} = e^{-ivr} \tag{32}$$

so that

$$| - irv^3 + \beta - \gamma r v^2|^2 = |\beta - arv^2|^2, \tag{33}$$

or

$$r^2 v^6 + (\beta - \gamma r v^2)^2 = (\beta - arv^2)^2, \tag{34}$$

and finally

$$rv^4 + r(\gamma^2 - a^2)v^2 = 2\beta(\gamma - a). \tag{35}$$

We can solve (35) uniquely for $v^2 = v^2(r)$,

$$v^2 = \frac{1}{2}\left(-(\gamma^2 - a^2) + \sqrt{(\gamma^2 - a^2)^2 + \frac{8\beta(\gamma - a)}{r}}\right) . \tag{36}$$

Thus we have shown that if $h(iv, r) = 0$, then $v = v\,(r)$ with $v > 0$ must satisfy (36). However, it is really the converse question that we must answer: if $v\,(r)$ is the positive branch of (36), do we have $h\,(iv\,(r)\,, r) = 0$? However, for $v > 0$, (36) is equivalent to (33), so that if we let $W\,(r)$ denote the left-hand side of (32) with $v = v(r)$, then $|W\,(r)| = 1$ for all $r > 0$.

Multiplying (36) by r^2 we have

$$(rv)^2 = \frac{1}{2}\left(-r^2(\gamma^2 - a^2) + \sqrt{r^4(\gamma^2 - a^2)^2 + 8r^3\beta(\gamma - a)}\right). \tag{37}$$

From this, one immediately sees that

$$\lim_{r\to 0}[rv(r)]^2 = 0. \tag{38}$$

Moreover, rationalizing the numerator in (37) we have

$$(rv)^2 = \frac{8r^3\beta(\gamma - a)}{2\left(r^2(\gamma^2 - a^2) + \sqrt{r^4(\gamma^2 - a^2)^2 + 8r^3\beta(\gamma - a)}\right)}$$

which yields

$$\lim_{r\to\infty}[rv(r)]^2 = \infty. \tag{39}$$

It follows from (38) and (39) that

$$\lim_{r\to 0} rv(r) = 0 \qquad \text{and} \qquad \lim_{r\to\infty} rv(r) = \infty. \tag{40}$$

Next from (34) we have that $\beta - rav^2(r) > 0$ for all $r > 0$. For if this inequality does not hold for some $r > 0$, then we have $|\beta - arv^2(r)| = arv^2(r) - \beta$. Since $\gamma > a$ we have $\gamma r v^2(r) - \beta > arv^2(r) - \beta$. It would therefore follow that

$$r^2 v^6(r) + (\beta - \gamma r v^2(r))^2 > (\beta - arv^2(r))^2,$$

contradicting (34).

Finally, substitute $v = v(r)$ into (32). Letting r increase from 0 to ∞ and using (40), we find that e^{-irv} on the right side traces out the unit circle infinitely often, while on the left side $W(r)$ remains in the lower half of the unit circle. Now one easily sees that there are infinitely many positive r's such that $h(iv(r), r) = 0$. □

LEMMA 10. *Let $\lambda(r)$ be a branch of zeros of $h(\lambda, r)$ defined on an interval I such that for some $0 < r_0 \in I$, $\lambda(r_0) = iv_0$ with $v_0 > 0$. Then v_0 is a simple zero and*

$$\mathrm{Re}\left(\frac{d\lambda(r_0)}{dr}\right) > 0.$$

Proof. Since $h(\lambda(r), r) \equiv 0$, we have

$$\frac{\partial h(\lambda, r)}{\partial \lambda}\frac{d\lambda}{dr} = -\frac{\partial h(\lambda, r)}{\partial r}.$$

Hence

$$\overline{\left[\frac{\partial h(\lambda, r)}{\partial \lambda}\right]}\frac{\partial h(\lambda, r)}{\partial \lambda}\frac{d\lambda}{dr} = -\overline{\left[\frac{\partial h(\lambda, r)}{\partial \lambda}\right]}\frac{\partial h(\lambda, r)}{\partial r},$$

where $\overline{z}$ denotes the conjugate of a complex number z. It follows from this equality that

$$\mathrm{sign}\left(\mathrm{Re}\frac{d\lambda(r_0)}{dr}\right) = \mathrm{sign}\left(\mathrm{Re}\left\{-\overline{\left[\frac{\partial h(iv_0, r_0)}{\partial \lambda}\right]}\frac{\partial h(iv_0, r_0)}{\partial r}\right\}\right).$$

For notational simplicity from now on we use v and r instead of v_0 and r_0. First from (29) we deduce that

$$iv^3 + \gamma v^2 - av^2 e^{-ivr} = \frac{\beta}{r}(1 - e^{-ivr}), \tag{41}$$

or equivalently, after complex conjugation

$$v^2 + i\gamma v - iave^{ivr} = \frac{i\beta}{rv}(1 - e^{ivr}). \tag{42}$$

Following a straightforward computation and with the use of (41) and (42) we have

$$\begin{aligned}\overline{\left[\frac{\partial h(iv, r)}{\partial \lambda}\right]} &= r\left[-3v^2 - i2\gamma v + i2ave^{ivr} + (\beta - arv^2)e^{ivr}\right] \\ &= r\left[-3(v^2 + i\gamma v - iave^{ivr}) + i(\gamma - a)v + iav(1 - e^{ivr})\right] \\ &\quad + r(\beta - arv^2)e^{ivr} \\ &= r\left[i(\gamma - a)v + i(av - \frac{3\beta}{rv})(1 - e^{ivr}) + (\beta - arv^2)e^{ivr}\right],\end{aligned} \tag{43}$$

$$\begin{aligned}-\frac{\partial h(iv, r)}{\partial r} &= iv^3 + \gamma v^2 - av^2 e^{-ivr} - iv(\beta - arv^2)e^{-ivr} \\ &= \frac{\beta}{r}(1 - e^{-ivr}) - iv(\beta - arv^2)e^{-ivr}.\end{aligned} \tag{44}$$

Let

$$\Phi = -\overline{\left[\frac{\partial h(iv,r)}{\partial \lambda}\right]}\frac{\partial h(iv,r)}{\partial r}.$$

It is clear that if $\mathrm{Re}\Phi > 0$, then $\frac{\partial h(iv,r)}{\partial \lambda} \neq 0$ which will imply that $iv = iv_0$ is a simple eigenvalue and $\mathrm{Re}\frac{d\lambda(r_0)}{dr} > 0$. Hence to complete the proof of Lemma 10, it is sufficient to show that $\mathrm{Re}\Phi > 0$. First by (43) and (44) we have

$$\begin{aligned}\mathrm{Re}\Phi &= \mathrm{Re}\left\{\beta(1-e^{-ivr})\left[i(\gamma-a)v + (\beta - arv^2)e^{ivr}\right]\right\}\\ &\quad + \mathrm{Re}\left\{-ivr(\beta-arv^2)e^{-ivr}\left[-3v^2 - i2\gamma v + i2ave^{ivr}\right]\right\}\\ &= \mathrm{Re}\Phi_1 + \mathrm{Re}\Phi_2\end{aligned}$$

with

$$\begin{aligned}\Phi_1 &= \beta(1-e^{-ivr})\left[i(\gamma-a)v + (\beta-arv^2)e^{ivr}\right],\\ \Phi_2 &= -ivr(\beta-arv^2)e^{-ivr}\left[-3v^2 - i2\gamma v + i2ave^{ivr}\right].\end{aligned}$$

Following a further calculation we have

$$\begin{aligned}\mathrm{Re}\Phi_1 &= -\beta(\gamma-a)v\sin(vr) - \beta(\beta - arv^2)[1-\cos(vr)],\\ \mathrm{Re}\Phi_2 &= (\beta - arv^2)\left[3rv^3\sin(vr) - 2\gamma rv^2\cos(vr) + 2arv^2\right].\end{aligned} \tag{45}$$

Next, by separating the real and imaginary parts of (32) we obtain

$$\begin{aligned}(\beta - arv^2)\cos(vr) &= \beta - \gamma rv^2,\\ (\beta - arv^2)\sin(vr) &= rv^3.\end{aligned} \tag{46}$$

With the use of (35), (45), and (46) we finally arrive at

$$\begin{aligned}\mathrm{Re}\Phi &= -\beta(\gamma-a)v\sin(vr) - \beta(\beta-arv^2) + \beta(\beta-\gamma rv^2)\\ &\quad + 3r^2v^6 - 2\gamma rv^2(\beta-\gamma rv^2) + 2arv^2(\beta-arv^2)\\ &= -\beta(\gamma-a)v\sin(vr) - 3r\beta v^2(\gamma-a) + 3r^2v^6 + 2r^2v^4(\gamma^2-a^2)\\ &= -\beta(\gamma-a)v\sin(vr) + r^2v^6 + r\beta v^2(\gamma-a)\\ &\quad + 2r^2v^6 + 2r^2v^4(\gamma^2-a^2) - 4rv^2\beta(\gamma-a)\\ &= -\beta(\gamma-a)v\sin(vr) + r^2v^6 + r\beta v^2(\gamma-a)\\ &> \beta(\gamma-a)v\left[rv - \sin(vr)\right]\\ &\geq 0. \qquad \square\end{aligned}$$

Now we are ready to prove the following result.

THEOREM 11. *Suppose that $\alpha > 1$. Then there is a sequence $\{r_n\}_{n=1}^{\infty}$ with*

$$0 < r_1 < r_2 < \cdots < r_n < \cdots$$

such that the following hold:

1. *If $0 < r < r_1$, then the positive equilibrium $(x^*, y^*) = (1/\alpha, z^*/r)$ of (22) is linearly stable. If $r > r_1$ then that positive equilibrium is linearly unstable.*
2. *For each $n \in \mathbf{N}$ there is a $v_n > 0$ such that $\lambda = \pm iv_n$ are eigenvalues associated with the equilibrium (x^*, y^*) of (22) at $r = r_n$ and there is a Hopf bifurcation there.*

Proof. Using Lemmas 9 and 10 with the application of classical results on stability switches (for example, see [7]) and on Hopf bifurcation (e.g., [10], p. 291), we need only to show that all solutions of the characteristic equation

$$\Delta(\lambda, r) = \lambda^2 + \gamma\lambda - a\lambda e^{-\lambda r} + \frac{\beta(1 - e^{-\lambda r})}{r\lambda} = 0$$

have negative real parts for all $r \in [0, r_1)$, where r_1 is taken to be the smallest positive delay at which the characteristic function has a purely imaginary zero. The existence of r_1 is guaranteed by Lemma 9.

Now for a characteristic root $\lambda = \lambda(r) = u + iv$ with $r \geq 0$ and $u \geq 0$ we have

$$\begin{aligned} \left|1 - e^{-\lambda r}\right| &\leq \left|1 - e^{-ur} + e^{-ur}\left(1 - \cos(vr) + i\sin(vr)\right)\right| \\ &\leq \left|1 - e^{-ur}\right| + \left|1 - \cos(vr)\right| + \left|\sin(vr)\right| \\ &\leq |ur| + 2|vr| \leq 3|\lambda r|. \end{aligned}$$

From this we can conclude that

$$\left|\frac{\beta\left(1 - e^{-\lambda r}\right)}{r\lambda}\right| \leq 3\beta.$$

It now follows from this that there is a sufficiently large constant C such that for all $\lambda \in \mathbf{C}$ with $\mathrm{Re}(\lambda) \geq 0$ and $|\lambda| \geq C$,

$$\lambda^2 + \gamma\lambda - a\lambda e^{-\lambda r} + \frac{\beta(1 - e^{-\lambda r})}{r\lambda} \neq 0.$$

Since $\Delta(\lambda, r)$, with $\Delta(\lambda, 0)$ as in (30), is analytic in $\lambda \neq 0$ and continuous in $r \geq 0$, it follows that its number of zeros on $\mathrm{Re}\lambda \geq 0$ is constant for $r \in [0, r_1)$. Since $\Delta(\lambda, 0)$ has only zeroes with negative real part, our result holds. ∎

5. Inactive juvenile predators. We now consider the effect of taking into account an inactive juvenile class of predator. Thus all predation is done by the adult predators which we still denote by P, but we change J to denote the juvenile predators and keep N to denote the prey. Taking the model (1) again as our starting point, we assume that the juvenile class (consisting of those predators from ages 0 to R) is the direct beneficiary of predation and is subject to a constant mortality rate d_1 and so is given by

$$J(T) = \int_{T-R}^{T} cN(s)\,P(s)\,e^{-d_1(T-s)}\,ds$$

with derivative

$$J'(T) = cN(T)\,P(T) - cN(T-R)\,P(T-R)\,e^{-d_1 R} - d_1 J(T). \tag{47}$$

Interpreting the three terms of the last expression, we find that the first is the current rate of juvenile births, while the second is the current rate of maturation of surviving juveniles to adulthood, and the third is current juvenile mortality. Thus

$$N'(T) = N(T)B(N(T)) - aN(T)P(T) - dN(T) \tag{48a}$$

$$P'(T) = cN(T-R)\,P(T-R)\,e^{-d_1 R} - d_P P(T). \tag{48b}$$

We suppose that the conditions (2a), (2b), and (7) on B continue to hold. As in section 2 we define N_0 by the condition

$$B(N_0) = d.$$

Since (48a) and (48b) can be decoupled and solved independently from (47), we can again restrict our attention to the differential equations (with delays) for (N, P).

We will again find that our analysis is facilitated by a scaling of the variables. If

$$\begin{array}{ccc} x = N/N_0, & t = d_P T, & \alpha = ce^{-d_1 R} N_0/d_P, \\ y = ae^{d_1 R} P/cN_0, & r = d_P R, & \gamma = d/d_P, \\ & b(x) = B(xN_0)/d_P, & \end{array} \tag{49}$$

then the system (48) takes the (nondimensional) form (for scaled prey and adult predators, respectively x and y)

$$\frac{dx}{dt} = x(t)\, b(x(t)) - \alpha x(t)\, y(t) - \gamma x(t) \tag{50a}$$

$$\frac{dy}{dt} = \alpha x(t-r)\, y(t-r) - y(t). \tag{50b}$$

Notice that in this scaling the transfer from x to y by the scaled predation is again perfectly efficient, and the mortality factor, $e^{-d_1 R}$, is scaled out. The properties of b take the same form as in (10). The boundary equilibria are again at $(0,0)$ and $(1,0)$. A somewhat simplified version of the proof in the second paragraph of the proof for Lemma 7 gives us

LEMMA 12. *For every positive solution* $(x(t), y(t))$, $\limsup_{t\to\infty} x(t) \le 1$.

COROLLARY 13. *Suppose* $\alpha < 1$. *Then all positive solutions of* (50) *converge to* $(1,0)$ *as* $t \to \infty$.

Proof. Let $(x(t), y(t))$ be a positive solution. First we show that $y(t) \to 0$. If β is chosen such that $\alpha < \beta < 1$ then by the above lemma there is some t^* such that $\alpha x(t-r) \le \beta$ for all $t \ge t^*$ and so

$$y'(t) = \alpha x(t-r)\, y(t-r) - y(t) \le \beta y(t-r) - y(t). \tag{51}$$

Since $0 < \beta < 1$, all solutions of

$$w'(t) = \beta w(t-r) - w(t) \tag{52}$$

tend to zero as $t \to \infty$ ([5], section 2). Furthermore all solutions of (52) with positive initial data remain positive for all positive t. From these considerations we can conclude ([20]) that if a solution $w(t)$ of (52) shares positive initial data with $y(t)$ then $0 < y(t) \le w(t)$ for all $t \ge 0$. So

$$y(t) \to 0 \tag{53}$$

as claimed.

Setting $M = \liminf_{t\to\infty} x(t) > 0$, one shows that $M > 0$ in exactly the same way as in Corollary 6. Now we show $M = 1$. (Recall $b(1) = \gamma$.) Suppose not. Then by the previous lemma, $M \le 1$, so $0 < M < 1$. Let M_1 be such that $M < M_1 < 1$. Then $b(M_1) - \gamma > b(1) - \gamma = 0$. By (53), for all t larger than some T_1,

we have $0 < \alpha y(t) < (b(M_1) - \gamma)/2$ and $x(t) \geq M/2$. Hence whenever $t > T_1$ and $\frac{1}{2}M \leq x(t) \leq M_1$, we have

$$\begin{aligned} x'(t) &= x(t)(b(x(t)) - \alpha y(t) - \gamma) \\ &\geq \frac{M}{2}\left(b(M_1) - \frac{(b(M_1) - \gamma)}{2} - \gamma\right) \\ &= \frac{M}{2}(b(M_1) - \gamma)/2 \end{aligned}$$

providing a uniform positive lower bound on $x'(t)$ whenever t is large and $\frac{1}{2}M \leq x(t) \leq M_1$ if $M < 1$, another contradiction. Hence $M = 1$ and

$$\liminf_{t\to\infty} x(t) = \limsup_{t\to\infty} x(t) = \lim_{t\to\infty} x(t) = 1,$$

as desired. □

Finally we consider the case $\alpha > 1$. (For an interpretation of α, see section 6.) In this case $b(1/\alpha) > b(1) = \gamma$ and there is a positive equilibrium (x^*, y^*),

$$\begin{aligned} x^* &= 1/\alpha, \\ y^* &= (b(1/\alpha) - \gamma)/\alpha. \end{aligned}$$

To simplify notation somewhat, we let $b_* = b(x^*)$ and $b'_* = b'(x^*)$. Note that $b_* > \gamma$ and $b'_* < 0$ and that both b_* and b'_* depend on α.

THEOREM 14. *Consider* (50) *in the case* $\alpha > 1$, *that is, in the case that there is an interior equilibrium* (x^*, y^*).

1. *If* $\alpha(b_* - \gamma) + 2b'_* < 0$ *the equilibrium* (x^*, y^*) *is linearly stable for all delays* $r \geq 0$;
2. *If* $\alpha(b_* - \gamma) + 2b'_* > 0$ *then there is a critical delay* $r_1 > 0$ *such that the equilibrium* (x^*, y^*) *is linearly stable for all delays* $0 \leq r < r_1$ *and is linearly unstable for all delays* $r > r_1$. *A Hopf bifurcation occurs as* r *increases through* r_1.

Proof. The linearization of (50) about (x^*, y^*) is

$$\begin{aligned} u'(t) &= x^* b'_* u(t) - \alpha x^* v(t) \\ &= b'_* u(t)/\alpha - v(t) \\ v'(t) &= \alpha y^* u(t - r) + \alpha x^* v(t - r) - v(t) \\ &= (b_* - \gamma) u(t - r) + v(t - r) - v(t), \end{aligned}$$

yielding the linear equation for solutions of the form $(u(t), v(t)) = e^{zt}(u_0, v_0)$,

$$ze^{zt}\begin{bmatrix} u_0 \\ v_0 \end{bmatrix} = e^{zt}\begin{bmatrix} b'_*/\alpha & -1 \\ (b_* - \gamma)e^{-zr} & (e^{-zr} - 1) \end{bmatrix}\begin{bmatrix} u_0 \\ v_0 \end{bmatrix}$$

and the characteristic equation

$$\begin{aligned} 0 &= \det\begin{bmatrix} b'_*/\alpha - z & -1 \\ (b_* - \gamma)e^{-zr} & e^{-zr} - 1 - z \end{bmatrix} \\ &= \frac{1}{\alpha}\left(-b'_* - (b'_* - \alpha)z + \alpha z^2 + e^{-zr}(b'_* + \alpha b_* - \alpha\gamma - \alpha z)\right) \end{aligned}$$

or

$$0 = \alpha z^2 - (b'_* - \alpha)z - b'_* + e^{-zr}(-\alpha z + \alpha b_* + b'_* - \alpha\gamma). \tag{54}$$

We apply the results of Cooke and van den Driessche [7] to analyze the roots of the characteristic equation. In their notation, we set

$$P(z) = \alpha z^2 - (b'_* - \alpha) z - b'_*, \tag{55}$$
$$Q(z) = -\alpha z + \alpha (b_* - \gamma) + b'_*.$$

According to [7], $z = iy$ is a purely imaginary characteristic root only if $F(y) = 0$ where $F(y) = |P(iy)|^2 - |Q(iy)|^2$, and furthermore:

1. If $F(y) = 0$ has no positive roots and the equilibrium is stable when $r = 0$, then it is stable for all $r > 0$.
2. If $F(y) = 0$ has a positive root and each positive root is simple, then there is an r_1 such that the equilibrium is unstable for $r > r_1$ and the direction of a characteristic root crossing the imaginary axis is given by the sign of $F'(y)$.

To apply these results, we find

$$\begin{aligned} F(y) &= \left|-\alpha y^2 - (b'_* - \alpha) iy - b'_*\right|^2 - \left|-\alpha iy + \alpha (b_* - \gamma) + b'_*\right|^2 \\ &= \alpha^2 y^4 + b'^2_* y^2 - \alpha^2 b^2_* + 2\alpha^2 b_* \gamma - 2\alpha b_* b'_* - \alpha^2 \gamma^2 + 2\alpha b'_* \gamma. \end{aligned}$$

Let's examine the constant term of $F(y)$,

$$\alpha \left(-\alpha b^2_* + 2\alpha b_* \gamma - 2 b_* b'_* + 2 b'_* \gamma - \alpha \gamma^2\right) = -\alpha (b_* - \gamma) (2 b'_* + \alpha (b_* - \gamma)).$$

As mentioned just before this theorem, the assumption of $\alpha > 1$ gives us $b_* > \gamma$. Since $b'_* < 0$ the sign of the factor $2b'_* + \alpha (b_* - \gamma)$ can vary depending on the implementation of $b(x)$.

Clearly if $2b'_* + \alpha (b_* - \gamma) < 0$, then F has no real zeros and so the characteristic equation has no purely imaginary roots. According to [7], since the equilibrium is stable under zero delay (Theorem 4), it is stable under all nonnegative delays.

However, if $2b'_* + \alpha (b_* - \gamma) > 0$, then F will have exactly one positive zero, corresponding to a characteristic root with positive imaginary part. In this case the direction of a characteristic root crossing the imaginary axis is given by the sign of $F'(y)$ [7] which is positive when y is positive. So any crossing is transverse, from left to right. The second conclusion of the theorem now follows exactly from [7] and standard bifurcation theory (e.g. [14], p. 332). □

Since Theorem 14 concerns the dimensionless system (50) obtained by scaling out the physical presence of the delay from (48), we should explain how this theorem can be applied to the dimensional system (48). First note that $\alpha > 1$ is equivalent to $d_P e^{d_1 R}/c < N_0$, so that the latter is equivalent to the existence of a positive equilibrium. Upon a substitution of the original parameters into the inequality of part 1 in Theorem 14, we conclude that the positive equilibrium of system (48) exists and is linearly stable if $d_P e^{d_1 R}/c < N_0$ and $ce^{-d_1 R} N_0 \left(B(d_P e^{d_1 R}/c) - d\right) + 2B'(d_P e^{d_1 R}/c) N_0 d_P < 0$.

However, for the bifurcation at $F = 0$ one must proceed cautiously when translating the results from the nondimensional case of Theorem 14 to the dimensional case which motivates it. We naturally inquire if a bifurcation, such as given by Theorem 14 in the nondimensional case with r increasing, is mirrored by one in the corresponding original system (48) with R as the bifurcation parameter and all other parameters (except r) held fixed. However, such variation of parameters is inherently contradictory! Since

$$\alpha = ce^{-d_1 R} N_0 / d_P, \tag{56}$$

we see that if we seek bifurcation in the original system (48) with only R varying, then we must examine Theorem 14 as α varies, instead of relying on it to be constant, with $\alpha > 1$, as assumed in the hypotheses of the theorem. In particular, *if* we increase R while holding fixed all the other parameters in (48), then $\alpha = ce^{-d_1 R}N_0/d_P$ must decrease, eventually violating the theorem's hypothesis of $\alpha > 1$.

Fortunately, a bifurcation diagram for system (48) can be calculated in the $(R\,, c)$ parameter plane (while all parameters other than R, c are fixed). To proceed with this for each given $\alpha > 0$, let the curve Γ_α be the level curve of (56) in the (R, c) plane, i.e., the graph of the function

$$c = \alpha d_P e^{d_1 R}/N_0, \qquad (R \geq 0)$$

with (d_1, d_P, N_0) fixed. There is a positive equilibrium corresponding to some (R, c) on Γ_α if and only if $\alpha > 1$. For such α, let $r_1 = r_1(\alpha)$ be the unique positive value defined in part 2 of Theorem 14 and let $R_1 = R_1(\alpha) = r_1(\alpha)/d_p$. Then from Theorem 14 it immediately follows that such an equilibrium is linearly stable (respectively, unstable) if $R < R_1$ (respectively, $R > R_1$). Moreover, a Hopf bifurcation occurs as the point (R, c) passes through (R_1, c_1) along Γ_α, where $c_1 = c_1(\alpha) = \alpha d_P e^{d_1 R_1(\alpha)}/N_0$.

Let $K\,(\alpha)$ denote the critical quantity distinguishing the cases of Theorem 14:

$$K(\alpha) = \alpha(b(1/\alpha) - \gamma) + 2b\prime(1/\alpha) \qquad (\alpha > 1).$$

It is clear that

$$K(\alpha) \to +\infty \quad \text{as} \quad \alpha \to \infty.$$

Let us first consider the simple case in which there is a unique $\alpha^* > 1$ such that $K(\alpha^*) = 0$ and $K(\alpha) > 0$ for $\alpha > \alpha^*$. (In section 6 α^* exists and is given for birth function b_i by $\alpha^* = \alpha_i, i = 1, 2, 3.$) Then for each $\alpha > \alpha^*$, there is a unique bifurcation point $(R_1(\alpha), c_1(\alpha))$ in the curve Γ_α. It is obvious that the bifurcation point $(R_1(\alpha), c_1(\alpha))$ is continous with respect to α (but formulas are supplied below). Thus by varying α from α^* to ∞ we obtain a simple bifurcation curve in the (R, c) parameter plane that does not intersect itself. In what follows we shall show that the function $R_1(\alpha)$ has a nice property that

$$\lim_{\alpha \to \alpha^*} R_1(\alpha) = +\infty, \qquad \lim_{\alpha \to \infty} R_1(\alpha) = 0.$$

One readily finds closed form expression for the bifurcation delay r_1 in the nondimensional equations (50) as it depends on α. When $K\,(\alpha) > 0$ the function F in the proof of Theorem 14 has a unique positive zero y_0 and r_1 is the least positive solution of $P + e^{-iy_0 r}Q = 0$. Thus, remembering that P, Q, F, y_0 depend on α, r_1 is a composition of two functions:

$$y_0\,(\alpha) = \sqrt{\frac{1}{2A}\left(-B + \sqrt{B^2 - 4AC}\right)}, \tag{57}$$

where

(58)

$$A = \alpha^2, \qquad B = b'\,(1/\alpha)^2\,, \qquad C = -\alpha\,(b\,(1/\alpha) - \gamma)\,(2b'\,(1/\alpha) + \alpha\,(b\,(1/\alpha) - \gamma))\,,$$

followed by

$$r_1\,(\alpha) = \rho\,(\alpha) = \frac{1}{y_0\,(\alpha)} \arg\left(\frac{-Q\,(iy_0\,(\alpha)\,, \alpha)}{P\,(iy_0\,(\alpha)\,, \alpha)}\right), \tag{59}$$

where arg denotes the argument of a complex number. Using the expressions of P, Q one is able to verify that for all sufficiently large α,

$$0 < \arg\left(\frac{-Q(iy_0(\alpha),\alpha)}{P(iy_0(\alpha),\alpha)}\right) < \pi. \tag{60}$$

Furthermore, (57) and (58) yield that

$$\lim_{\alpha\to\infty} y_0(\alpha) = \sqrt{b(0)-\gamma}, \qquad \lim_{\alpha\to\infty} \frac{-Q(iy_0(\alpha),\alpha)}{P(iy_0(\alpha),\alpha)} = 1 \tag{61}$$

and it follows from (59) and (61) that

$$\lim_{\alpha\to\infty} R_1(\alpha) = \frac{1}{d_P}\lim r_1(\alpha) = \frac{1}{d_P\sqrt{b(0)-\gamma}}\arg(1) = 0. \tag{62}$$

Finally we easily see that

$$\lim_{\alpha\to\alpha^*} y_0(\alpha) = 0, \qquad \lim_{\alpha\to\alpha^*} \frac{-Q(iy_0(\alpha),\alpha)}{P(iy_0(\alpha).\alpha)} = -1 \tag{63}$$

which implies that

$$R_1(\alpha) \to +\infty \quad \text{as} \quad \alpha \searrow \alpha^*. \tag{64}$$

With the use of (62) and (64) we can describe the bifurcation diagram of the dimensional system (48) in the (R, c) plane (see Figure 1). The positive quadrant of the (R, c) plane is divided into three parts by the curves $\Gamma_1 : c = d_P e^{d_1 R}/N_0$ and $\Gamma = \{(R_1(\alpha), c_1(\alpha)) : \alpha \in (\alpha^*, \infty)\}$. The curve Γ is asymptotic to the graph $\Gamma_{\alpha^*} : c = \alpha^* d_P e^{d_1 R}/N_0$ as $\alpha \searrow \alpha^*$ and is asymptotic to the vertical line $R = 0$ as $\alpha \to \infty$. Region I corresponds to system (48) having no positive equilibrium; region II corresponds to system (48) having a stable positive equilibrium; and region III corresponds to system (48) with an unstable positive equilibrium. Γ is a hopf bifurcation curve.

We finally remark that if there is any pair of two numbers $1 < \alpha_1 < \alpha_2$ such that $K(\alpha_1) = K(\alpha_2) = 0$, with $K(\alpha) > 0, \alpha \in (\alpha_1, \alpha_2)$ then the curve $\{(R_1(\alpha), c_1(\alpha)) : \alpha \in (\alpha_1, \alpha_2)\}$ gives rise to a Hopf bifurcation curve that is asymptotic to the graph Γ_{α_i} as $\alpha \to \alpha_i$ for $i = 1, 2$.

The parameterization of Γ, $(R_1(\alpha), c_1(\alpha))$ can be found with $R_1(\alpha)$ explicitly and $c_1(\alpha)$ explicitly, using (59) and $c = \alpha d_P e^{d_1 R}/N_0$. Since

$$R_1 = r_1/d_P = \rho(\alpha)/d_P = \rho\left(ce^{-d_1 R_1} N_0/d_P\right)/d_P,$$

we have

$$R_1(\alpha) = r_1(\alpha)/d_P = \rho(\alpha)/d_P, \tag{65}$$

$$c_1(\alpha) = d_p e^{d_1 R_1(\alpha)} \sigma(d_P R_1(\alpha))/N_0 \tag{66}$$

if σ inverts ρ. When one has ρ numerically, σ just reverses the coordinates, so this is easy to plot.

We consider two examples. Since the physical presence of the delay is a key feature in (48), we simplify everything else as much as possible (but with some minimal care to respect biological interpretation of the parameters: see below), taking

$$a = d_1 = d_P = N_0 = 1,$$

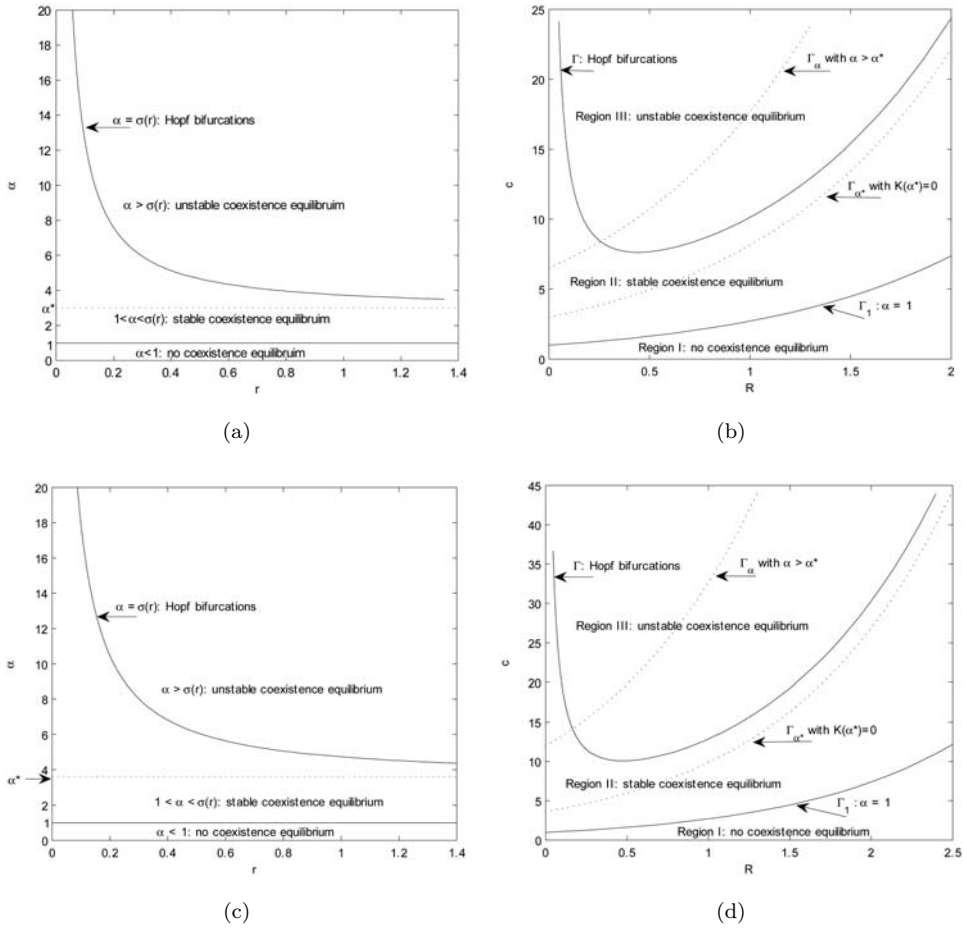


FIG. 1. (a) *Scaled system,* $b(x) = 8(1 - x/4)$; (b) *Dimensional system,* $B(X) = 8(1 - X/4)$; (c) *Scaled system,* $b(x) = 9/(1 + 0.5x)$; (d) *Dimensional system,* $B(X) = 9/(1 + 0.5X)$.

so that in the corresponding dimensionless system

$$\alpha = ce^{-R}, \quad t = T, \quad x(t) = N(t), \quad y(t) = e^{R}P(t)/c, \quad b(x) = B(x).$$

In the case of $B(X) = 8(1 - X/4)$ giving $\gamma = b(1) = 6$, we find bifurcation diagrams as in (a) and (b) of Figure 1, and then, when $B(X) = 9/(1 + 0.5X)$, again with $\gamma = b(1) = 6$, we find bifurcation diagrams as in (c) and (d) of Figure 1.

Notice in both examples that for appropriately large, fixed c with R increasing from 0 the dimensional system (48) goes first through a supercritical Hopf bifurcation, and then back through a subcritical one.

6. Numerical examples.

6.1. Interpretation of parameters. In choosing parameter values for numerical examples, it is helpful to think about the ideas they represent. Thus we consider:

1. t which is time scaled by $1/d_P$, which we may take as a measure of the expected lifespan for predators.
2. α, which has been a focus of attention, beginning with the condition $\alpha > 1$ as our condition for the existence of an interior equilibrium at $(x^*, y^*) =$

$(1/\alpha, (b(1/\alpha) - \gamma)/\alpha)$, and the key to an additional condition for a Hopf bifurcation at this equilibrium. We find that α is similar to R_0 in some ODE models, by first observing that in a complete absence of predators, N is steady at N_0 which functions as its carrying capacity. Then, if a very small number of predators is introduced and holds roughly constant (near the equilibrium), we find $ce^{-d_1 R} - d_P$ to be the net per unit rate of change of P and hence $\alpha = ce^{-d_1 R} N_0/d_P$ is the number of newly matured predators per adult predator, produced during a predator lifetime at prey carrying capacity. Thus we may interpret it as a *replacement ratio* and indeed similar to R_0. Likewise, $1/\alpha$ can be thought of as the fraction of an average adult predator's lifetime needed for self-replacement, at prey carrying capacity. All such considerations are with respect to a common unit of measure for predator and prey, for example, biomass.

3. $\gamma = d/d_P$, which is $b(1)$ in consequence of our scaling of x. Similarly to interpreting $1/d_P$ above, we find that γ is the ratio of the lifetime of a predator to that of a prey. In scenarios such as the present in which maturation time of predators is deemed important to include in the model, while that of prey is not, we expect $\gamma > 1$.
4. $b(x^*)/b(1) = B(x^* N_0)/d$ is the prey lifetime recruitment (per unit) at equilibrium.

6.2. Comparison of birth functions: Affine, concave up, and concave down. In this section we numerically investigate three implementations of the birth function,

$$b_1(x) = a\left(1 - \frac{x}{b}\right), \tag{67a}$$

$$b_2(x) = \frac{c}{1 + dx}, \tag{67b}$$

$$b_3(x) = p\left(1 + \frac{1}{x - q}\right), \tag{67c}$$

in order to better understand the behavior generally of the nondimensional models from section 5, especially with regard to the bifurcations guaranteed by Theorem 14. We determine the pairs of coefficients $((a, b), (c, d), (p, q))$ so that the implementations of the b_i have the same interior equilibrium (x^*, y^*), the same value of $b_i(x^*)$ and the same model coefficients α, γ, and with values so that the conditions for Hopf bifurcation are satisfied. Notice that over the domain of interest, b_1 is affine; b_2 is concave up; and b_3 is concave down. Although b_1 does not satisfy the condition $(xb(x))' \geq 0$ if $x > b/2$, we use it only on $0 < x < b/2$ where that condition is satisfied. The function can be redefined elsewhere to satisfy the condition if desired, with no effect on our computations. All the numerical computations use the Matlab-based package DDE-BIFTOOL [18].

Implementation. The interpretations above guide us in choosing values for $\beta = b(x^*)$ and $\gamma = b(1)$. The parameter α determines the equilibrium (x^*, y^*) through

$$x^* = 1/\alpha$$
$$y^* = (b(x^*) - b(1))\, x^* = (b(1/\alpha) - \gamma)/\alpha.$$

We then solve for the two parameters in each of the birth functions b_i. There are elementary but tedious details to be checked that all the requirements on the birth function and α are satisfied, which we omit. The resulting restrictions are listed here.

PROPOSITION 15. *For $i = 1, 2, 3$, there is an $\alpha_i > 1$ such that the condition (2) of Theorem 14 for bifurcation in the model based on b_i is satisfied if and only if $\alpha > \alpha_i$. (Thus the α_i satisfy the criteria of α^* above.) Furthermore, whenever $\beta > \gamma > 0$ and the restrictions immediately below hold, positive parameters for the birth function b_i are uniquely determined such that with $x^* = 1/\alpha < 1/\alpha_i$ we have $b_i(x^*) = \beta$ and $b_i(1) = \gamma$.*

Table 1 summarizes the relevant relations and conditions.

TABLE 1

$b_1(x) = a\left(1 - \frac{x}{b}\right)$	$a = \frac{\alpha\beta-\gamma}{\alpha-1}, b = \frac{\alpha\beta-\gamma}{\alpha(\beta-\gamma)}$ $(xb(x))' > 0$ if $\gamma < \beta < 2\gamma$ $\alpha_1 = 3$
$b_2(x) = \frac{c}{1+dx}$	$c = \beta\frac{\alpha-1}{\alpha-\beta/\gamma}, d = \alpha\frac{\beta/\gamma-1}{\alpha-\beta/\gamma}$ Both c, d are positive iff $\alpha > \beta/\gamma > 1$. $\alpha_2 = \left(3 + d + \sqrt{9 + 10d + d^2}\right)\Big/2$
$b_3(x) = p\left(1 + \frac{1}{x-q}\right)$	p = larger zero of $(\alpha-1)Z^2 - (\beta(2\alpha-1)-\gamma)Z + (\alpha-1)\beta\gamma$ $q = (\beta - p - p\alpha)/(\alpha(\beta-p))$ $(xb(x))' > 0$ if $q > \frac{3}{2} + \frac{1}{2}\sqrt{5} \approx 2.62$ $\alpha_3 = \frac{1}{2q}\left(3q - 1 + \sqrt{9q^2 - 10q + 1}\right)$

Discussion. We saw above that we must respect $0 < \gamma < \beta$, while numerically we observed that this together with $\beta \leq 3\gamma/2$ was necessary and sufficient for each b_i to be implemented over an interval for α including $[\alpha_i, 7]$ in its interior. For each b_i in combination with each of the pairs $(\beta, \gamma) = (4.5, 3), (8, 6)$ we computed a locus of Hopf bifurcation points in the (α, r) plane which included $5 \leq \alpha \leq 7$ in all cases. Then in each of the cases $\alpha = 5, 7$, we computed bifurcation diagrams with the delay r as bifurcation parameter, and plotted: (1) bifurcation diagrams, (2) profiles of the bifurcating periodic solutions (traces of the solutions $(x(t), y(t))$), and (3) the largest size of a nontrivial Floquet multiplier. (See [6] for more extensive graphics.) These computations were done for values of the delay r beginning at the bifurcation value and ending at about $r = 2$. Considering that we scaled time by the measure of a predator lifetime, $1/d_P$, a delay of $r = 2$ corresponds to a juvenile predator maturation that is twice this magnitude, and hence is more than adequate for biological considerations.

There were remarkable similarities across these computations.

1. All the bifurcation loci in the (α, r) plane were decreasing, concave up.
2. All the bifurcation diagrams of the x-amplitude of bifurcating period solutions vs. r were increasing, concave down, with the amplitude approaching 1 as r increased. (Recall that the "carrying capacity" for x is normalized to 1.)
3. All the bifurcation diagrams of the period of bifurcation periodic solutions vs. r were increasing and almost linear for a long range of r. However, this broke down as discussed below.
4. All the profile plots of the periodic solutions (that is, projections of periodic solutions $(x(t), y(t)$ into an (x, y) plane) showed initial nesting of each surrounding those of lesser r, expanding with increasing r but contained within a triangle adjacent to the origin and coordinate axes in the (x, y) plane, and then developing an "overhang." See Figure 2 for a typical scenario.
5. Floquet multipliers (These are eigenvalues of the linearized Poincaré map associated with each bifurcating periodic solution and always "trivially" in-

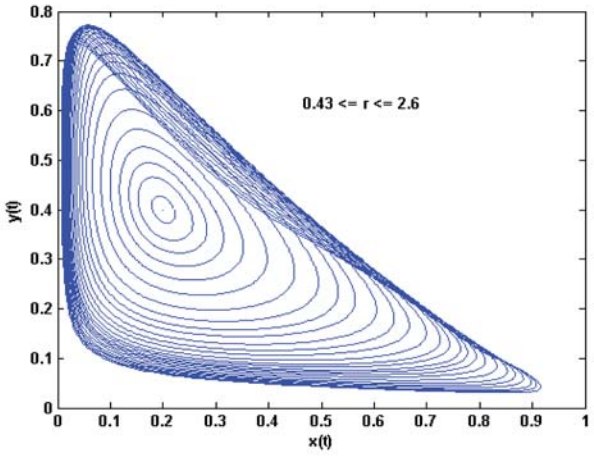


FIG. 2. *Overhang in bifurcating family of periodic solution profiles.*

clude 1 corresponding to the initial condition of the periodic solution. Only a finite number can lie outside any disk centered at the origin in the complex plane.): In plotting the largest magnitude of these (excepting the trivial 1) vs. r, we found rapid decrease to small values across all our examples, followed in some circumstances by eventual increase (but not in ranges of r that are biologically compelling).

Prompted by the eventual increase of Floquet multipliers for $r > 2$ we singled out the example case of birth function b_1 with $\alpha = 5$, $\beta = 8$, $\gamma = 6$ for further investigation and extended the delay r from the bifurcation value of about 0.43 to over 8. Some interesting aspects emerged, which although not of direct biological interest in the current context, might easily emerge in others. They include:

1. Self-crossover of periodic orbit profiles: At $r \approx 3.8$ we observe the beginnings of self-intersection within profiles of the bifurcating periodic orbits. See the sequence in Figure 3. Since the periods of these solutions increased approximately affinely with the delay over this regime, perhaps this extra looping of the profile can be understood as a mechanism by which the longer period is accommodated in a limited spatial region. There were no remarkable aspects of the Floquet multipliers (measures of stability) apparently associated with these behaviors.
2. The profiles of the self-intersecting orbits also approach the equilibrium at $(1, 0)$ (thereby slowing down) and develop folds and spikes in their upper left corner, again with the effect of enabling longer periods. See the sequence in Figure 3.
3. It is quite remarkable that over $0.43 < r < 8$ the bifurcation diagram of period vs. r is approximately affine with the notable exception of $7.12 < r < 7.29$ where it doubles over in a backwards S. Moreover, the vertical tangencies (in the period vs. delay bifurcation diagrams) at $r = 7.12, 7.29$ are accompanied by a Floquet multiplier leaving the unit disk in the complex plane. Notice that the self-crossovers remarked on above appear to be completely independent of the branch of multipliers that leaves the unit disk over $4.8 < r < 6.3$. See Figure 4.

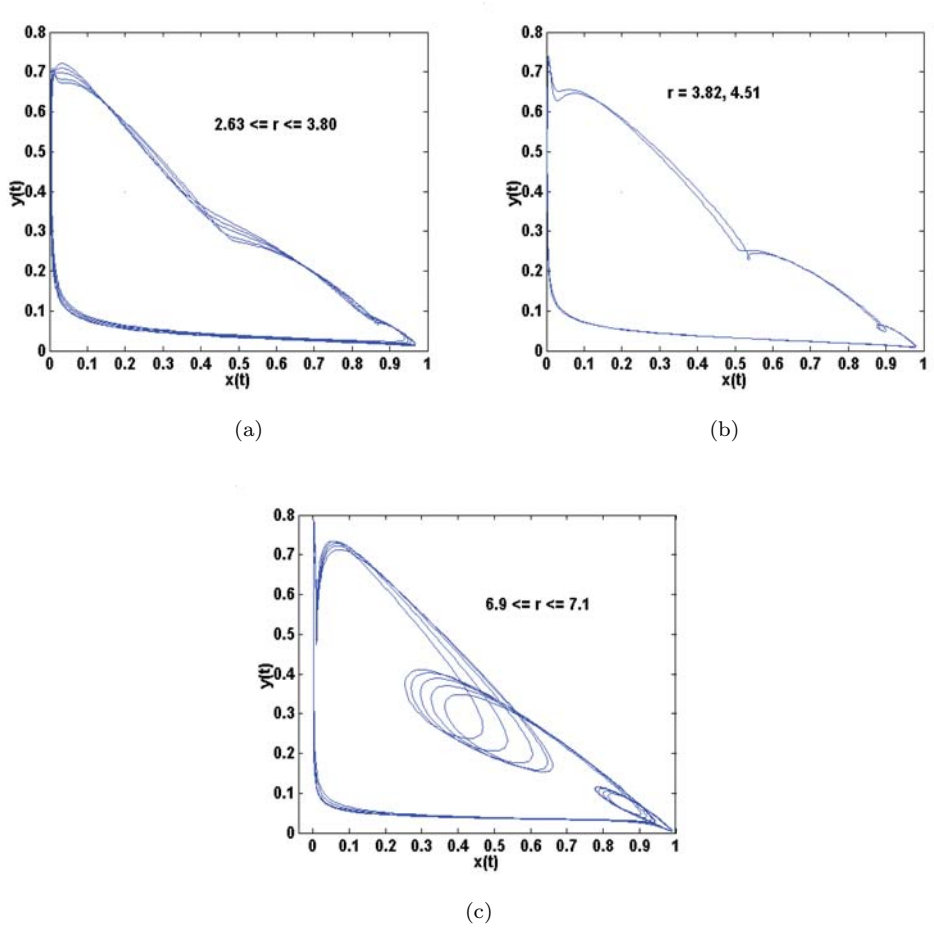


FIG. 3. *Progression of self-intersecting profiles of bifurcating period orbits. (a) As r increases to* 3.8*, a cusp develops (lower right) and early stages of a fold are seen (upper left). (b) Intermediate stages of both self-intersection and fold. (c) Late stages at large r.*

7. Equivalence of dimensional and nondimensional results. There are some concerns that might arise concerning the scaling of the original models (5) in the host-parasite case (and (48) in the predator-prey case) and how analysis of the resulting dimension-free resulting models (respectively, (9) and (50)) applies back to those original models. In particular, there might be some concern about the use of the scaled delay, r, as a bifurcation parameter in (50) after a coefficient in (48) which "physically" contained the original delay, R, has been scaled out. Furthermore, there might arise concern about the validity of the bifurcation analysis since the dimension of the parameter space is reduced from eight (including the delay and N_0) in the original models to three (including the scaled delay) in the corresponding dimension-free ones. Since the relation between (48) and (50) scales out the "physical" presence of the delay and is therefore more complicated, we will focus on that situation, leaving the other host-parasite situation of section 3 as an easy corollary. We should note that Beretta and Kuang [1] address these issues without passing to a scaled version of the model. Our situation, however, can be addressed directly.

Let us refer to (48a–48c), together with the assumptions (2a), (2b), and (7), as the dimensional model. Recall that we defined N_0 by the condition $B(N_0) = d$. Then

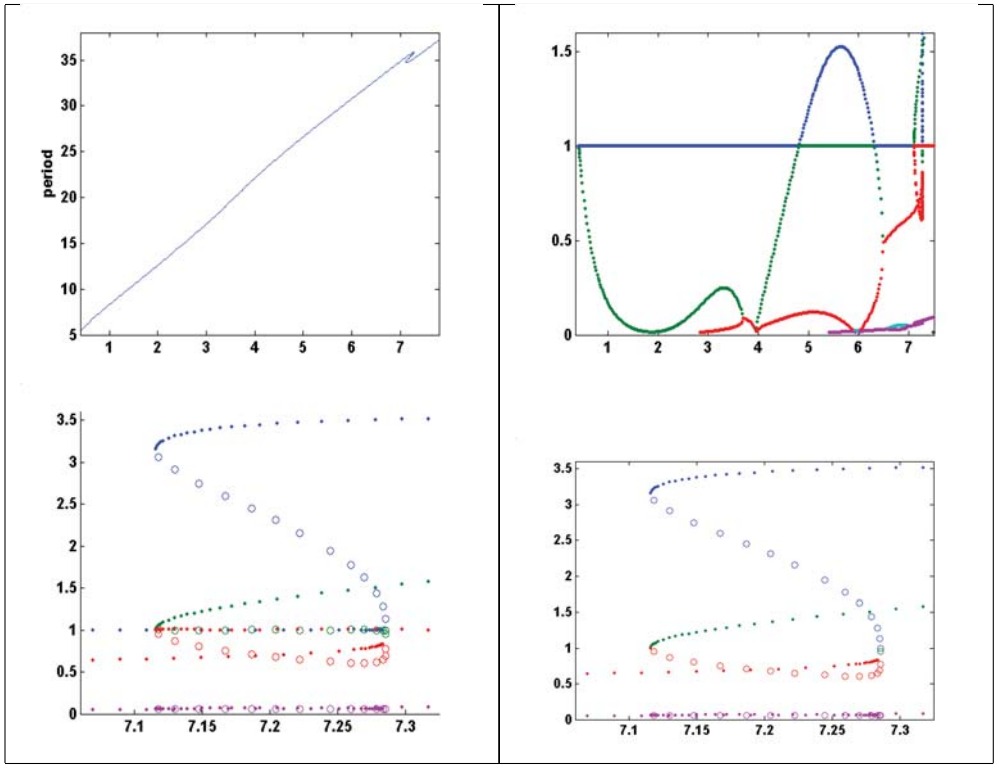


FIG. 4. *With the maturation delay, r, on the horizontal axis: Upper left shows periods, including the backwards S. All other frames show absolute values of Floquet multipliers larger than 0.01: Upper right shows values < 7.5. Lower frames have dots plotted as both r and the period increase, first over $7.05 < r < 7.29$ (corresponding to the top of the backwards S in the bifurcation diagram), and later over $7.12 < r < 7.35$ (corresponding to the bottom of the backwards S). To contrast circles are plotted as r decreases from 7.29 to 7.12 through the middle of the backwards S. The lower right is the same as lower left with values of 1 removed to fascilitate observation of a branch of complex conjugate multipliers (shown only in absolute value) becoming real at $r = 7.29$, with one leaving the unit disk there, and the other leaving after r decreases to 7.12.*

we scaled variables and parameters according to (49), resulting in the nondimensional model

$$\frac{dx}{dt} = x(t)\, b(x(t)) - \alpha x(t)\, y(t) - \gamma x(t), \tag{68a}$$

$$\frac{dy}{dt} = \alpha x(t-r)\, y(t-r) - y(t), \tag{68b}$$

satisfying

$$b(1) = \gamma. \tag{68c}$$

The purpose of this section is to show that bifurcation in the latter system (68) with respect to r implies bifurcation in the former, (48), with respect to R, and conversely.

Suppose now that $(\alpha, \gamma, r, b(\cdot))$ are given and consider the corresponding system (68). Also consider a system (48) in which $(a, c, d, d_1, d_P, N_0, R, B(\cdot))$ are given such

that the parametric relations of (49) hold, namely,

$$\begin{aligned} ce^{-d_1 R} N_0/d_P &= \alpha, \\ d/d_P &= \gamma, \\ d_P R &= r, \\ B\left(xN_0\right)/d_P &= b\left(x\right). \end{aligned} \tag{69}$$

Then, given a solution $(x(t), y(t))$ of (68), if

$$\begin{aligned} T &= t/d_P, \\ N(T) &= N_0 x\left(d_P T\right), \\ P(T) &= d_P \alpha y\left(d_P T\right)/a, \end{aligned} \tag{70}$$

it is easy to see that $(N(T), P(T))$ solves (48). In this way (while parameters are fixed and appropriately related), solutions of the two systems are related by a bicontinuous linear isomorphism. In this context, stability of equilibria and periodic solutions carries over from (68) to (48), and conversely.

In examining implications for bifurcation, we have already seen that it is not possible with our choice of scaling to vary only the delays, r and R. However, suppose again that $(\alpha, \gamma, r, b(\cdot))$ are given with $(\alpha, \gamma, b(\cdot))$ fixed, but with r varying, and let $(c, d, d_1, d_P, N_0, B(\cdot))$ be smooth functions of $R = r/d_P$ such that (69) and $B(N_0) = d$ continue to hold. Again, given a solution $(x(t), y(t))$ of (68), if (69) holds it is easy to see that $(N(T), P(T))$ solves (48) for each value of r with $R = r/d_P$. Moreover the triples $(x(\cdot), y(\cdot), r)$ and $(N(\cdot), P(\cdot), R)$ are related by a bicontinuous bijection, based on the bicontinuous linear isomorphism between $(x(\cdot), y(\cdot))$ and $(N(T), P(T))$ above, but with nonlinear inclusion of r. This correspondence maintains equilibrium, periodicity, and stability properties of solutions, so that if a bifurcation occurs in (68) with respect to r, then a corresponding one in (48) must also occur with respect to R.

To address the converse question regarding implications of bifurcations in (48) for bifurcations in (68), let us consider the case of R as bifurcation parameter in (48) with (a, c, d, d_1, d_P, N_0) held fixed. Then assuming our usual relations between that system and (68), we again have the triples $(N(\cdot), P(\cdot), R)$ and $(x(\cdot), y(\cdot), r)$ related by a bicontinuous bijection that preserves equilibrium, periodicity, and stability properties. Thus any bifurcation in (48) with respect to R is mirrored by one in (68) with respect to r, but with α also varying (as a function of r, through (56)), perhaps with consequences, e.g., for the existence of a coexistence equilibrium as α decreases through unity.

REFERENCES

[1] Beretta, E. and Y. Kuang, *Geometric stability switch criteria in delay differential systems with delay dependent parameters*, SIAM J. Math. Anal., 33 (2002), pp. 1144–1165.

[2] C. J. Briggs, R. M. Nisbet, and W. W. Murdoch, *Delayed feedback and multiple attractors in a host-parasitoid system*, J. Math. Biol., 38 (1999), pp. 317–345.

[3] C. J. Briggs and H. C. J. Godfray, *The dynamics of insect-pathogen interactions in stage-structured populations*, Amer. Naturalist, 145 (1995), pp. 855–887.

[4] M. Cavani, M. Lizana, and H. Smith, *Stable periodic orbits for a predator-prey model with delay*, J. Math. Anal. Appl., 249 (2000), pp. 324–339.

[5] K. L. Cooke and Z. Grossman, *Discrete delay, distributed delay and stability switches*, J. Math. Anal. Appl., 86 (1982), pp. 592–627.

[6] K. L. Cooke, R. H. Elderkin, and W. Huang, *Appendix of graphics*, http://www.pages.pomona.edu/~relderkin/pub/GraphAppend.pdf (2004).
[7] K. L. Cooke and P. van den Driessche, *On zeroes of some transcendental equations*, Funkcial. Ekvac., 29 (1986), pp. 77–90.
[8] K. L. Cooke, P. van den Driessche, and X. Zou, *Interaction of maturation delay and nonlinear birth in population and epidemic models*, J. Math. Biol., 39 (1999), pp. 332–352.
[9] J. M. Cushing, *Integrodifferential Equations and Delay Models in Population Dynamics*, Lecture Notes in Biomath. 20, Springer-Verlag, Berlin/Heidelberg/New York, 1977.
[10] O. Diekmann, S. A. van Gils, S. M. V. Lunel, and H.-O. Walther, *Delay Equations: Functional-, Complex-, and Nonlinear Analysis*, Springer-Verlag, New York, 1995.
[11] H. Freedman, *I. Deterministic Mathematical Models in Population Ecology*, 2nd ed., HIFR Consulting, Edmonton, 1987.
[12] K. Gopalsamy, *Stability and Oscillations in Delay Differential Equations of Population Dynamics*, Kluwer Academic Publishers, Dordrecht/Boston/London, 1992.
[13] S. A. Gourley and Y. Kuang, *A stage structured predator-prey model and its dependence on maturation delay and death rate*, J. Math. Biol., 49 (2004), pp. 188–200.
[14] J. K. Hale and S. M. V. Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.
[15] M. P. Hassell and S. W. Pacala, *Heterogeneity and the dynamics of host-parasitoid interactions*, Philos. Trans. R. Soc. Lond. Ser. B, 330 (1990), pp. 203–220.
[16] A. Hastings, *Age-dependent predation is not a simple process. I. Continuous time models*, Theoret. Pop. Biol., 23 (1983), pp. 347–362.
[17] A. Hastings, *Delays in recruitment at different trophic levels: Effects on stability*, J. Math. Biol., 21 (1984), pp. 35–44.
[18] K. Engelborghs, T. Luzyanina, and D. Roose, *Numerical bifurcation analysis of delay differential equations using DDE-BIFTOOL*, ACM Trans. Math. Software, 28 (2002), pp. 1–21.
[19] Y. Kuang, *Delay Differential Equations*, Academic Press, Boston, 1993.
[20] S. Leela, and V. Lakshmikantham, *Differential and Integral Inequalities*, Academic Press, New York, 1969.
[21] W. W. Murdoch, C. J. Briggs, and R. M. Nisbet, *Dynamical effects of host size- and parasitoid state-dependent attacks by parasitoids*, J. Animal Ecol., 66 (1997), pp. 542–556.
[22] W. W. Murdoch, R. M. Nisbet, S. P. Blythe, W. S. C. Gurney, and J. D. Reeve, *An invulnerable age class and stability in delay-differential parasitoid-host models*, Amer. Naturalist, 129 (1987), pp. 263–282.
[23] J. D. Murray, *Mathematical Biology*, Springer-Verlag, Berlin, 1989.
[24] R. M. Nisbet and W. S. C. Gurney, *Modelling Fluctuating Populations*, John Wiley & Sons, New York, 1982.
[25] L. Nunney, *The effect of long time delays in predator-prey systems*, Theoret. Pop. Biol., 27 (1985a), pp. 202–221.
[26] L. Nunney, *Absolute stability in predator-prey models*, Theoret. Pop. Biol., 28 (1985b), pp. 209–232.
[27] L. Nunney, *Short time delays in population models: A role in enhancing stability*, Ecology, 66 (1985c), pp. 1849–1858.
[28] P. Turchin, A. D. Taylor, and J. D. Reeve, *Dynamical role of predators in population cycles of a forest insect: An experimental test*, Science, 285 (1999), pp. 1068–1071.

SIAM J. APPL. MATH.
Vol. 66, No. 3, pp. 1080–1099

EXISTENCE AND STABILITY OF SPHERICALLY LAYERED SOLUTIONS OF THE DIBLOCK COPOLYMER EQUATION*

XIAOFENG REN† AND JUNCHENG WEI‡

Abstract. The relatively simple Ohta–Kawasaki density functional theory for diblock copolymer melts allows us to construct and analyze exact solutions to the Euler–Lagrange equation by singular perturbation techniques. First, we consider a solution of a single sphere pattern that models a cell in the spherical morphology. We show the existence of the sphere pattern and find a stability threshold, so that if the sphere is larger than the threshold value, the sphere pattern becomes unstable. Next we study a spherical lamellar pattern, which may be regarded as a defective lamellar pattern. We reduce the existence and the stability problems to some finite dimensional problems which are accurately solved with the help of a computer. We find two thresholds. Only when the size of the sample is larger than the first threshold value does a spherical lamellar pattern exist. This patten is stable only when the size of the sample is less than the second threshold value. As the stability of the spherical lamellar pattern changes at the second threshold, a bifurcating branch with a pattern of wriggled spherical interfaces appears. The free energy of the latter pattern is lower than that of the first pattern. A similar bifurcation phenomenon also occurs in the single sphere pattern at its stability threshold.

Key words. Ohta–Kawasaki diblock copolymer theory, sphere pattern, optimal size, spherical lamellar pattern, existence threshold, stability threshold, bifurcation, wriggled sphere pattern, wriggled spherical lamellar pattern

AMS subject classifications. 34E05, 82D60

DOI. 10.1137/040618771

1. Introduction.

A diblock copolymer melt is a soft material, characterized by fluid-like disorder on the molecular scale and a high degree of order at a longer length scale. A molecule in a diblock copolymer is a linear subchain of A-monomers grafted covalently to another subchain of B-monomers. Because of the repulsion between the unlike monomers, the different types of subchains tend to segregate, but as they are chemically bonded in chain molecules, segregation of subchains cannot lead to a macroscopic phase separation. Only a local microphase separation occurs: microdomains rich in A-monomers and microdomains rich in B-monomers emerge as a result. These microdomains form morphology patterns/phases.

There are two types of phase separations in a diblock copolymer system: weak segregation and strong segregation. The weak segregation occurs when the temperature is relatively high. The microdomains are small and there are no clear interfaces separating them. When the temperature is lower, strong segregation is observed. The microdomains become larger and they are separated by narrow interfaces.

The self-consistent mean field theory [11, 13, 14, 15, 17, 18] is the most successful theory in modeling and capturing aspects of the phase separation. It consists of five equations for five field variables: two density fields of A- and B-monomers, two mean fields on A- and B-monomers simulating the interaction between the molecular chains, and a Lagrange multiplier field. Two of the five equations are nonlocal, while

*Received by the editors November 11, 2004; accepted for publication (in revised form) November 30, 2005; published electronically March 10, 2006.
http://www.siam.org/journals/siap/66-3/61877.html

†Corresponding author. Department of Mathematics and Statistics, Utah State University, Logan, UT 84322-3900 (ren@math.usu.edu). This author was supported in part by NSF grant DMS-0509725.

‡Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong (wei@math.cuhk.edu.hk). This author was supported in part by an Earmarked Grant of RGC of Hong Kong (RGC Project 402304).

the remaining three are algebraic [22]. The theory is derived from a microscopic description of interacting polymer chains. Based on a variational principle, the Gibbs canonical distribution is approximated by the distribution generated by the mean fields [7]. This theory is quite complex to which only numerical studies have been done. One of them is the spectral method of Matsen and Schick [22] that yields predictions with striking resemblances to experiments.

A limitation of such techniques or other test field based methods is that they proceed by assuming a periodic structure, computing its free energy and then comparing that free energy to the free energy of other candidate structures [3]. The patterns found by such methods in general do not exactly solve the self-consistent equations. However, finding analytic solutions to these equations is very difficult due to the complexity of the two nonlocal equations.

The density functional theory of Ohta and Kawasaki [28] is a much simpler model. The free energy of a diblock copolymer melt is an elegant functional of the A-monomer density field only. Unlike an earlier density functional theory of Leibler [20] that deals only with the weak segregation region, the Ohta–Kawasaki theory deals with both the weak- and strong-segregation phenomena. The Euler–Lagrange equation of the Ohta–Kawasaki free energy is an integro-differential equation (see (2.9)), which can also be viewed as a system of two elliptic partial differential equations (see (2.11)–(2.12)).

A close examination of the derivation of the density functional theory shows that it is a simplified version of the self-consistent mean field theory. We refer the reader to [7] for a detailed study of the simplification procedure. Here we briefly summarize the results of [7]. There are two approximation steps. First, we consider the relationship between the A-monomer density field u_a, the B-monomer density field u_b, and the mean fields U_A, U_B that act, respectively, on A- and B- monomers. In the self-consistent mean field theory one can express u_a and u_b in terms of U_A and U_B with the help of Feynman integrals, i.e., by solving some parabolic partial differential equations. In the density functional theory we simplify this relationship via linearization. This approximation is accurate if the temperature is not too low. Then we reverse the linearized relationship between u_a, u_b and U_A, U_B to express U_A, U_B in terms of u_a, u_b. An analysis in the Fourier space shows that this reversed relationship is described by a pseudodifferential, nonlocal operator. In the second approximation step we keep the long wave and the short wave parts of this operator and discard the intermediate wave effects. This way we end up with a sum of the Laplace operator $-\Delta$ and the inverse Laplace operator $(-\Delta)^{-1}$. When we finally express the free energy as a functional of u_a and u_b in the density functional theory, the Laplace operator gives rise to the local part of the functional and the inverse Laplace operator leads to the nonlocal part of the functional (see (2.2)).

Despite the shortcomings associated with these approximations, the density functional theory at least qualitatively captures the properties of diblock copolymers [21, 10]. Ohta and Kawasaki used their theory to study the common lamellar, cylindrical, and spherical phases [28]. More recently Teramoto and Nishiura found the less common double gyroid morphology by numerically simulating the theory [43]. Although Ohta and Kawasaki applied their theory only to test fields and did not construct exact solutions of the Euler–Lagrange equation, we will show that the simplicity of the theory actually makes it possible to study exact solutions analytically.

The weak segregation regime may be studied by the bifurcation theory rather easily. One starts with a uniform state and linearizes the Euler–Lagrange equation at the uniform state. For some parameter values the principal eigenvalue of the linearized problem is zero. Then a nonuniform state bifurcates from the uniform state. If one

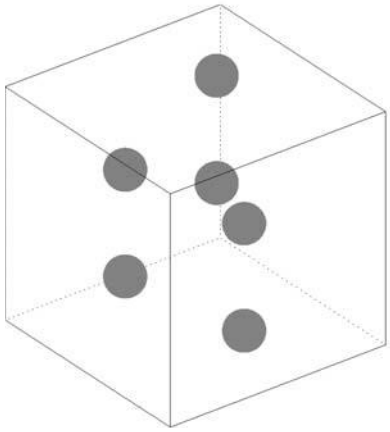


FIG. 1. *The spherical phase. The B-monomers form dark spheres and the A-monomers occupy the background. Not reflected in this figure is the body centered cubic pattern in which the spheres pack.*

can show that this nonuniform state is stable, then it gives the profile of a weakly segregated pattern.

In this paper we use the density functional theory to study the more complex strong segregation phenomenon. Strongly segregated patterns are too different from the uniform state to be treated as their bifurcating branches. The appropriate mathematical tool is the singular perturbation theory in calculus of variations and differential equations. We find exact solutions, or at least leading order terms of exact solutions, to the Euler–Lagrange equation of the free energy functional [26, 30, 32, 31, 35, 39, 40, 9, 6, 16]. Often these solutions may be carefully analyzed and their stability in space can be determined [33, 38].

The first strongly segregated pattern we study is the single sphere pattern. This pattern arises from the spherical phase of a diblock copolymer. When the monomer fraction is skewed in favor of A-monomers, the B-monomers form spherical microdomains; see Figure 1. We find a pattern with one sphere of B-monomers as a solution to the Euler–Lagrange equation. There is an optimal size for the sphere. The sphere of optimal size has lower free energy than those of other spheres. Then we linearize the Euler–Lagrange equation at the sphere solution and study the spectrum of the linearized operator. We will show that there is an upper bound for the size of the sphere. Beyond the upper bound, the sphere cannot be stable.

The second pattern is the spherical lamellar pattern. This pattern may be viewed as a defective lamellar pattern. Other defective patterns are considered in [44], where a model of a fourth-order differential equation is used. Given the number of interfaces we look for a solution that consists of spherical layers of microdomains separated by narrow interfaces. In this case the existence and stability problems are reduced by singular perturbation techniques to some finite dimensional problems. The reduced problems are easily solved with the help of a computer. Note that here we apply numerical methods to the reduced finite dimensional problems only. This way we obtain far more accurate and reliable results compared to results found from direct numerical simulations of infinite dimensional problems. There is an existence threshold. Only when the sample is greater than this threshold does a spherical lamellar pattern exist. There is also a stability threshold which is greater than the existence threshold.

When the sample is larger than the stability threshold, the spherical lamellar pattern becomes unstable.

We emphasize that all the results presented in this paper are mathematically rigorous. The informal style adopted here when we describe perturbation expansions can be changed to a strict mathematical framework, part of which is known as the Γ-convergence theory [8, 24, 23, 19].

2. Free energy. We review the density functional theory of Ohta and Kawasaki [28] in this section. We consider a diblock copolymer melt that occupies a region D in space. The system has the following parameters.

1. The polymerization index N that is the number of all the monomers in a chain molecule.
2. The A-monomer number N_A and the B-monomer number N_B in a chain. Note that $N_A + N_B = N$.
3. The number of chain molecules in the melt n. The average total monomer number density is $\rho_0 = \frac{nN}{V}$.
4. The Kuhn statistical length l measuring the average distance between two adjacent monomers in a chain molecule.
5. The inverse absolute temperature β.
6. The dimensionless Flory–Huggins parameter χ that measures the repulsion between unlike monomers; it is defined by

$$\chi = \beta\left(V_{AB} - \frac{V_{AA} + V_{BB}}{2}\right), \tag{2.1}$$

   where V_{AB} (and V_{AA} and V_{BB}, respectively) is the energy cost to bring an A-monomer (A-monomer and B-monomer, respectively) and a B-monomer (A-monomer and B-monomer, respectively) close to each other. This number is positive because the repulsion force between unlike monomers is stronger than those between like ones. Note that χ is inversely proportional to temperature.
7. The volume V of the sample. The domain D is nondimensionalized so that the size of D, denoted by $|D|$, is a convenient value. In this paper D is a ball, so we take the radius of D to be 1 and the size of D to be $|D| = 4\pi/3$.

The main field in the Ohta–Kawasaki theory is the relative A-monomer number density field $u(x)$. The melt is assumed to be incompressible, so when $u(x) = 1$ (or $u(x) = 0$, respectively), the point x in D is occupied by A-monomers only (or B-monomers only, respectively); if $0 < u(x) < 1$, a mixture of A- and B-monomers occupies x. The free energy F of the system is a functional of $u(x)$. In a dimensionless form we write

$$\frac{\beta F}{\chi \rho_0 V} = \int_D \left[\frac{\epsilon^2}{2}|\nabla u|^2 + W(u) + \frac{\epsilon\gamma}{2}|(-\Delta)^{-1/2}(u-a)|^2\right] dx. \tag{2.2}$$

On the right side of (2.2) we have introduced three dimensionless parameters:

$$\epsilon^2 = \frac{|D|^{2/3} l^2}{12a(1-a)\chi V^{2/3}}, \tag{2.3}$$

$$\gamma = \frac{18\sqrt{3}V}{|D| a^{3/2}(1-a)^{3/2}\chi^{1/2}N^2 l^3}, \tag{2.4}$$

$$a = \frac{N_A}{N}. \tag{2.5}$$

Note that the parameter $a = N_A/N$ is the average A-monomer density. The field u must satisfy the constraint

$$\overline{u} = a, \tag{2.6}$$

where $\overline{u} = \frac{1}{|D|}\int_D u(x)\,dx$ is the average of u.

The exact form of W is not given in [28]. In [7] an approximation

$$W(u) = \begin{cases} u - u^2 & \text{if } u \in [0,1], \\ \infty & \text{otherwise} \end{cases} \tag{2.7}$$

is found. A more accurate W should be a smooth double well function of equal depth. It must have a global minimum value 0 achieved at 0 and 1. It must have the symmetry $W(u) = W(1-u)$. 0 and 1 are nondegenerate: $W''(0) = W''(1) > 0$.

Central in (2.2) is the third term in the integrand. It is nonlocal and models the long range interaction between monomers due to the connectivity of the molecular chains. The operator $(-\Delta)^{-1/2}$ is the square root of the inverse of $-\Delta$ with the natural boundary condition. Alternatively in (2.2) one may write

$$\int_D |(-\Delta)^{-1/2}(u-a)|^2\,dx = \int_D\int_D (u(x)-a)G(x,y)(u(y)-a)\,dxdy, \tag{2.8}$$

where G is the Green's function of $-\Delta$ with the natural boundary condition.

The second term in (2.2) can be regarded as the internal energy field of the system, and the first and the third terms give the entropy of the system. As mentioned in the introduction we have taken only the long wave and short wave effects, modeled by the ∇ and $(-\Delta)^{-1/2}$ operators, into consideration in this model.

When we minimize (2.2), the first term in the integrand of (2.2) penalizes any space nonuniformity. The second term favors u being close either to 0 or close to 1 everywhere. The best profile for the third term is to have u close to a everywhere. However, this is not a good profile for the second term. The second best profile for the third term is for u to have many oscillations. Local minimizers of the free energy result from these three competing preferences.

The Euler–Lagrange equation of (2.2) is a nonlinear integro-differential equation

$$-\epsilon^2\Delta u + f(u) + \epsilon\gamma(-\Delta)^{-1}(u-a) = \eta \text{ in } D \tag{2.9}$$

subject to the natural boundary condition

$$\partial_\nu u = 0 \text{ on } \partial D. \tag{2.10}$$

Here $f = W'$. The constant η on the right side of (2.9) is a Lagrange multiplier coming from the constraint (2.6). Equation (2.9) may also be written as a system of elliptic partial differential equations

$$-\epsilon^2\Delta u + f(u) + \epsilon\gamma v = \eta, \tag{2.11}$$

$$-\Delta v = u - a \tag{2.12}$$

subject to the conditions

$$\partial_\nu u = \partial_\nu v = 0 \text{ on } \partial D, \quad \overline{u} - a = \overline{v} = 0. \tag{2.13}$$

Note that (2.9) always has the uniform solution $u(x) = a$. When ϵ is large, corresponding to high temperature, this solution is stable and it models the disordered phase. One may decrease ϵ to a value so that the principal eigenvalue of the linearized problem at $u(x) = a$ becomes 0. Then one finds a nonuniform solution bifurcating out of the uniform solution. This bifurcation solution explains the weak segregation phenomenon and the corresponding ϵ identifies the parameter range for weak segregation.

However, in the strong segregation regime ϵ is much smaller. In this case the free energy (2.2) is most easily analyzed in the parameter range

$$\epsilon \ll 1, \tag{2.14}$$

$$\gamma \sim 1. \tag{2.15}$$

Here the uniform solution $u(x) = a$ has much higher free energy than those of many other states and is hence thermodynamically unfavored. Under (2.14)–(2.15), we are in the strong segregation regime and have taken the volume of the sample to be of order

$$V \sim a^{3/2}(1-a)^{3/2}\chi^{1/2}N^2l^3. \tag{2.16}$$

We will see that in the parameter range (2.14)–(2.15) the number of microdomains is of order 1. Therefore the right side of (2.16) also predicts the size of a microdomain. Particularly we find the domain size $V^{1/3}$ to be proportional to $N^{2/3}$, which is the celebrated $N^{2/3}$ law [28].

Having a small ϵ makes (2.9) a singular perturbation problem. Although in mathematics the singular perturbation theory is much harder and less mature than the regular perturbation theory, a great deal of quantitative properties of solutions to (2.9) can be obtained, using the existing techniques in the theory. Many problems can be solved exactly in the leading order, and many other problems can be reduced to much simpler finite dimensional problems that are solved with the help of a computer.

3. Sphere pattern. When the monomer fraction a is close to 1, the diblock polymer typically exists in the spherical phase. B-rich microdomains form spheres and pack in a body centered cubic (BCC) pattern. Here we study a single sphere (see Figure 2) based on the model (2.2). Mathematically this must be done before we can analyze the BCC pattern. In a future publication we will "connect" several single sphere patterns to construct a BBC pattern solution in a general domain.

3.1. Existence. When one takes the domain D to be a unit ball, a radially symmetric solution $u(r)$ of (2.9) is found where u now is a function of $r = |x|$; see Figure 2. A narrow interface, whose thickness is of order ϵ, exists at some r_1, where $u(r_1) = 1/2$. The leading order of r_1 is determined by (2.6):

$$r_1 = (1-a)^{1/3} + O(\epsilon). \tag{3.1}$$

Inside the interface $u(r)$ is close to 0 and outside $u(r)$ is close to 1. The profile of u near r_1 is described by the inner expansion

$$u(r) = H\left(\frac{r-r_1}{\epsilon}\right) + \epsilon P\left(\frac{r-r_1}{\epsilon}\right) + O(\epsilon^2). \tag{3.2}$$

The leading order term H is the solution of

$$-H'' + f(H) = 0 \tag{3.3}$$

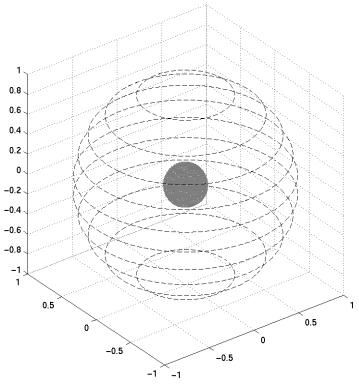


FIG. 2. *A single sphere solution with the natural boundary condition in the unit ball.*

with the conditions $H(-\infty) = 0$, $H(\infty) = 1$, and $H(0) = 1/2$. The next term P is the solution of

$$-P'' + f'(H)P - \frac{2H'}{r_1} + \frac{2\tau}{r_1} = 0, \quad P(0) = 0. \tag{3.4}$$

The definition of P involves r_1. Because $1/r_1$ is the mean curvature of the interface, the curvature affects the inner expansion of u in the ϵ order but not in the leading order, for H is independent of r_1.

In (3.4) we have a constant τ, which is the interface tension. For a general W,

$$\tau = \int_0^1 \sqrt{2W(s)}dx, \tag{3.5}$$

and for (2.7) we have

$$\tau = \frac{\sqrt{2}\pi}{8}. \tag{3.6}$$

The interface tension may also be calculated in the self-consistent mean field theory [12]. The value obtained there differs slightly from (3.6) of the density functional theory.

The free energy of this solution may be viewed as a sum of two parts. The first part comes from the two local terms of (2.2) and is equal to

$$4\pi r_1^2 \tau\epsilon + O(\epsilon^2). \tag{3.7}$$

Note that the first term on the right side is the area of the interface times τ times ϵ.

The second part of the free energy comes from the nonlocal term of (2.2) and is equal to

$$\frac{2\pi r_1^5 (r_1^3 - 3r_1 + 2)\gamma\epsilon}{15} + O(\epsilon^2). \tag{3.8}$$

Note that the free energy of the disordered phase modeled by the uniform solution $u(x) = a$ is $W(a)|D|$, a quantity of order 1, which is much larger than the free energy of the sphere pattern solution, which is of order ϵ. Hence under the condition (2.14)–(2.15) the system is in an ordered phase.

TABLE 1
The values of γ_o for various a.

a	.5	.525	.55	.575	.6	.625	.65	.675	.7	.725
γ_o	140	130	122	115	109	104	100	97	95	93

a	.75	.775	.8	.825	.85	.875	.9	.925	.95	.975
γ_o	92	92	93	95	99	106	117	137	176	290

3.2. Optimal size. There is a sphere pattern solution of (2.9) for any γ as long as ϵ is sufficiently small. This means that there is a solution for a wide range of V of the sample (the volume of the B-monomer sphere in the middle is then $(1-a)V$). It is natural to ask for which value of V the sphere pattern is most energetically favored. Intuitively we know that V cannot be too large or too small. By (2.3)–(2.4) we write $\epsilon = \tilde{\epsilon} V^{-1/3}$ and $\gamma = \tilde{\gamma} V$ so that $\tilde{\epsilon}$ and $\tilde{\gamma}$ no longer depend on V. Then by (3.7)–(3.8) we find that the leading term of the rescaled free energy of a sphere pattern is

$$4\pi r_1^2 \tau \tilde{\epsilon} V^{-1/3} + \frac{2\pi r_1^5 (r_1^3 - 3r_1 + 2)\tilde{\gamma}}{15} \tilde{\epsilon} V^{2/3}. \tag{3.9}$$

With respect to V, (3.9) is minimized at

$$V = V_o = \frac{15\tau}{r_1^3 (r_1^3 - 3r_1 + 2)\tilde{\gamma}}. \tag{3.10}$$

The optimal size of the sample is now given by (3.10). It is more convenient to express this in terms of the dimensionless γ. The optimal γ is denoted by γ_o, which is just

$$\gamma_o = \tilde{\gamma} V_o = \frac{15\tau}{r_1^3 (r_1^3 - 3r_1 + 2)}. \tag{3.11}$$

Table 1 reports the values of γ_o for various a.

3.3. Stability. We return to a sphere pattern solution with a general γ which is not necessarily equal to γ_o. Although a sphere solution of (2.9) is found for every γ, we will see that it is stable only if γ is not too large. A stable solution of (2.9) is a free energy local minimizer, which corresponds to a metastable state of the physical system. An unstable solution cannot be observed in experiments.

The stability analysis requires that we solve the eigenvalue problem

$$-\epsilon^2 \Delta \varphi + f'(u)\varphi - \overline{f'(u)\varphi} + \epsilon\gamma(-\Delta)^{-1}\varphi = \lambda\varphi. \tag{3.12}$$

The left side of (3.12) comes from linearizing the Euler–Lagrange equation (2.9) at a sphere pattern solution u. The eigenvalues λ are classified by the mode $l = 0, 1, 2, \ldots$. The eigenvalues whose modes are l are denoted by λ_l. Their corresponding eigenfunctions take the form

$$\varphi(x) = \phi_l(r) Y_{lm}(\theta, \omega), \tag{3.13}$$

where $m = 0, \pm 1, \ldots, \pm l$, and the Y_{lm}'s are the spherical harmonics. An eigenvalue either approaches 0, a critical eigenvalue, or stays positively away from 0 when $\epsilon \to 0$. Hence it suffices to consider the critical eigenvalues.

For the $l = 0$ mode there is one critical eigenvalue of order ϵ. It is of multiplicity 1 and has the form

$$\lambda_0 = \frac{3f'(0)r_1^2\epsilon}{\tau} + O(\epsilon^2). \tag{3.14}$$

This eigenvalue is positive, and $l = 0$ is a stable mode. The eigenfunction associated with this eigenvalue is radially symmetric. We denote it by $\phi_0(r)$. It has the expansion

$$\phi_0(r) = H'\left(\frac{r-r_1}{\epsilon}\right) + \epsilon P'\left(\frac{r-r_1}{\epsilon}\right) - \overline{\left[H'\left(\frac{r-r_1}{\epsilon}\right) + \epsilon P'\left(\frac{r-r_1}{\epsilon}\right)\right]} + O(\epsilon^2). \tag{3.15}$$

Here H' and P' are the derivatives of H and P, defined in (3.3)–(3.4), respectively.

For $l = 1$ there is one critical eigenvalue of order ϵ^2. It has multiplicity 3 and is of the form

$$\lambda_1 = \frac{\gamma r_1^4\epsilon^2}{\tau} + O(\epsilon^3). \tag{3.16}$$

This mode is again stable. The eigenfunctions associated with this eigenvalue are $(x/r)\phi_1(r)$, $(y/r)\phi_1(r)$, and $(z/r)\phi_1(r)$, where ϕ_1 has the expansion

$$\phi_1(r) = H'\left(\frac{r-r_1}{\epsilon}\right) + \epsilon P'\left(\frac{r-r_1}{\epsilon}\right) + O(\epsilon^2). \tag{3.17}$$

For each l greater than 1, there is one critical eigenvalue of order ϵ^2. This eigenvalue has multiplicity $2l + 1$ and has the form

$$\lambda_l = \left[\frac{l(l+1)-2}{r_1^2} + \frac{\gamma}{\tau}\left(\frac{r_1^4 - r_1}{3} + \frac{(l+1)r_1^{2l+2}}{l(2l+1)} + \frac{r_1}{2l+1}\right)\right]\epsilon^2 + O(\epsilon^3). \tag{3.18}$$

The quantity in (3.18) may not always be positive. One finds a threshold γ_s so that when $\gamma < \gamma_s$ all the eigenvalues in (3.18) are positive, but when $\gamma > \gamma_s$ at least for one l the eigenvalue λ_l in (3.18) is negative. Therefore the sphere solution u is stable if $\gamma < \gamma_s$ and unstable if $\gamma > \gamma_s$. The eigenfunctions associated with λ_l are $\phi_l(r)Y_{lm}$ with $m = 0, \pm 1, \ldots, \pm l$. ϕ_l has the same expansion as in (3.17).

The leading order of γ_s is determined from (3.18) following these steps:

1. For each $l \geq 2$, set the leading term

$$\frac{l(l+1)-2}{r_1^2} + \frac{\gamma}{\tau}\left(\frac{r_1^4 - r_1}{3} + \frac{(l+1)r_1^{2l+2}}{l(2l+1)} + \frac{r_1}{2l+1}\right) \tag{3.19}$$

   in (3.18) to be 0, and solve for γ. Denote the solution for γ by $\hat{\gamma}_l$. If this $\hat{\gamma}_l$ is less than or equal to 0, the mode l does not yield a zero eigenvalue. Discard such $\hat{\gamma}_l$.
2. Minimize the remaining $\hat{\gamma}_l$'s from the last step with respect to l. The minimum is achieved at a $\hat{\gamma}_l$ which is the leading order of γ_s.

Table 2 reports the leading order of γ_s for various a. At $\gamma = \gamma_s$ the smallest eigenvalue is 0. The mode l of this eigenvalue is also given in Table 2.

We compare the stability threshold γ_s to the optimal size γ_o in Table 1. All the γ_o's are significantly less than the corresponding γ_s's. Therefore, not surprisingly, the sphere with optimal size is stable.

TABLE 2
The (leading order) values of γ_s for various a and the corresponding mode l.

a	.5	.525	.55	.575	.6	.625	.65	.675	.7	.725
γ_s	463	425	372	336	312	296	276	250	234	225
l	5	5	4	4	4	4	3	3	3	3

a	.75	.775	.8	.825	.85	.875	.9	.925	.95	.975
γ_s	222	225	232	216	209	215	237	283	387	714
l	3	3	3	2	2	2	2	2	2	2

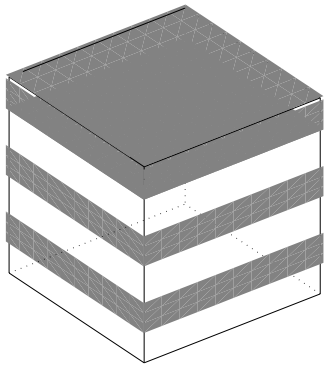


FIG. 3. *A perfect lamellar pattern.*

4. Spherical lamellar pattern. When a is close to $1/2$, the diblock copolymer exists in the lamellar phase. The perfect lamellar pattern consists of microdomains separated by parallel flat planes; see Figure 3. However, one often observes the lamellar pattern with topological defects such as dislocations, disclinations, grain boundaries, and tilt boundaries [44]. In this section we consider the spherical lamellar pattern (see Figure 4), which we view as a defective lamellar pattern.

Because it involves many interfaces, the study in this section is more complex. Nevertheless, we will show that by singular perturbation argument, solving the Euler–Lagrange equation (2.9) and analyzing the stability of the solution are reduced to studying some finite dimensional problems.

4.1. Existence. Unlike the existence problem for the sphere pattern, where no condition on γ is needed, the existence of a spherical lamellar pattern as a solution of (2.9) requires that γ is not too small. We now have an existence threshold $\gamma_{K,e}$. Given the number of interfaces $K \geq 2$ a K-interface spherical lamellar pattern solution of (2.9) exists if $\gamma > \gamma_{K,e}$. If $\gamma < \gamma_{K,e}$, there is no K-interface spherical lamellar solution.

When $\gamma > \gamma_{K,e}$, we define the interfaces r_j, $j = 1, 2, \ldots, K$, to be the radii, where $u(r_j) = 1/2$. They have the expansion

$$r_j = r_j^0 + O(\epsilon). \tag{4.1}$$

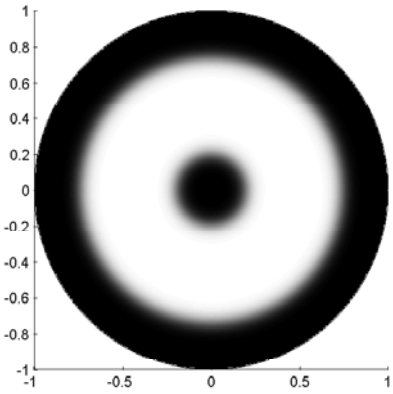


FIG. 4. *A cross section of a spherical lamellar pattern with two interfaces.*

The leading order r_j^0's are determined by solving a system of $K+1$ nonlinear equations

$$\frac{\tau}{r_j^0} + \frac{(-1)^j\gamma}{2}\mathcal{V}(r_j^0; r_1^0, r_2^0, \dots, r_K^0) = (-1)^j\eta^0, \quad j = 1, 2, \dots, K,$$

$$\sum_{j=1}^{K}(-1)^j(r_j^0)^3 + \frac{1-(-1)^K}{2} = a \tag{4.2}$$

for r_1^0, $r_2^0, \dots, r_K^0$, and η^0. Here η^0 is a Lagrange multiplier. The function $\mathcal{V}$ in (4.2) is the solution of

$$-\mathcal{V}'' - \frac{2}{r}\mathcal{V} = \mathcal{U} - a,\ \mathcal{V}'(0) = \mathcal{V}'(1) = 0,\ \overline{\mathcal{V}} = 0, \tag{4.3}$$

where

$$\mathcal{U}(r) = 0,\ \text{if } r \in (0, r_1^0),\ \text{and } \mathcal{U}(r) = 1 \text{ if } r \in (r_1^0, r_2^0), \dots . \tag{4.4}$$

Denote this solution by $\mathcal{V}(r; r_1^0, \dots, r_K^0)$, where we emphasize in its arguments that $\mathcal{V}$ depends on $r_1^0, \dots, r_K^0$. In (4.2) this $\mathcal{V}$ is evaluated at $r = r_j^0$.

The system (4.2) is the Euler–Lagrange equations of the minimizer of the function

$$J(q_1, q_2, \dots, q_K) = 3\tau\sum_{k=1}^{K} q_k^2 + \frac{3\gamma}{2}\int_0^1 \mathcal{V}'(r; q_1, \dots, q_K)^2 r^2 dr \tag{4.5}$$

subject to the constraint

$$-q_1^2 + q_2^3 + \cdots + (-1)^K q_K^3 + \frac{1-(-1)^K}{2} = a. \tag{4.6}$$

In the mathematics literature J is known as the Γ-limit of $(4\pi\epsilon/3)^{-1}I$. The Γ-limit theory thus reduces the study of the infinite dimensional problem I to the study of the finite dimensional problem J [30, 31].

Whether J has a minimizer depends on γ. In general, J has a minimizer only if γ is large. The border line is exactly the leading order of $\gamma_{K,e}$. For $K = 2$, Table 3 reports the leading order of $\gamma_{2,e}$ for various a.

TABLE 3
The leading order values of $\gamma_{2,e}$ for various a.

a	.5	.525	.55	.575	.6	.625	.65	.675	.7	.725
$\gamma_{2,e}$	171	175	180	186	194	204	216	230	249	271

a	.75	.775	.8	.825	.85	.875	.9	.925	.95	.975
$\gamma_{2,e}$	300	337	386	453	549	694	932	1379	2432	6590

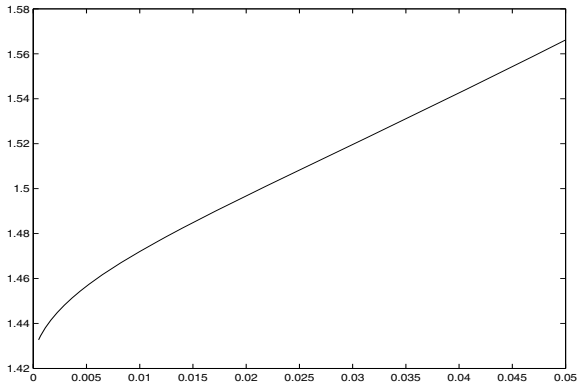


FIG. 5. *J as a function of y is increasing when $a = 1/2$ and $\gamma = 100$.*

Let us consider the $K = 2$ case in more detail. We introduce y so that

$$y = q_1^3, \quad y + a = q_2^3. \tag{4.7}$$

J can be viewed as a function of y only without the constraint (4.6). Take $a = 1/2$. For $\gamma = 100$, J is monotonically increasing in y; see Figure 5. Since $\gamma < \gamma_{2,e} \approx 171$ from Table 3, there is no 2-interface spherical lamellar solution of (2.9).

When $\gamma = 180 > \gamma_{2,e}$, J is no longer monotone; see Figure 6. In this case, J has a local minimum, and (2.9) has a 2-interface spherical lamellar solution.

If we further increase γ to 200, the local minimum of J becomes a global minimum; see Figure 7. The spherical lamellar solution continues to exist.

We now return to the general solution with K interfaces. Near each interface r_j the solution u again has a profile

$$u(r) = H\left(\frac{r - r_j}{\epsilon}\right) + \epsilon P_j\left(\frac{r - r_j}{\epsilon}\right) + O(\epsilon^2) \tag{4.8}$$

when j is odd and

$$u(r) = H\left(-\frac{r - r_j}{\epsilon}\right) + \epsilon P_j\left(-\frac{r - r_j}{\epsilon}\right) + O(\epsilon^2) \tag{4.9}$$

when j is even. H is the same function defined in (3.3) and P_j is defined by (3.2) with r_1 replaced by r_j.

The free energy of this solution is

$$\left[4\pi\tau \sum_{j=1}^{K} r_j^2 + 2\pi\gamma \int_0^1 \mathcal{V}'(r)^2 r^2 \, dr\right]\epsilon + O(\epsilon^2). \tag{4.10}$$

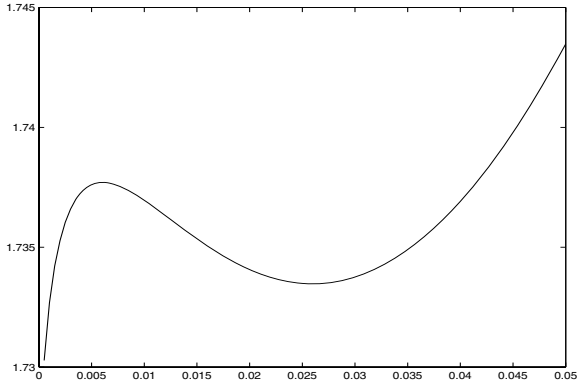


FIG. 6. *J as a function of y has a local minimum when $a = 1/2$ and $\gamma = 180$.*

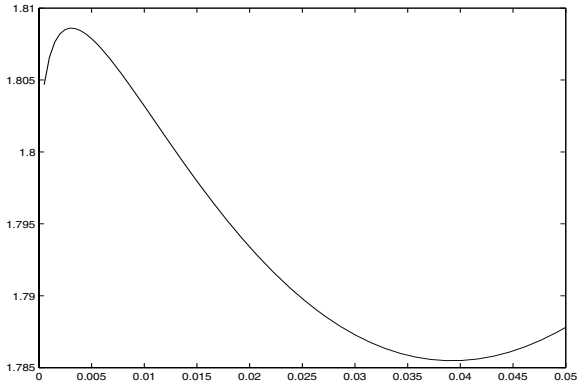


FIG. 7. *J as a function of y has a global minimum when $a = 1/2$ and $\gamma = 200$.*

4.2. Stability. Similar to the sphere pattern solution, a K-interface spherical lamellar solution is stable only if γ is not too large. More precisely, for any given number of interfaces K, there is a stability threshold $\gamma_{K,s}$, which is larger than the existence threshold $\gamma_{K,e}$, such that the K-interface spherical lamellar solution is stable if $\gamma_{K,e} < \gamma < \gamma_{K,s}$. The solution becomes unstable if $\gamma > \gamma_{K,s}$.

To verify these statements and determine $\gamma_{K,s}$ we again turn to the linearized problem (3.12). But this time u is the K-interface spherical lamellar solution found in section 4.1. The eigenvalues are again classified by the mode $l = 0, 1, 2, \ldots$. Denote the eigenvalues whose modes are l by λ_l. For each l the noncritical eigenvalues all stay positively away from 0, so it suffices to find the critical eigenvalues to determine whether u is stable.

When l is equal to 0, there exist K critical eigenvalues. All of them are simple. One of them is of order ϵ and has the expansion

$$\lambda_0 = \frac{3f'(0)\sum_{k=1}^{K}(r_k^0)^2}{\tau}\epsilon + O(\epsilon^2), \tag{4.11}$$

which is positive. The associated eigenfunction is

$$\phi_0(r) = \sum_{j=1}^{K} \left\{ H'\left(\frac{r-r_j}{\epsilon}\right) + \epsilon P_j'\left(\frac{r-r_j}{\epsilon}\right) - \overline{\left[H'\left(\frac{r-r_j}{\epsilon}\right) + \epsilon P_j'\left(\frac{r-r_j}{\epsilon}\right)\right]} \right\} + O(\epsilon^2). \tag{4.12}$$

The remaining $K-1$ eigenvalues of mode $l=0$ are of order ϵ^2. Let us expand them as

$$\lambda_0 = \mu_0 \epsilon^2 + O(\epsilon^3). \tag{4.13}$$

The determination of μ_0 is more complex.

First, we define a K by K matrix M whose kj-entry is

$$\begin{cases} (-\frac{2\tau}{(r_k^0)^2} + \gamma(-1)^k \mathcal{V}'(r_k^0)) + \gamma G_0(r_k^0, r_k^0) & \text{if} \quad k = j, \\ \gamma G_0(r_k^0, r_j^0) & \text{if} \quad k \neq j, \end{cases} \tag{4.14}$$

where G_0 is a Green's function:

$$G_0(r,s) = \begin{cases} \frac{s^2 r^2}{2} + s - \frac{9s^2}{5} + \frac{s^4}{2} & \text{if} \quad r < s, \\ \frac{s^2 r^2}{2} + \frac{s^2}{r} - \frac{9s^2}{5} + \frac{s^4}{2} & \text{if} \quad r \geq s. \end{cases} \tag{4.15}$$

Then we set a nonstandard inner product

$$\langle A, B \rangle = \sum_{k=1}^{K} A_k B_k (r_j^0)^2. \tag{4.16}$$

Let $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K$ be an orthonormal basis with respect to the inner product (4.16) and

$$\mathbf{e}_1 = \frac{(1,1,\ldots,1)}{\sqrt{\langle (1,1,\ldots,1), (1,1,\ldots,1) \rangle}}. \tag{4.17}$$

The μ_0's are determined from a $K-1$ dimensional eigenvalue problem:

$$\sum_{m=2}^{K} d_m N_{mn} = \mu_0 \tau d_n, \quad n = 2, 3, \ldots, K. \tag{4.18}$$

The $K-1$ by $K-1$ matrix N is obtained by projecting the K by K matrix M into the $K-1$ dimensional subspace spanned by $\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_K$:

$$N_{mn} = \langle M \mathbf{e}_m, \mathbf{e}_n \rangle, \quad m, n = 2, 3, \ldots, K. \tag{4.19}$$

The inner product in (4.19) is the one defined in (4.16).

These critical eigenvalues all turn out to be positive. This follows as a consequence of section 4.1. That N is positive definite is equivalent to the fact that r_j^0 minimizes J, defined in (4.5). The latter condition is fulfilled when $\gamma > \gamma_{K,e}$. Hence $l=0$ is a stable mode. To each one of these $k-1$ λ_0's, the corresponding eigenfunction $\phi_0(r)$ is given with the help of the eigenvector $(d_2, d_3, \ldots, d_K)$ of (4.18):

$$\phi_0(r) = \sum_{j=1}^{K} c_j \left[H'\left(\frac{r-r_j}{\epsilon}\right) + \epsilon P_j'\left(\frac{r-r_j}{\epsilon}\right) \right] + O(\epsilon^2), \tag{4.20}$$

TABLE 4
The (leading order) values of $\gamma_{2,s}$ for various a and the corresponding mode l.

a	.5	.525	.55	.575	.6	.625	.65	.675	.7	.725
$\gamma_{2,s}$	1162	1067	1002	956	931	919	924	943	951	939
l	4	3	3	3	3	3	3	3	2	2

a	.75	.775	.8	.825	.85	.875	.9	.925	.95	.975
$\gamma_{2,s}$	952	999	1077	1199	1393	1711	2241	3254	5684	15454
l	2	2	2	2	2	2	2	2	2	2

where

$$(c_1, c_2, \dots, c_K) = \sum_{n=2}^{K} d_n \mathbf{e}_n. \tag{4.21}$$

When l is greater than 0, there are K critical eigenvalues for each l. They all have multiplicity $2l+1$. All of them are of order ϵ^2. If we write

$$\lambda_l = \epsilon^2 \mu_l + O(\epsilon^3), \tag{4.22}$$

then the μ_l's are found by solving the K-dimensional eigenvalue problem

$$\left[\frac{(l(l+1)-2)\tau}{(r_k^0)^2} + (-1)^k \gamma \mathcal{V}(r_k^0)\right] c_k + \gamma \sum_{j=1}^{K} G_l(r_k^0, r_j^0) c_j = \mu_l \tau c_k,$$
$$k = 1, 2, \dots, K, \tag{4.23}$$

where G_l is another Green's function:

$$G_l(r,s) = \begin{cases} (\frac{s^{1-l}}{2l+1} + \frac{(l+1)s^{2+l}}{l(2l+1)}) r^l & \text{if} \quad r < s, \\ s^{2+l}(\frac{r^{-1-l}}{2l+1} + \frac{(l+1)r^l}{l(2l+1)}) & \text{if} \quad r \geq s. \end{cases} \tag{4.24}$$

These critical eigenvalues are not always positive. There exists a stability threshold $\gamma_{K,s}$ so that when $\gamma_{K,e} < \gamma < \gamma_{K,s}$ all the critical eigenvalues are positive, and hence the K-interface solution u is stable, and when $\gamma > \gamma_{K,s}$ at least one critical eigenvalue is negative and the K-interface solution u is unstable. To each λ_l the corresponding eigenfunctions are $\phi_l(r) Y_{lm}$, $m = 0, \pm 1, \dots, \pm l$. ϕ_l is determined with the help of the eigenvectors c_k in (4.23):

$$\phi_l(r) = \sum_{j=1}^{K} c_j \left[H'\left(\frac{r - r_j}{\epsilon}\right) + \epsilon P_j'\left(\frac{r - r_j}{\epsilon}\right)\right] + O(\epsilon^2). \tag{4.25}$$

Table 4 reports the stability threshold values for various a. Note that the $\gamma_{2,s}$'s are greater than the corresponding $\gamma_{2,e}$. Hence there is a range $(\gamma_{2,e}, \gamma_{2,s})$ for γ where the 2-interface spherical lamellar pattern is stable.

5. Discussion. The single sphere pattern studied in section 3 gives only a limited picture of the spherical phase of a diblock copolymer, where multiple spheres coexist. Moreover, the spheres are observed to pack in the BCC pattern. An analytic study of such a multisphere pattern requires more refined singular perturbation techniques. The main difficulty is that the spheres in such a phase are only approximately round. The following argument illustrates this point.

It is known that even in a general domain, which we call Ω, (2.9) has a singular limit as $\epsilon \to 0$ [30]. The leading order outer expansion u^0 of u, a solution of (2.9), has the property that for a.e. $x \in \Omega$ $u^0(x) = 0$ or $u^0(x) = 1$ and $\overline{u^0} = a$. Let S be the union of the interfaces that separate the regions $u^0 = 0$ (B-rich microdomains) from the regions $u^0 = 1$ (A-rich microdomains), and $v^0 = (-\Delta)^{-1}(u^0 - a)$. In the singular limit an interface is a two-dimensional surface, with no thickness. At every $x \in S$,

$$\tau\kappa(x) + \gamma v^0(x) = \eta^0, \tag{5.1}$$

where $\kappa(x)$ is the mean curvature of S at x viewed from the $u^0 = 1$ side, and η^0 is a Lagrange multiplier to be determined. Equation (5.1) is a generalization of (4.2). The constraint (4.2) is replaced by $\overline{u^0} = a$. If the free boundary problem (5.1) admits an isolated stable solution u^0, then near u^0 there exists a local minimizer solution u of (2.9). However, (5.1) is still a challenging nonlocal geometric problem. Even though Figure 1 suggests that we look for solutions with multiple spheres, (5.1) implies that for such a solution the curvature of the interface of a sphere is in general not constant (there is the impact of v^0), i.e., the spheres are not exactly round, unless we deal with the one sphere or the spherically lamellar solutions in a ball as in this paper.

Nevertheless, if we consider the situation where a is close to 0 (or 1), then v^0 is near constant throughout Ω, and hence κ becomes close to a constant and the spheres are approximately round. The spherical phase in Figure 1 is thus heuristically explained. One must realize that in the small a case, i.e., small droplet/high curvature case, the parameter ϵ should be significantly less than a; otherwise we cannot have morphologies with microdomains separated by sharp interfaces. It was shown in [40] that the borderline range for a in one dimension is $a \sim \epsilon^{1/2}$. It is not clear at the moment what the borderline values for a are in two and three dimensions.

The stability threshold γ_s (or $\gamma_{K,s}$) is related to a strong segregation bifurcation phenomenon (not to be confused with the bifurcation analysis in the weak segregation regime). When γ passes γ_s (or $\gamma_{K,s}$) a second solution bifurcates out of the sphere (or spherical lamellar) solution. The new solution differs from the old one by a quantity which is roughly proportional to the eigenfunction of the 0 principal eigenvalue at $\gamma = \gamma_s$ (or $\gamma_{K,s}$). Because the eigenfunction has the form (3.13), the new solution has a wiggling interface (or interfaces). The wiggles are determined by the spherical harmonics Y_{lm} in (3.13); see Figure 8. Such a wiggling interface solution can be regarded as another defective pattern. If we consider the free energy during the bifurcation process, the bifurcating branch lowers the nonlocal part of the free energy by introducing more oscillation but increases the interface energy. The overall free energy of the bifurcating branch is lower than that of the first branch.

We did not discuss the dynamics of a diblock copolymer system. The purpose of studying the critical eigenvalues of a solution of (2.9) in this paper is to determine whether the solution is a local minimizer of (2.2). However, the same critical eigenvalues also determine the local dynamics, near the solution, of the evolution equation

$$u_t = \epsilon^2\Delta u - f(u) - \epsilon\gamma(-\Delta)^{-1}(u - a) + \overline{f(u)}, \quad \partial_\nu u = 0 \text{ on } \partial D. \tag{5.2}$$

Note that $\int_D u(x)\,dx$ is conserved under (5.2) because $\frac{d}{dt}\int_D u(x)\,dx = 0$ after one integrates (5.2) over D. The eigenfunctions of the critical eigenvalues give the directions along which the dynamics of (5.2) runs slowly (the eigenfunctions of the noncritical eigenvalues are directions of fast dynamics).

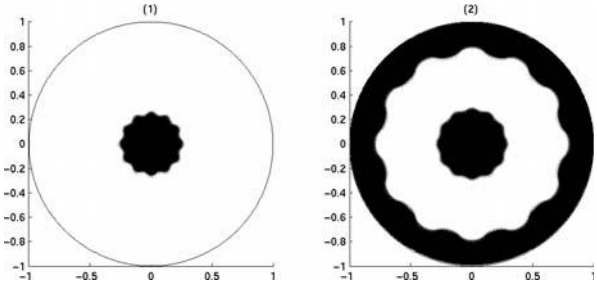


FIG. 8. *A cross section of a wiggling sphere solution and a cross section of a wriggled spherical lamellar solution with two interfaces.*

The critical eigenvalues in this context admit geometric interpretations. For the sphere pattern the eigenfunction of the critical eigenvalue λ_0 in (3.14) is radially symmetric. So the dynamics in this direction involves only the shape change of u in the radial direction. The eigenfunctions of the critical eigenvalue λ_1 in (3.16) have the forms $(x/r)\phi_1(r)$, $(y/r)\phi_1(r)$, and $(z/r)\phi_1(r)$. They lead to translations of u in x-, y-, or z-directions in the dynamics. Finally, the eigenfunctions of the critical eigenvalues λ_l, $l > 1$, in (3.18) give rise to oscillations of the interfaces. The same interpretations are also valid for the spherical lamellar pattern. Because all these eigenfunctions concentrate at the interface r_j by (3.15), (3.17), (4.12), (4.20), and (4.25), the dynamics along the directions of the critical eigenvalues is seen as the motion of the interfaces.

However, (5.2) is only one dynamical law that we can associate with (2.2). A more realistic one is the fourth-order partial differential equation [2]

$$u_t = \Delta(-\epsilon^2\Delta + f(u)) - \epsilon\gamma(u-a), \quad \partial_\nu u = \partial_\nu \Delta u = 0 \text{ on } \partial D. \tag{5.3}$$

Equation (5.3) generalizes the well-known Cahn–Hilliard equation [5]. Based on a spectral comparison argument [4], one shows that a steady state is stable under (5.2) if and only if it is stable under (5.3) [27]. Hence our results on the stability of the various steady states in this paper remain valid in the dynamics (5.3). An even more complex dynamical law considers a diblock copolymer melt as a fluid. It adds the velocity field and couples (5.3) with the Navier–Stokes equation of the velocity field [2].

We are mainly interested in stable solutions of (2.9). They are local minimizers of (2.2). We do not know whether or not any of the solutions found in this paper is a global minimizer. There also exist unstable spherical lamellar solutions even for $\gamma \in (\gamma_{K,e}, \gamma_{K,s})$. In Figures 6 and 7, when $\gamma > \gamma_{2,e}$, in addition to the minimum of J there exists a local maximum of J. This maximum point corresponds to an unstable spherical lamellar solution. This unstable solution exists for $\gamma \geq \gamma_{2,s}$ as well. The instability of this solution is caused by the $m = 0$ mode.

In the functional (2.2) the key ingredient is the nonlocal term. It describes a long range interaction. Many other important physical systems that exhibit self-organization and pattern formation share the same phenomenon [42]. Examples include charged Langmuir monolayers [1] and smectic liquid crystal films [41]. Many of the singular perturbation techniques presented here may be applied to these problems [29, 34, 37, 36].

The nonlocal interaction in (2.2) is of Coulomb type [25]. Some of the above-mentioned problems have different nonlocal interactions. In the charged monolayer

problem the nonlocal term is written as

$$\int_D \int_D (u(x) - a) G_c(x, y)(u(y) - a)\, dxdy, \tag{5.4}$$

which assumes the same form as (2.8). However, the kernel G_c is different. If D is a square, i.e., $(0, 1)^2$, with the periodic boundary condition, then G_c is translation invariant so that $G_c(x, y) = G_c(x - y)$. The Fourier series of G_c is

$$\hat{G}_c(\xi) = \frac{1}{|\xi|}. \tag{5.5}$$

Note that for the diblock copolymer problem the corresponding G on a square is

$$\hat{G}(\xi) = \frac{1}{|\xi|^2}. \tag{5.6}$$

Hence as $|\xi| \to \infty$, (5.6) has a faster decay rate than (5.5). Many properties, such as the optimal size discussed in section 3.2, are sensitive to these decay rates. In general with a slower decay rate, one finds smaller microdomains [34].

In the smectic liquid crystal film problem, the nonlocal interaction comes from a coupling effect with the director field. In this case, because of the unit length constraint on the director field, the nonlocal interaction is no longer quadratic [36].

6. Conclusion. We used asymptotic analysis to study the Ohta–Kawasaki density functional theory for diblock copolymers. We constructed a single sphere pattern in a unit ball. Such a pattern is a cell in the spherical morphology. We showed the existence of the sphere pattern as a solution of the Euler–Lagrange equation of the free energy. We identified the optimal size of such a cell with the least free energy. We also found a stability threshold. The sphere is stable if it is less than the threshold value and unstable if it is greater than the threshold value. The stability threshold value is greater than the optimal size. At the stability threshold, there is another solution, a bifurcating branch. It has an interface of a wriggled sphere. This solution has lower free energy than that of the first solution.

Next we studied a spherical lamellar pattern, which we view as a defective lamellar pattern. Singular perturbation analysis allowed us to reduce the existence and stability problems in infinite dimensions to existence and matrix problems in finite dimensions. We found two thresholds: an existence threshold and a larger, stability threshold. There is a spherical lamellar pattern only when the size of the sample is larger than the existence threshold value. This patten is stable only when the size of the sample is between the existence threshold and the stability threshold. At the stability threshold, there is a bifurcating branch with a pattern of wriggled spherical interfaces. The bifurcating branch again has lower free energy.

REFERENCES

[1] D. Andelman, F. Broçhard, and J.-F. Joanny, *Phase transitions in Langmuir monolayers of polar molecules*, J. Chem. Phys., 86 (1987), pp. 3673–3681.

[2] M. Bahiana and Y. Oono, *Cell dynamical system approach to block copolymers*, Phys. Rev. A, 41 (1990), pp. 6763–6771.

[3] F. S. Bates and G. H. Fredrickson, *Block copolymers–designer soft materials*, Phys. Today, 52 (1999), pp. 32–38.

[4] P. W. Bates and P. C. Fife, *Spectral comparison principles for the Cahn–Hilliard and phase-field equations, and times scales for coarsening*, Phys. D, 43 (1990), pp. 335–348.

[5] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system.* I. *Interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–267.
[6] R. Choksi, *Scaling laws in microphase separation of diblock copolymers*, J. Nonlinear Sci., 11 (2001), pp. 223–236.
[7] R. Choksi and X. Ren, *On the derivation of a density functional theory for microphase separation of diblock copolymers*, J. Statist. Phys., 113 (2003), pp. 151–176.
[8] E. De Giorgi, *Sulla convergenza di alcune successioni d'integrali del tipo dell'area*, Rend. Mat. (6), 8 (1975), pp. 277–294.
[9] P. C. Fife and D. Hilhorst, *The Nishiura–Ohnishi free boundary problem in the* 1*D case*, SIAM J. Math. Anal., 33 (2001), pp. 589–606.
[10] I. W. Hamley, *The Physics of Block Copolymers*, Oxford University Press, Oxford, UK, 1998.
[11] E. Helfand, *Theory of inhomogeneous polymers: Fundamentals of Gaussian random-walk model*, J. Chem. Phys., 62 (1975), pp. 999–1005.
[12] E. Helfand and Y. Tagami, *Theory of the interface between immiscible polymers* II, J. Chem. Phys., 56 (1972), pp. 3592–3601.
[13] E. Helfand and Z. R. Wasserman, *Block copolymer theory.* 4. *Narrow interphase approximation*, Macromolecules, 9 (1976), pp. 879–888.
[14] E. Helfand and Z. R. Wasserman, *Block copolymer theory.* 5. *Spherical domains*, Macromolecules, 11 (1978), pp. 960–966.
[15] E. Helfand and Z. R. Wasserman, *Block copolymer theory.* 6. *Cylindrical domains*, Macromolecules, 13 (1980), 994–998.
[16] M. Henry, *Singular limit of a fourth order problem arising in the microphase separation of diblock copolymers*, Adv. Differential Equations, 6 (2001), pp. 1049–1114.
[17] K. M. Hong and J. Noolandi, *Theory of inhomogeneous multicomponent polymer systems*, Macromolecules, 14 (1981), pp. 727–736.
[18] K. M. Hong and J. Noolandi, *Theory of phase equilibria in systems containing block copolymers*, Macromolecules, 16 (1983), pp. 1083–1093.
[19] R. Kohn and P. Sternberg, *Local minimisers and singular perturbations*, Proc. Roy. Soc. Edinburgh Sect. A, 111 (1989), pp. 69–84.
[20] L. Leibler, *Theory of microphase separation in block copolymers*, Macromolecules, 13 (1980), pp. 1602–1617.
[21] M. W. Matsen and F. S. Bates, *Unifying weak- and strong-segregation block copolymer theories*, Macromolecules, 29 (1996), pp. 1091–1098.
[22] M. W. Matsen and M. Schick, *Stable and unstable phases of a diblock copolymer melt*, Phys. Rev. Lett., 72 (1994), pp. 2660–2663.
[23] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Rational Mech. Anal., 98 (1987), pp. 123–142.
[24] L. Modica and S. Mortola, *Un esempio di* Γ^-*-convergenza*, Boll. Un. Mat. Ital. B (5), 14 (1977), pp. 285–299.
[25] C. B. Muratov, *Theory of domain patterns in systems with long-range interactions of Coulomb type*, Phys. Rev. E, 66 (2002), 066108.
[26] Y. Nishiura and I. Ohnishi, *Some mathematical aspects of the microphase separation in diblock copolymers*, Phys. D, 84 (1995), pp. 31–39.
[27] I. Ohnishi and Y. Nishiura, *Spectral comparison between the second and fourth order equations of conservative type with nonlocal terms*, Japan J. Indust. Appl. Math., 15 (1998), pp. 253–262.
[28] T. Ohta and K. Kawasaki, *Equilibrium morphology of block copolymer melts*, Macromolecules, 19 (1986), pp. 2621–2632.
[29] X. Ren and L. Truskinovsky, *Finite scale microstructures in nonlocal elasticity. In recognition of the* 60*th birthday of Roger L. Fosdick (Blacksburg, VA,* 1999*)*, J. Elasticity, 59 (2000), pp. 319–355.
[30] X. Ren and J. Wei, *On the multiplicity of solutions of two nonlocal variational problems*, SIAM J. Math. Anal., 31 (2000), pp. 909–924.
[31] X. Ren and J. Wei, *Concentrically layered energy equilibria of the diblock copolymer problem*, European J. Appl. Math., 13 (2002), pp. 479–496.
[32] X. Ren and J. Wei, *On energy minimizers of the diblock copolymer problem*, Interfaces Free Bound., 5 (2003), pp. 193–238.
[33] X. Ren and J. Wei, *On the spectra of three-dimensional lamellar solutions of the diblock copolymer problem*, SIAM J. Math. Anal., 35 (2003), pp. 1–32.
[34] X. Ren and J. Wei, *Soliton-stripe patterns in charged Langmuir monolayers*, J. Nonlinear Sci., 13 (2003), pp. 603–624.

[35] X. Ren and J. Wei, *Triblock copolymer theory: Ordered ABC lamellar phase*, J. Nonlinear Sci., 13 (2003), pp. 175–208.
[36] X. Ren and J. Wei, *Chiral symmetry breaking and the soliton-stripe pattern in Langmuir monolayers and smectic films*, Nonlinearity, 17 (2004), pp. 617–632.
[37] X. Ren and J. Wei, *The soliton-stripe pattern in the Seul–Andelman membrane*, Phys. D, 188 (2004), pp. 277–291.
[38] X. Ren and J. Wei, *Stability of spot and ring solutions of the diblock copolymer equation*, J. Math. Phys., 45 (2004), pp. 4106–4133.
[39] X. Ren and J. Wei, *Wriggled lamellar solutions and their stability in the diblock copolymer problem*, SIAM J. Math. Anal., 37 (2005), pp. 455–489.
[40] X. Ren and J. Wei, *Droplet solutions in the diblock copolymer problem with skewed monomer composition*, Calc. Var. Partial Differential Equations, 25 (2006), pp. 333–359.
[41] J. V. Selinger, Z.-G. Wang, R. F. Bruinsma, and C. M. Knobler, *Chiral symmetry breaking in Langmuir monolayers and smectic films*, Phys. Rev. Lett., 70 (1993), pp. 1139–1142.
[42] M. Seul and D. Andelman, *Domain shapes and patterns: The phenomenology of modulated phases*, Science, 267 (1995), pp. 476–483.
[43] T. Teramoto and Y. Nishiura, *Double gyroid morphology in a gradient system with nonlocal effects*, J. Phys. Soc. Japan, 71 (2002), pp. 1611–1614.
[44] Y. Tsori, D. Andelman, and M. Schick, *Defects in lamellar diblock copolymers: Chevron- and Ω-shaped boundaries*, Phys. Rev. E, 61 (2000), pp. 2848–2858.

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1101–1129

A STAGE-STRUCTURED PREDATOR-PREY MODEL OF BEDDINGTON–DEANGELIS TYPE*

SHENGQIANG LIU† AND EDOARDO BERETTA‡

Abstract. We formulate and study a robust stage structured predator-prey model of Beddington–DeAngelis-type functional response. The time delay is the time taken from birth to maturity. The Beddington–DeAngelis functional response is similar to the Holling type 2 functional response but contains an extra term describing mutual interference by predators. First we show that the predator coexists with prey permanently if and only if the predator's recruitment rate at the peak of prey abundance is larger than its death rate. Second, we show that if the system is permanent, then a sufficiently large degree of the predator interference can not only stabilize the system but also guarantee the stability of the system against the increase of the carrying capacity of prey and the increase of birth rate of the adult predator. Third, we show both analytically and numerically that stability switches of interior equilibrium may occur as maturation time delay increases: stability may change from stable to unstable to finally stable, implying that a large delay can be stabilizing.

Key words. delay, predator-prey, stage structure, Beddington–DeAngelis

AMS subject classifications. 34C11, 92D25

DOI. 10.1137/050630003

1. Introduction.

The goal in this paper is to study a stage structured predator-prey model with Beddington–DeAngelis-type functional response. It is a central goal for ecologists to understand the relationship between predator and prey, and one significant component of the predator-prey relationship is the predator's functional responses or so-called predator's rate of feeding upon prey in other references [39], i.e., the rate of prey consumption by an average predator.

As for the mathematical predator-prey models, the description of a predator's instantaneous, per capita feeding rate, f, as a function of prey abundance x, is the classic definition of a predator's functional responses. There have been several famous functional response type: Holling types I–III [19], [20]; Hassell–Varley type [17]; Beddington–DeAngelis type by Beddington [6] and DeAngelis, Goldstein, and Neill [13] independently; the Crowley–Martin type [12]; and the recent well-known ratio-dependence type by Arditi and Ginzburg [3] later studied by Kuang and Beretta [27]. Of them, the Holling type I–III was labeled "prey-dependent" and the other types that consider the interference among predators were labeled "predator-dependent" by Arditi and Ginzburg [3]. Recently, "predator-dependent" type models have received much support from theoretical and empirical work in biology (see [4], [5], [11], [37], [38], [39], and the references therein). In [5], Abrams and Ginzburg even pointed out that "precise prey dependence and ratio dependence will both be rare" while "predator dependence will be common." In [39], by comparing the statistical evidence from

*Received by the editors April 26, 2005; accepted for publication (in revised form) November 16, 2005; published electronically March 17, 2006.
http://www.siam.org/journals/siap/66-4/63000.html

†Department of Mathematics, Xiamen University, Xiamen, 361005, China (sqliu@xmu.edu.cn). The work of this author was partially performed while he was a visiting professor at Istituto di Biomatematica of Universitá di Urbino. The work of this author was also supported by National Natural Science Foundation of China grant 10371127.

‡Istituto di Biomatematica, Universitá di Urbino, I-61029, Urbino, Italy (e.beretta@mat. uniurb. it). The work of this author was supported by Italian research project FIRB RBAUO1K7M2_001.

19 predator-prey systems with three predator-dependent functional responses, Skalski and Gilliam pointed out that the predator-dependent functional responses can provide better descriptions of predator feeding over a range of predator-prey abundances, and in some cases, the Beddington–DeAngelis-type functional response (hereafter the BD model) performed even better. Although the predator-dependent models that they considered fit those data reasonably well, no single functional response best describes all the data sets. The Beddington–DeAngelis response can be generated by a number of natural mechanisms [6], [11], [38] and because it admits rich but biologically reasonable dynamics [9], it is worthy for us to further study the BD model.

The per capita feeding rate of BD model takes the form [6]

$$f(x,y)=\frac{bx}{1+k_1x+k_2y}, \tag{1.1}$$

and thus the BD model takes the form

$$\begin{cases} x'(t)=rx(t)\left(1-\dfrac{x(t)}{K}\right)-\dfrac{bx(t)y(t)}{1+k_1x(t)+k_2y(t)},\\ y'(t)=\dfrac{nbx(t)y(t)}{1+k_1x(t)+k_2y(t)}-dy(t), \end{cases} \tag{1.2}$$

where x and y represent prey and predator densities; b (units: 1/time) and k_1 (units: 1/prey) are positive constants that describe the effects of capture rate and handling time, respectively, on the feeding rate; n is the birth rate of the predator; and $k_2 \geq 0$ (units: 1/predator) is a constant describing the magnitude of interference among predators [13]. The BD model is similar to the well-known Holling II type functional response (hereafter the H2 model) but has an extra term k_2y in the denominator modeling mutual interference among predators. Hence this kind of type functional response given in (1.1) is affected by both predator and prey, i.e., the so-called predator dependence by Arditi and Ginzburg [3]. Dynamics for the H2 model have been much studied ([22], [25], and references therein). Then how the mutual interference term affects the dynamic of the whole system is an interesting problem.

Many recent works have contributed to the BD model (1.2) [9], [10], [14], [23], [24], [33], [36]. Cantrell and Cosner [9] considered (1.2) and obtained the criteria for permanence, extinction, global stability of the interior equilibrium, and existence of periodic orbits. They showed that k_2 does affect the location and stability of the equilibrium for (1.2): (1) Adequate increase of k_2 may change the positive equilibrium from unstable to stable. (2) Having $k_2 > 0$ can stabilize the system by reducing the extent to which trajectories can exhibit "boom-bust" behavior. Therefore, the effect of k_2 in system (1.2) is to introduce a self-limiting term into the predator equation.

Hwang [23] showed that the interior equilibrium of system (1.2) is globally stable provided it is locally asymptotically stable. In [24], Hwang obtained the sufficient conditions for the uniqueness of limit cycles of (1.2).

Liu and Yuan [33] considered time delay τ in the response term $f(x,y)$ of (1.1) in the predator equation, that is,

$$\begin{cases} x'(t)=rx(t)\left(1-\dfrac{x(t)}{K}\right)-\dfrac{bx(t)y(t)}{1+k_1x(t)+k_2y(t)},\\ y'(t)=y(t)\left[-d+\dfrac{nbx(t-\tau)}{1+k_1x(t-\tau)+k_2y(t-\tau)}\right]. \end{cases} \tag{1.3}$$

In the second equation of (1.3), there is a nondelay term $y(t)$, which is similar to the delay models studied by Kuang [26] and Beretta and Kuang [7], where τ can be regarded as a gestation period or reaction time of the predations (see Martin and Ruan, [34]). By choosing the delay τ as a parameter, Liu and Yuan [33] showed that as τ crosses some critical values, Hopf bifurcation about the stability of interior equilibrium in (1.3) can occur.

Although much progress has been seen in the above work on BD model, such models are not well studied yet in the sense that all the known results are for models that ignore the enormous diversity during the life histories of the predator. Unfortunately, this is not realistic due to the following reasons:

1. Juvenile predators have a time lag from their birth to maturity.
2. Young predators are raised by their parents or are dependent on the nutrition from the eggs they stay in and they are much weaker than the mature predators, hence the juvenile predators cannot hunt the prey, nor can they breed.
3. Young predators reach maturity after surviving the immature stage; if the juvenile death rate (through-stage death rate) is nonzero, then not all immature predators can survive the juvenile stage.

Therefore, it is realistic and interesting for us to construct the stage-structured predator-prey model and study the combined effects of stage structure and mutual interference by predators. Most existing stage structure models (see [1], [28], [29], [30], [35], [2], and the references therein) deal with single species growth that assume a constant resource supply [15]. Recently, Gourley and Kuang [15] formulated a robust stage-structured predator-prey model with the assumption that stage-structured consumer species growth is a combined result of birth and death processes, both of which are closely linked to the dynamical supply of resource. Enlightened by the modeling methods in [15], we formulate the robust stage-structured BD model as follows:

$$
\begin{cases}
x'(t) = rx(t)\left(1 - \dfrac{x(t)}{K}\right) - \dfrac{bx(t)y(t)}{1 + k_1x(t) + k_2y(t)}, \\
y'(t) = \dfrac{nbe^{-d_j\tau}x(t-\tau)y(t-\tau)}{1 + k_1x(t-\tau) + k_2y(t-\tau)} - dy(t), \\
y_j'(t) = \dfrac{nbx(t)y(t)}{1 + k_1x(t) + k_2y(t)} - \dfrac{nbe^{-d_j\tau}x(t-\tau)y(t-\tau)}{1 + k_1x(t-\tau) + k_2y(t-\tau)} - d_jy_j(t), \\
x(\theta),\ y(\theta) \geq 0 \text{ are continuous on } -\tau \leq \theta \leq 0, \text{ and } x(0), y(0), y_j(0) > 0,
\end{cases} \tag{1.4}
$$

where $x(t)$ and $y(t)$ represent prey and the mature predator densities, respectively, and $y_j(t)$ denotes the immature or juvenile predator densities. We assume that juveniles suffer a mortality rate of d_j(the through-stage death rate) and take τ units of time to mature; thus $e^{-d_j\tau}$ is the surviving rate of each immature predator to reach maturity. And for the continuity of the solutions to system (1.4), in this paper, we require

$$
y_j(0) = bn\int_{-\tau}^{0} \frac{e^{d_j s}x(s)y(s)}{1 + k_1x(s) + k_2y(s)}\mathrm{d}s. \tag{1.5}
$$

By the third equation of system (1.4), the initial conditions (1.5), and using the arguments similar to Lemma 3.1 in [31, p. 672], we have

$$
y_j(t) = bn\int_{-\tau}^{0} \frac{e^{d_j s}x(t+s)y(t+s)}{1 + k_1x(t+s) + k_2y(t+s)}\mathrm{d}s, \tag{1.6}
$$

i.e., $y_j(t)$ is completely determined by $x(t), y(t)$, and thus the following system can be separated from system (1.4):

$$\begin{cases} x'(t) = rx(t)\left(1 - \dfrac{x(t)}{K}\right) - \dfrac{bx(t)y(t)}{1 + k_1x(t) + k_2y(t)}, \\ y'(t) = \dfrac{nbe^{-d_j\tau}x(t-\tau)y(t-\tau)}{1 + k_1x(t-\tau) + k_2y(t-\tau)} - dy(t), \\ x(\theta),\ y(\theta) \geq 0 \text{ is continuous on } -\tau \leq \theta \leq 0, \text{ and } x(0), y(0) > 0. \end{cases} \tag{1.7}$$

In the present paper, our main purpose is to study the global dynamics of our system (1.4) and to consider how the stage structure parameters $d_j\tau$ and the predator interference parameter k_2 affect the dynamical behaviors of system (1.4).

This paper is organized as follows. In the next section, we consider the equilibria for system (1.4) and give the conditions for the existence of the positive equilibrium. In section 3, we obtain the necessary and sufficient conditions for the extinction of predator and for the permanence of system (1.4). This is followed by a section on the global attractiveness of the positive equilibrium. The local stability of the equilibria of (1.4) is considered in section 5. The analysis of positive equilibrium is highly nontrivial and we provide only generic conditions for its stability switch, but we manage to give the sufficient conditions for the asymptotic stability of the positive equilibrium which require large k_2. To complement this analytic work, we also present some carefully designed simulation results. The paper ends with a discussion.

2. Equilibria. Because of (1.6), we only consider the equilibria (x, y) of system (1.7), which are solutions of the system

$$\begin{cases} rx\left(1 - \dfrac{x}{K}\right) - \dfrac{bxy}{1 + k_1x + k_2y} = 0, \\ \dfrac{nbe^{-d_j\tau}xy}{1 + k_1x + k_2y} - dy = 0. \end{cases} \tag{2.1}$$

It is easy to see that for all parameter values (1.7) has the equilibria $E_0 = (0, 0),\ E_1 = (K, 0)$. By (2.1), system (1.7) has the positive equilibrium $E = (x^*, y^*)$ iff

$$\frac{nbe^{-d_j\tau}K}{1 + k_1K} > d, \tag{2.2}$$

where

$$x^* = \frac{1}{2}\left(-B + \sqrt{B^2 + 4C}\right), \quad y^* = \frac{x^*(nbe^{-d_j\tau} - dk_1) - d}{dk_2} \tag{2.3}$$

with

$$B = \frac{K}{r}\left(\frac{nbe^{-d_j\tau} - dk_1}{ne^{-d_j\tau}k_2} - r\right); \quad C = \frac{Kd}{rne^{-d_j\tau}k_2}.$$

Hence the positive equilibrium E exists for all predation maturation times τ in the interval $I = [0, \tau^*)$, where

$$\tau^* = \frac{1}{d_j}\log\frac{Knb}{d(1 + Kk_1)}. \tag{2.4}$$

Increase of τ in I will lower y^* until E will coincide with E_1 at the finite value τ^*, and for higher τ there is no positive equilibrium.

On the other hand, k_2 does not affect the existence of the positive equilibrium since k_2 is not involved in (2.2). However, (2.3) indicates that increase of k_2 will lower y^* until E will coincide with E_1 at the infinite value k_2.

3. Permanence and extinction. Here the following results give conditions which are both necessary and sufficient for the permanence, extinction, of system (1.4), respectively.

THEOREM 3.1. $\lim_{t\to\infty}(x(t), y(t), y_j(t)) = (K, 0, 0)$ *holds true iff* $\frac{nbe^{-d_j\tau}K}{1+k_1K} \le d$ *holds true.*

THEOREM 3.2. *System* (1.4) *is permanent iff it satisfies* (2.2).

Theorem 3.1 and Theorem 3.2 directly generalize Theorem 3.1 in [9] into the stage-structured case.

Mathematically, Theorem 3.2 suggests that the permanence of (1.4) is equivalent to its existence of positive equilibrium. Biologically, Theorem 3.2 shows that the predator coexists with prey permanently iff the predator's recruitment rate at the peak of prey abundance is more than its death rate.

Let $k_2 = 0$, i.e., $f(x, y)$ in (1.1) becomes the H2 functional response. By Theorem 3.1 and Theorem 3.2, we directly have the following corollary.

COROLLARY 3.3. *Given system* (1.4) *with* $k_2 = 0$, *then* $\lim_{t\to\infty}(x(t), y(t), y_j(t)) = (K, 0, 0)$ *holds true iff* $\frac{nbe^{-d_j\tau}K}{1+k_1K} \le d$ *holds true.*

COROLLARY 3.4. *Given system* (1.4) *with* $k_2 = 0$, *then it is permanent iff it satisfies* $\frac{nbe^{-d_j\tau}K}{1+k_1K} > d$.

Let $k_1 = k_2 = 0$, i.e., $f(x, y)$ becomes the Holling type I functional response. By Theorem 3.1 and Theorem 3.2, we directly have the next corollary.

COROLLARY 3.5. *Given system* (1.4) *with* $k_1 = k_2 = 0$, *then* $\lim_{t\to\infty}(x(t), y(t), y_j(t)) = (K, 0, 0)$ *holds true iff* $nbe^{-d_j\tau}K \le d$ *holds true.*

COROLLARY 3.6. *Given system* (1.4) *with* $k_1 = k_2 = 0$, *then it is permanent iff it satisfies* $nbe^{-d_j\tau}K > d$.

For stage-structured predator-prey system with functional response of Holling I–II, Gourley and Kuang [15, Theorem 3.1] give the necessary and sufficient conditions on extinction of predator, which is included in our Corollaries 3.3 and 3.5. Further, Corollary 3.4 suggests that the stage-structured H2 and BD systems share the same permanence and extinction conditions. Hence, mutual interference coefficient k_2 does not affect the permanence of system (1.4), which is also an extension of the corresponding conclusions in [9].

To prove the above main results, we need some preliminary results. By arguments similar to Lemma 1 in [28], we have the following.

LEMMA 3.7. *Suppose* $y(\theta) \ge 0$ *is continuous on* $-\tau \le \theta \le 0$, *and* $x(0), y(0), y_j(0) > 0$. *Then the solution of system* (1.4) *satisfies* $x(t), y(t), y_j(t) > 0$ *for all* $t > 0$.

LEMMA 3.8. *Permanence of* $x(t), y(t)$ *in system* (1.4) *implies that of* $y_j(t)$.

Proof. Since $x(t), y(t)$ have positive ultimately upper and lower boundaries, using (1.6) we get

$$0 < \underline{\lim_{t\to\infty}} y_j(t) \le \overline{\lim_{t\to\infty}} y_j(t) < \infty,$$

proving Lemma 3.8. □

LEMMA 3.9. *System* (1.4) *is always dissipative in the first quadrant.*

Proof. By the first equation of system (1.4), $\dot{x}(t) < rx(t)(1 - \frac{x(t)}{K})$; thus we have

$$\overline{\lim_{t\to\infty}} x(t) \le K. \tag{3.1}$$

Let $W(t) = ne^{-d_j\tau}x(t) + y(t+\tau)$, then we have

$$\begin{aligned}\dot{W}(t)|_{(1.4)} &= -dy(t+\tau) + ne^{-d_j\tau}rx(t)\left(1 - \frac{x(t)}{K}\right)\\ &= -dW(t) + nde^{-d_j\tau}x(t) + ne^{-d_j\tau}rx(t)\left(1 - \frac{x(t)}{K}\right)\end{aligned}$$

By (3.1), there exist some positive constant B, T, such that $\dot{W}(t)|_{(1.4)} \le B - dW(t)$ for all $t \ge T$. Thus $\overline{\lim_{t\to\infty}} W(t) \le B/d$, and consequently $x(t), y(t)$ are ultimately bounded; using (1.6) we also have that $y_j(t)$ is ultimately bounded, proving Lemma 3.9. □

Proof of Theorem 3.1. For the sufficiency of the theorem, we consider two cases.

Case 1. $\frac{nbe^{-d_j\tau}K}{1+k_1K} < d$.

By Lemma 3.9, for the sufficiently small positive constant ϵ with $\frac{nbe^{-d_j\tau}(K+\epsilon)}{1+k_1(K+\epsilon)} < d$, there exists a $T = T_\epsilon > 0$ such that $x(t) < K + \epsilon$ for all $t > T$; substitute it into the second equation of (1.4), we get that for all $t > T + \tau$, there is

$$y'(t) < \frac{nbe^{-d_j\tau}(K+\epsilon)y(t-\tau)}{1 + k_1(K+\epsilon) + k_2y(t-\tau)} - dy(t) < \frac{nbe^{-d_j\tau}(K+\epsilon)y(t-\tau)}{1 + k_1(K+\epsilon)} - dy(t).$$

Since $\frac{nbe^{-d_j\tau}(K+\epsilon)}{1+k_1(K+\epsilon)} < d$, Lemma 2 of Liu et al. [28, p. 131] implies that the solution for the comparison equation

$$u'(t) = \frac{nbe^{-d_j\tau}(K+\epsilon)u(t-\tau)}{1 + k_1(K+\epsilon)} - du(t)$$

satisfies $\lim_{t\to\infty} u(t) = 0$, which with Lemma 3.7 proves $\lim_{t\to\infty} y(t) = 0$. Therefore by the third equation of (1.4) and the arguments by Liu et al. [28, p. 128], we get that $\lim_{t\to\infty} y(t) = 0$ implies $\lim_{t\to\infty} y_j(t) = 0$, hence by the first equation of system (1.4), we get $\lim_{t\to\infty} x(t) = K$, proving Case 1.

Case 2. $\frac{nbe^{-d_j\tau}K}{1+k_1K} = d$.

By the first equation of system (1.4), $x(t)$ is always decreasing when above K. We can prove that if there exists some $t_0 > 0$ such that $x(t_0) < K$, then $x(t) < K$ for all $t > t_0$. Otherwise there must exist some $t_1 > t_0$ such that $x(t_1) = K$ and $x'(t_1) \ge 0$. This is impossible. Hence, there are two possible cases, either

(1) $x(t) > K$ and $x(t) \to K$ as $t \to \infty$, or

(2) there exists some $t_0 > 0$ such that $x(t_0) < K$.

For the first of these cases, we only need to prove that $\lim_{t\to\infty} y(t) = 0$, since this implies $\lim_{t\to\infty} y_j(t) = 0$. Integrating the equation for $x(t)$ in (1.4), we have

$$\begin{aligned}x(t) - x(0) &= \int_0^t rx(s)\left(1 - \frac{x(s)}{K}\right)\mathrm{d}s - \int_0^t \frac{bx(s)y(s)}{1 + k_1x(s) + k_2y(s)}\mathrm{d}s\\ &< \int_0^t \underbrace{rx(s)\left(1 - \frac{x(s)}{K}\right)}_{x(s)\ge K}\mathrm{d}s - \int_0^t \frac{bKy(s)}{1 + k_1K + k_2y(s)}\mathrm{d}s\end{aligned}$$

for all $t \geq t_0$, and then

$$\int_0^t \frac{bKy(s)}{1+k_1K+k_2y(s)}\mathrm{d}s < x(0) - x(t) + \int_0^t \underbrace{rx(s)\left(1-\frac{x(s)}{K}\right)}_{\leq 0}\mathrm{d}s < x(0).$$

By the boundedness of $y(t)$, then $\int_0^t y(s)\mathrm{d}s$ is bounded for all $t \geq t_0$, and this implies $\lim_{t\to\infty} y(t) = 0$.

For the second of these cases, consider the function

$$V = y(t) + d\int_{t-\tau}^t y(s)\mathrm{ds}.$$

Then for all $t \geq t_0 + \tau$, we have

$$\begin{aligned}
\frac{dV}{dt}|_{(1.4)} &= \frac{nbe^{-d_j\tau}x(t-\tau)y(t-\tau)}{1+k_1x(t-\tau)+k_2y(t-\tau)} - dy(t) + d(y(t)-y(t-\tau)) \\
&= y(t-\tau)\cdot\left(\underbrace{\frac{nbe^{-d_j\tau}x(t-\tau)}{1+k_1x(t-\tau)+k_2y(t-\tau)}}_{x(t-\tau)<K} - d\right) \\
&< y(t-\tau)\cdot\left(\frac{nbe^{-d_j\tau}K}{1+k_1K+k_2y(t-\tau)} - d\right) \\
&= -\frac{dk_2y^2(t-\tau)}{1+k_1K+k_2y(t-\tau)} < 0,
\end{aligned}$$

which with Lemma 3.7 proves $\lim_{t\to\infty} y(t) = 0$. This proves $\frac{nbe^{-d_j\tau}K}{1+k_1K} \leq d$ is the sufficient condition for $\lim_{t\to\infty}(x(t), y(t), y_j(t)) = (K, 0, 0)$.

Now, we prove $\lim_{t\to\infty}(x(t), y(t), y_j(t)) = (K, 0, 0) \Longrightarrow \frac{nbe^{-d_j\tau}K}{1+k_1K} \leq d$. Assume the contrary, i.e., $\frac{nbe^{-d_j\tau}K}{1+k_1K} > d$; then system (1.4) has a positive equilibrium (x^*, y^*, y_j^*), contradicting $\lim_{t\to\infty}(x(t), y(t), y_j(t)) = (K, 0, 0)$ for all solution $(x(t), y(t), y_j(t))$. Hence there must be $\frac{nbe^{-d_j\tau}K}{1+k_1K} \leq d$, and this proves Theorem 3.1. □

To prove Theorem 3.2, we engage the persistence theory by Hale and Waltmann [16] for infinite dimensional systems; we also refer to Thieme [40]. Now, we present the persistence theory [16] as follows.

Consider a metric space X with metric d. T is a continuous semiflow on X, i.e., a continuous mapping $T:\ [0,\infty)\times X \to X$ with the following properties:

$$T_t \circ T_s = T_{t+s}, \quad t, s \geq 0; \quad T_0(x) = x,\ x \in X.$$

Here T_t denotes the mapping from X to X given by $T_t(x) = T(t, x)$. The distance $d(x, y)$ of a point $x \in X$ from a subset Y of X is defined by

$$d(x, y) = \inf_{y\in Y} d(x, y).$$

Recall that the positive orbit $\gamma^+(x)$ through x is defined as $\gamma^+(x) = \cup_{t\geq 0}\{T(t)x\}$, and its $\omega-$limit set is $\omega(x) = \cap_{\tau\geq 0} CL\cup_{t\geq\tau}\{T(t)x\}$, where CL means closure. Define $W^s(A)$ the stable set of a compact invariant set A as

$$W^s(A) = \{x:\ x \in X,\ \omega(x) \neq \phi,\ \omega(x) \subset A\};$$

define $\widetilde{A_\partial}$ the particular invariant sets of interest as

$$\widetilde{A_\partial} = \bigcup_{x \in A_\partial} \omega(x).$$

(H_1). Assume X is the closure of open set X^0; ∂X^0 is nonempty and is the boundary of X^0. Moreover the C^0-semigroup $T(t)$ on X satisfies

$$T(t): \ X^0 \to X^0, \ \ T(t): \ \partial X^0 \to \partial X^0.$$

LEMMA 3.10 (see [16, Theorem 4.1, p. 392]). *Suppose $T(t)$ satisfies* (H_1) *and*
(i) *there is a $t_0 \geq 0$ such that $T(t)$ is compact for $t > t_0$;*
(ii) *$T(t)$ is point dissipative in X;*
(iii) *$\widetilde{A_\partial}$ is isolated and has an acyclic covering M.*
Then $T(t)$ is uniformly persistent iff for each $M_i \in M$, $W^s(M_i) \bigcap X^0 = \phi$.

Proof of Theorem 3.2.

Claim 1. The condition (2.2) leads to the permanence of system (1.4).

We begin by showing Claim 1 holds true for system (1.7), the subsystem of system (1.4), as the first step, we verify that the boundary planes of $R^2_+ = \{(x, y) : \ x \geq 0, \ y \geq 0\}$ repel the positive solutions to system (1.7) uniformly.

Let $C^+([-\tau, 0], R^2_+)$ denote the space of continuous functions mapping $[-\tau, 0]$ into R^2_+. We choose

$$C_1 = \{(\varphi_0, \varphi_1) \in C^+([-\tau, 0], R^2_+): \ \varphi_0(\theta) \equiv 0, \ \varphi_1(\theta) > 0, \ \theta \in [-\tau, 0]\},$$

$$C_2 = \{(\varphi_0, \varphi_1) \in C^+([-\tau, 0], R^2_+): \ \varphi_0(\theta) > 0, \ \varphi_1(\theta) \equiv 0, \ \theta \in [-\tau, 0]\}.$$

Denote $C = C_1 \bigcup C_2$, $X = C^+([-\tau, 0], R^2_+)$, and $X^0 = \text{Int}C^+([-\tau, 0], R^2_+)$; then $C = \partial X^0$. It is easy to see that system (1.7) possesses, two constant solutions in $C = \partial X^0$: $\widetilde{E_0} \in C_1$, $\widetilde{E_1} \in C_2$ with

$$\widetilde{E_0} = \{(\varphi_0, \varphi_1) \in C^+([-\tau, 0], R^2_+): \ \varphi_0(\theta) \equiv \varphi_1(\theta) \equiv 0, \ \theta \in [-\tau, 0]\},$$

$$\widetilde{E_1} = \{(\varphi_0, \varphi_1) \in C^+([-\tau, 0], R^2_+): \ \varphi_0(\theta) \equiv K, \ \varphi_1(\theta) \equiv 0, \ \theta \in [-\tau, 0]\}.$$

We verify below that the conditions of Lemma 3.10 are satisfied. By the definition of X^0 and ∂X^0 and system (1.7), it is easy to see that conditions (i) and (ii) of Lemma 3.10 are satisfied and that X^0 and ∂X^0 are invariant. Hence (H_1) is also satisfied.

Consider condition (iii) of Lemma 3.10. We have

$$\dot{x}(t)|_{(\varphi_0, \varphi_1) \in C_1} \equiv 0,$$

thus $x(t)|_{(\varphi_0, \varphi_1) \in C_1} \equiv 0$ for all $t \geq 0$. Hence we have

$$\dot{y}(t)|_{(\varphi_0, \varphi_1) \in C_1} = -dy(t) \leq 0,$$

from which follows that all points in C_1 approach $\widetilde{E_0}$, i.e., $C_1 = W^s(\widetilde{E_0})$. Similarly we can prove that all points in C_2 approach $\widetilde{E_1}$, i.e., $C_2 = W^s(\widetilde{E_1})$. Hence $\widetilde{A_\partial} = \widetilde{E_0} \bigcup \widetilde{E_1}$ and clearly it is isolated. Noting that $C_1 \bigcap C_2 = \phi$, it follows from these structural features that the flow in $\widetilde{A_\partial}$ is acyclic, satisfying condition (iii) of Lemma 3.10.

Now we show that $W^s(\widetilde{E_i}) \bigcap X^0 = \phi$, $i = 0, 1$. By Lemma 3.7, we have $x(t), y(t) > 0$ for all $t > 0$. Assume $W^s(\widetilde{E_0}) \bigcap X^0 \neq \phi$, i.e., there exists a positive solution

$(x(t), y(t))$ with $\lim_{t\to\infty}(x(t), y(t)) = (0, 0)$, then using the first equation of (1.7), we get

$$\frac{\mathrm{d}(\ln x(t))}{\mathrm{d}t} = r(1 - \frac{x(t)}{K}) - \frac{by(t)}{1 + k_1x(t) + k_2y(t)} > \frac{r}{2}$$

for all sufficiently large t. Hence we have $\lim_{t\to\infty} x(t) = +\infty$, contradicting $\lim_{t\to\infty} x(t) = 0$; this proves $W^s(\widetilde{E_0}) \bigcap X^0 = \phi$.

Now we verify $W^s(\widetilde{E_1}) \bigcap X^0 = \phi$; assume the contrary, i.e., $W^s(\widetilde{E_1}) \bigcap X^0 \neq \phi$. Then there exists a positive solution $(x(t), y(t))$ to system (1.7) with $\lim_{t\to\infty}(x(t), y(t)) = (K, 0)$, and for sufficiently small positive constant ε with

$$\varepsilon < \min\left\{\frac{nbe^{-d_j\tau}K - d - dKk_1}{2(nbe^{-d_j\tau} - dk_1 + dk_2)}, \frac{nbe^{-d_j\tau}K - d - dKk_1}{2k_2d}\right\},$$

there exists a positive constant $T = T(\varepsilon)$ such that

$$x(t) > K - \varepsilon > 0, \quad y(t) < \varepsilon \ \text{ for all } t \geq T.$$

By the second equation of (1.7) we have

$$y'(t) > \frac{nbe^{-d_j\tau}(K - \varepsilon)y(t - \tau)}{1 + k_1(K - \varepsilon) + k_2y(t - \tau)} - dy(t), \quad t \geq T + \tau. \tag{3.2}$$

Consider the equation

$$\begin{cases} v'(t) = \dfrac{nbe^{-d_j\tau}(K - \varepsilon)v(t - \tau)}{1 + k_1(K - \varepsilon) + k_2v(t - \tau)} - dv(t), \quad t \geq T + \tau, \\ v(t) = y(t), \ t \in [T, T + \tau]. \end{cases} \tag{3.3}$$

By (3.2) and the comparison theorem, we have $y(t) \geq v(t)$ for all $t > T$. On the other hand, using Theorem 4.9.1 of [26, p. 159], we have $\lim_{t\to\infty} v(t) = v^*$ for all solutions to system (3.3), where $v^* = \frac{nbe^{-d_j\tau}(K-\varepsilon)-d-dk_1(K-\varepsilon)}{dk_2} > \varepsilon$ is the unique positive equilibrium of system (3.3). Hence we get $\lim_{\overline{t\to\infty}} y(t) \geq v^* > \varepsilon$, contradicting $y(t) < \varepsilon$ as $t \geq T$. Thus we have $W^s(\widetilde{E_i}) \bigcap X^0 = \phi$, $i = 0, 1$. Now we get that system (1.7) satisfies all conditions of Lemma 3.10, thus $(x(t), y(t))$ is uniformly persistent, i.e., there exists positive constants ϵ and $T = T(\epsilon)$ such that $x(t), y(t) \geq \epsilon$ for all $t \geq T$; noting Lemma 3.9 shows that (x, y) are ultimately bounded, and this proves the permanence of system (1.7). By Lemma 1.6, $y_j(t)$ is permanent, and this proves the permanence of system (1.4).

We verify below that permanence of system (1.4) indicates (2.2). Assume the contrary, i.e., $\frac{nbe^{-d_j\tau}K}{1+k_1K} \leq d$; then by Theorem 3.1, $x(t) \to K$, $y(t) \to 0$ as $t \to \infty$, contradicting permanence of (1.4). This proves Theorem 3.2. □

4. Global attractiveness. In this section, we consider the global stability of the interior equilibrium in system (1.7). We have the following result.

THEOREM 4.1. *The positive equilibrium E in system* (1.7) *is globally attractive provided that system* (1.7) *is permanent and*

$$k_2 > \max\left\{\frac{bK(nbe^{-d_j\tau} - k_1d)}{r[(nbe^{-d_j\tau} - dk_1)K - d]}, \frac{bK(nbe^{-d_j\tau} - k_1d)}{rd}, \frac{b}{r}\right\} \tag{4.1}$$

holds true.

Consider the following single species system with delay:

$$v'(t) = \frac{a_1 v(t-\tau)}{1 + a_2 v(t-\tau)} - a_3 v(t),\ v(t) = \phi(t) \geq 0,\ v(0) > 0,\quad t \in [-\tau, 0], \tag{4.2}$$

where $a_i > 0,\ i = 1, 2, 3$. Similar to Lemma 3.7, we have $v(t) > 0$ for all $t \geq 0$. From Theorem 4.9.1 in [26] we directly have the following lemma.

LEMMA 4.2. *System* (4.2) *has a unique positive equilibrium* $v^* = \frac{a_1 - a_3}{a_2 a_3}$ *that is globally asymptotically stable provided* $a_1 > a_3$.

Proof of Theorem 4.1. By first condition and Theorem 3.2, we have that (2.2) holds. By the first equation of (1.7) and the arguments to Lemma 3.9, for sufficiently small $\varepsilon > 0$, there is a $T_1 > 0$ such that $x(t) < K + \varepsilon = \overline{x_1}$ for $t \geq T_1$. Replacing this inequality into the second equation of (1.7), we have

$$y'(t) < \frac{nbe^{-d_j\tau}\overline{x_1} y(t-\tau)}{1 + k_1\overline{x_1} + k_2 y(t-\tau)} - dy(t),\quad t \geq T_1 + \tau.$$

Consider the system

$$\begin{cases} v'(t) = \dfrac{nbe^{-d_j\tau}\overline{x_1} v(t-\tau)}{1 + k_1\overline{x_1} + k_2 v(t-\tau)} - dv(t),\ t \geq T_1 + \tau, \\ v(t) \equiv y(t),\ t \in [T_1, T_1 + \tau]. \end{cases}$$

Noting $nbe^{-d_j\tau}\overline{x_1} - d(1 + k_1\overline{x_1}) > nbe^{-d_j\tau}K - d(1 + k_1 K) > 0$. Thus by Lemma 4.2, we have

$$\lim_{t\to\infty} v(t) = \frac{nbe^{-d_j\tau}\overline{x_1} - d(1 + k_1\overline{x_1})}{k_2 d} > 0.$$

By the comparison theorem, we have $y(t) \leq v(t),\ t \geq T_1 + \tau$. Then for the sufficiently small $\varepsilon > 0$, there exists $T_2 > T_1 + \tau$ such that

$$y(t) < \frac{nbe^{-d_j\tau}\overline{x_1} - d(1 + k_1\overline{x_1})}{k_2 d} + \varepsilon = \overline{y_1},\ t \geq T_2. \tag{4.3}$$

Replacing (4.3) into the first equation of (1.7), we have

$$x'(t) > rx(t)(1 - \frac{x(t)}{K}) - \frac{bx(t)\overline{y_1}}{1 + k_2\overline{y_1}},\quad t \geq T_2.$$

By (4.1), $r > \frac{b}{k_2} > \frac{b\overline{y_1}}{1 + k_2\overline{y_1}}$. Using the comparison theorem, for sufficiently small $\varepsilon > 0$, there is a $T_3 > T_2$ such that

$$x(t) > z^* - \varepsilon = \underline{x_1} > 0,\ t \geq T_3, \tag{4.4}$$

where $z^* = K \cdot [1 - \frac{b\overline{y_1}}{r(1 + k_2\overline{y_1})}] > 0$ is the positive root for the equation

$$rx(t)(1 - \frac{x(t)}{K}) - \frac{bx(t)\overline{y_1}}{1 + k_2\overline{y_1}} = 0.$$

Replacing (4.4) into the second equation of (1.7), we have

$$y'(t) > \frac{nbe^{-d_j\tau}\underline{x_1} y(t-\tau)}{1 + k_1\underline{x_1} + k_2 y(t-\tau)} - dy(t),\quad t \geq T_3 + \tau.$$

By (4.4), we have

$$\begin{aligned} nbe^{-d_j\tau}\underline{x_1} - d(1+k_1\underline{x_1}) &= (nbe^{-d_j\tau} - dk_1)\cdot\left\{K[1-\frac{b\overline{y_1}}{r(1+k_2\overline{y_1})}] - \varepsilon\right\} - d \\ &> (nbe^{-d_j\tau} - dk_1)\cdot\left\{K[1-\frac{b}{rk_2}] - \varepsilon\right\} - d \\ &= \frac{(nbe^{-d_j\tau} - dk_1)(K-\varepsilon) - d}{k_2} \\ &\quad\cdot\left\{k_2 - \frac{bK(nbe^{-d_j\tau} - dk_1)}{r[(nbe^{-d_j\tau} - dk_1)(K-\varepsilon) - d]}\right\}. \end{aligned}$$

Using (4.1), we can get

$$nbe^{-d_j\tau}\underline{x_1} - d(1+k_1\underline{x_1}) > 0 \quad \text{for sufficiently small } \varepsilon. \tag{4.5}$$

By Lemma 4.2 and the similar arguments to $\overline{y_1}$, for the above selected $\varepsilon > 0$, there exists $T_4 > T_3 + \tau$ such that

$$y(t) > \frac{nbe^{-d_j\tau}\underline{x_1} - d(1+k_1\underline{x_1})}{k_2 d} - \varepsilon = \underline{y_1} > 0,\ t \geq T_4. \tag{4.6}$$

Therefore we have that

$$\underline{x_1} < x(t) < \overline{x_1},\ \underline{y_1} < y(t) < \overline{y_1},\quad t \geq T_4,$$

hold for system (1.7).

Replacing (4.6) into the first equation of (1.7), we have

$$x'(t) < rx(t)(1-\frac{x(t)}{K}) - \frac{bx(t)\underline{y_1}}{1+k_2\underline{y_1}},\quad t \geq T_4.$$

Since $r - \frac{b\underline{y_1}}{1+k_2\underline{y_1}} > r - \frac{b\overline{y_1}}{1+k_2\overline{y_1}} > 0$, by the comparison theorem, for sufficiently small $\varepsilon > 0$, there is a $T_5 > T_4$ such that

$$x(t) < z_1^* + \varepsilon = \overline{x_2} > 0,\ t \geq T_5, \tag{4.7}$$

with $z_1^* = K\cdot[1 - \frac{b\underline{y_1}}{r(1+k_2\underline{y_1})}] > 0$. From the definition of $\overline{x_2}$ we get

$$\overline{x_2} < K < \overline{x_1}.$$

Replacing (4.7) into the second equation of (1.7), we have

$$y'(t) < \frac{nbe^{-d_j\tau}\overline{x_2}y(t-\tau)}{1+k_1\overline{x_2}+k_2y(t-\tau)} - dy(t),\quad t \geq T_5+\tau.$$

Since $\overline{x_2} > \underline{x_1}$ and noting (4.5), we have $nbe^{-d_j\tau}\overline{x_2} - d(1+k_1\overline{x_2}) > nbe^{-d_j\tau}\underline{x_1} - d(1+k_1\underline{x_1}) > 0$. Thus using arguments similar to above, for the sufficiently small $\varepsilon > 0$, there is a $T_6 > T_5 + \tau$ such that

$$y(t) < \frac{nbe^{-d_j\tau}\overline{x_2} - d(1+k_1\overline{x_2})}{k_2 d} + \varepsilon = \overline{y_2},\ t \geq T_6, \tag{4.8}$$

by (4.3), (4.8) we get $\overline{y_2} < \overline{y_1}$.

Replacing (4.8) into the first equation of (1.7), we have

$$x'(t) > rx(t)(1-\frac{x(t)}{K}) - \frac{bx(t)\overline{y_2}}{1+k_2\overline{y_2}}, \quad t \geq T_6.$$

From (4.1), $r > \frac{b}{k_2} > \frac{b\overline{y_1}}{1+k_2\overline{y_1}} > \frac{b\overline{y_2}}{1+k_2\overline{y_2}}$. Then by the comparison theorem, for sufficiently small $\varepsilon > 0$, there is a $T_7 > T_6$ such that

$$x(t) > z_2^* - \varepsilon = \underline{x_2} > 0, \ t \geq T_7, \tag{4.9}$$

with $z_2^* = K \cdot [1 - \frac{b\overline{y_2}}{r(1+k_2\overline{y_2})}] > 0$. By the definition of $\underline{x_2}$, we have $\underline{x_2} > \underline{x_1}$.

Replacing (4.9) into the second equation of (1.7), then by arguments similar to those for $\overline{y_2}$, we get that there exists a $T_8 > T_7 + \tau$ such that

$$y(t) > \frac{nbe^{-d_j\tau}\underline{x_2} - d(1+k_1\underline{x_2})}{k_2 d} - \varepsilon = \underline{y_2} > 0, \ t \geq T_8, \tag{4.10}$$

and we get $\underline{y_2} > \underline{y_1}$.

Therefore, we have

$$0 < \underline{x_1} < \underline{x_2} < x(t) < \overline{x_2} < \overline{x_1}, \quad 0 < \underline{y_1} < \underline{y_2} < y(t) < \overline{y_2} < \overline{y_1}, \quad t \geq T_8. \tag{4.11}$$

Repeating the above arguments, we get the four sequences $\{\overline{x_n}\}_{n=1}^{\infty}, \{\underline{x_n}\}_{n=1}^{\infty}$, $\{\overline{y_n}\}_{n=1}^{\infty}, \{\underline{y_n}\}_{n=1}^{\infty}$ with

$$\begin{aligned} &0 < \underline{x_1} < \underline{x_2} < \cdots < \underline{x_n} < x(t) < \overline{x_n} < \cdots < \overline{x_2} < \overline{x_1}, \\ &0 < \underline{y_1} < \underline{y_2} < \cdots < \underline{y_n} < y(t) < \overline{y_n} < \cdots < \overline{y_2} < \overline{y_1}, \quad t \geq T_{4n}. \end{aligned} \tag{4.12}$$

From (4.12) follows that the limit of each sequence in $\{\overline{x_n}\}_{n=1}^{\infty}, \{\underline{x_n}\}_{n=1}^{\infty}, \{\overline{y_n}\}_{n=1}^{\infty}$, $\{\underline{y_n}\}_{n=1}^{\infty}$ exists. Denote

$$\overline{x} = \lim_{n\to\infty} \overline{x_n}, \quad \overline{y} = \lim_{n\to\infty} \overline{y_n}, \quad \underline{x} = \lim_{n\to\infty} \underline{x_n}, \quad \underline{y} = \lim_{n\to\infty} \underline{y_n};$$

thus we get $\overline{x} \geq \underline{x}, \ \overline{y} \geq \underline{y}$. To complete the proof, it suffices to prove $\overline{x} = \underline{x}, \quad \overline{y} = \underline{y}$.

By the definition of $\overline{y_n}, \underline{y_m}$ we have

$$\overline{y_n} = \frac{nbe^{-d_j\tau}\overline{x_n} - d(1+k_1\overline{x_n})}{k_2 d} + \varepsilon, \quad \underline{y_m} = \frac{nbe^{-d_j\tau}\underline{x_m} - d(1+k_1\underline{x_m})}{k_2 d} - \varepsilon;$$

then we get

$$\overline{y_n} - \underline{y_m} = \frac{nbe^{-d_j\tau} - dk_1}{k_2 d} \cdot (\overline{x_n} - \underline{x_m}) + 2\varepsilon. \tag{4.13}$$

By the definition of $\overline{x_n}, \ \underline{x_n}$ and using (4.13), we have

$$\begin{aligned} \overline{x_n} - \underline{x_n} &= K \cdot \left[1 - \frac{b\underline{y_{n-1}}}{r(1+k_2\underline{y_{n-1}})}\right] - K \cdot \left[1 - \frac{b\overline{y_n}}{r(1+k_2\overline{y_n})}\right] + 2\varepsilon \\ &= \frac{bK}{r} \cdot \left[\frac{\overline{y_n} - \underline{y_{n-1}}}{(1+k_2\underline{y_{n-1}})(1+k_2\overline{y_n})}\right] + 2\varepsilon \\ &= \frac{bK}{r} \cdot \frac{[nbe^{-d_j\tau} - dk_1]/k_2 d \cdot (\overline{x_n} - \underline{x_{n-1}}) + 2\varepsilon}{(1+k_2\underline{y_{n-1}})(1+k_2\overline{y_n})} + 2\varepsilon \\ &< \frac{bK}{k_2 dr} \cdot [nbe^{-d_j\tau} - dk_1] \cdot (\overline{x_n} - \underline{x_{n-1}}) + 2\varepsilon\left(1 + \frac{bK}{r}\right). \end{aligned} \tag{4.14}$$

Let $n \to \infty$; then we have

$$\overline{x} - \underline{x} \le \frac{bK}{k_2 dr} \cdot [nbe^{-d_j\tau} - dk_1] \cdot (\overline{x} - \underline{x}) + 2\varepsilon \left(1 + \frac{bK}{r}\right),$$

thus

$$\left\{1 - \frac{bK}{k_2 dr} \cdot [nbe^{-d_j\tau} - dk_1]\right\} (\overline{x} - \underline{x}) \le 2\varepsilon \left(1 + \frac{bK}{r}\right).$$

From (4.1), we have $1 - \frac{bK}{k_2 dr} \cdot [nbe^{-d_j\tau} - dk_1] > 0$, and noting that ε can be arbitrarily small, then we have $\overline{x} = \underline{x}$. By (4.13) and let $n\ , m \to \infty$, we get $\overline{y} = \underline{y}$. This proves Theorem 4.1. □

5. Stability switches. Considering the characteristic equation of (1.7), we write (1.7) as

$$\underline{x}'(t) = \underline{F}(\underline{x}(t), \underline{x}(t-\tau))$$

and denote

$$G = \left(\frac{\partial \underline{F}}{\partial \underline{x}(t)}\right)_{\underline{x}^*}, \quad H = \left(\frac{\partial \underline{F}}{\partial \underline{x}(t-\tau)}\right)_{\underline{x}^*}.$$

Thus characteristic equation of (1.7) at the equilibrium $\underline{x}^*$ takes the form as follows:

$$\det(G + He^{-\lambda\tau} - \lambda I) = 0. \tag{5.1}$$

We have

$$G = \begin{pmatrix} r - 2\frac{r}{K}x - \frac{\partial g}{\partial x} & -\frac{\partial g}{\partial y} \\ 0 & -d \end{pmatrix}, \quad H = \begin{pmatrix} 0 & 0 \\ ne^{-d_j\tau}\frac{\partial g}{\partial x} & ne^{-d_j\tau}\frac{\partial g}{\partial y} \end{pmatrix},$$

where

(5.2)

$$g(x,y) = \frac{bxy}{1 + k_1 x + k_2 y}, \ \frac{\partial g(x,y)}{\partial x} = \frac{by(1+k_2 y)}{(1 + k_1 x + k_2 y)^2}, \ \frac{\partial g(x,y)}{\partial y} = \frac{bx(1+k_1 x)}{(1 + k_1 x + k_2 y)^2}.$$

Thus the characteristic equation of system (1.7) at some equilibrium (x^0, y^0) is as follows:

$$\begin{vmatrix} r - 2\frac{r}{K}x^0 - g'_x(x^0,y^0) - \lambda & -g'_y(x^0,y^0) \\ ne^{-(\lambda+d_j)\tau}g'_x(x^0,y^0) & ne^{-(\lambda+d_j)\tau}g'_y(x^0,y^0) - d - \lambda \end{vmatrix} = 0. \tag{5.3}$$

At the equilibrium $E_0 = (0,0)$ we have $g'_x(0,0) = g'_y(0,0) = 0$ and the characteristic equation (5.3) reduces to $\begin{vmatrix} r-\lambda & 0 \\ 0 & -d-\lambda \end{vmatrix} = 0$, i.e., E_0 is an unstable saddle point.

THEOREM 5.1. *The equilibrium $E_1 = (K, 0)$ is*

(i) *unstable if $\frac{nbe^{-d_j\tau}K}{1+k_1K} > d$;*

(ii) *linearly neutrally stable if $\frac{nbe^{-d_j\tau}K}{1+k_1K} = d$;*

(iii) *asymptotically stable if $\frac{nbe^{-d_j\tau}K}{1+k_1K} < d$.*

By Theorem 5.1 and the arguments to Theorem 3.1, we directly have that equilibrium $(K,0)$ of system (1.7) is globally asymptotically stable iff $\frac{nbe^{-d_j\tau}K}{1+k_1K} \le d$ holds true. Using (1.6), we can easily prove that the global asymptotic stability of $(K,0)$ in system (1.7) is equivalent to that of $(K,0,0)$ in system (1.4). Thus we have the next corollary.

COROLLARY 5.2. *The equilibrium* $(K,0,0)$ *of system* (1.4) *is globally asymptotically stable iff* $\frac{nbe^{-d_j\tau}K}{1+k_1K} \le d$ *holds true.*

Proof of Theorem 5.1. By (5.3), we get that the characteristic equation of (1.7) at the equilibrium E_1 is

$$(\lambda + r)[ne^{-(\lambda+d_j\tau)}bg'_y(K,0) - d - \lambda] = 0. \tag{5.4}$$

Hence, one characteristic root is $\lambda = -r < 0$. Since $g'_y(K,0) = \frac{K}{1+k_1K}$, then the other are the roots of

$$g(\lambda) = \lambda + d - \frac{nbK}{1+k_1K} \cdot e^{-d_j\tau}e^{-\lambda\tau} = 0.$$

(i) Assume that $\frac{nbe^{-d_j\tau}K}{1+k_1K} > d$; then $g(0) = d - \frac{nbe^{-d_j\tau}K}{1+k_1K} < 0$, and $g(+\infty) = \infty$. Hence $g(\lambda)$ has at least one positive root and E_1 is unstable.

(ii) As $\frac{nbe^{-d_j\tau}K}{1+k_1K} = d$, $g(\lambda) = \lambda + d - de^{-\lambda\tau}$ and $\lambda = 0$ is a root of $g(\lambda) = 0$. Furthermore, since $g'(\lambda) = 1 + \tau de^{-\lambda\tau}$, we have $g'(0) > 0$. Then, the root $\lambda = 0$ is simple.
Then if the other roots are $\lambda = \alpha + i\omega$ they must satisfy

$$(\alpha + d)^2 + \omega^2 = d^2e^{-2\alpha\tau}.$$

Hence we must have $\alpha \le 0$, i.e., all the other roots have real nonpositive parts. Therefore E_1 is linearly neutrally stable.

(iii) Assume now that $\frac{nbe^{-d_j\tau}K}{1+k_1K} < d$, i.e.,

$$d - \frac{nbK}{1+k_1K} \cdot e^{-d_j\tau}e^{-\lambda\tau} > 0.$$

Then $g(\lambda) = 0$ implies that

$$\lambda + d = \frac{nbK}{1+k_1K} \cdot e^{-d_j\tau}e^{-\lambda\tau}.$$

If $\mathrm{Re}(\lambda) \ge 0$, then

$$|\lambda + d| > d > \frac{nbK}{1+k_1K} \cdot e^{-d_j\tau}e^{-\lambda\tau}.$$

This shows that all roots of $g(\lambda) = 0$ must have negative real parts, and therefore E_1 is asymptotically stable, proving (iii). ∎

Now, we consider the stability switches of the interior equilibrium $E = (x^*, y^*)$ as maturation time delay τ increases. We will adopt the following nomenclature:

$$g^* = g(x^*,y^*), \quad g'_{x^*} = g'_x(x^*,y^*), \quad g'_{y^*} = g'_y(x^*,y^*).$$

By (5.3), we get that the characteristic equation at E is as follows:

$$D(\lambda,\tau) = P(\lambda,\tau) + Q(\lambda,\tau)e^{-\lambda\tau} = 0, \tag{5.5}$$

where

$$\begin{cases} P(\lambda,\tau)=\lambda^2+P_1(\tau)\lambda+P_0(\tau), \\ P_1(\tau)=d-R+g'_{x^*}, \\ P_0(\tau)=(-R+g'_{x^*})d, \end{cases} \tag{5.6}$$

$$\begin{cases} Q(\lambda,\tau)=\lambda Q_1(\tau)+Q_0(\tau), \\ Q_1(\tau)=-ne^{-d_j\tau}g'_{y^*}, \\ Q_0(\tau)=Rne^{-d_j\tau}g'_{y^*}, \end{cases} \tag{5.7}$$

where $R=r-2\frac{r}{K}x^*$.

Of course, the characteristic equation (5.5) must be considered in the interval $I=[0,\tau^*)$ of existence of the positive equilibrium.

First verify that $\lambda=0$ cannot be a root of (5.5) for any $\tau\in I$, i.e.,

$$P(0,\tau)+Q(0,\tau)\neq 0.$$

Noting

$$P(0,\tau)+Q(0,\tau)=P_0(\tau)+Q_0(\tau)=(-R+g'_{x^*})d+Rne^{-d_j\tau}g'_{y^*},$$

and

$$g'_y=\frac{g}{y}\left(1-\frac{k_2}{b}\frac{g}{x}\right), \qquad g'_x=\frac{g}{x}\left(1-\frac{k_1}{b}\frac{g}{y}\right),$$
$$R=\frac{g^*}{x^*}-\frac{r}{K}x^*, \qquad ne^{-d_j\tau}\frac{g^*}{y^*}=d,$$

we get

$$\begin{aligned} P(0,\tau)+Q(0,\tau) &= -Rd+dg'_{x^*}+Rne^{-d_j\tau}\frac{g^*}{y^*}\left(1-\frac{k_2}{b}\frac{g^*}{x^*}\right) \\ &= dg'_{x^*}-Rd\frac{k_2}{b}\frac{g^*}{x^*} \\ &= d\left(\frac{g^*}{x^*}\left(1-\frac{k_1}{b}\frac{g^*}{y^*}\right)-\frac{k_2}{b}\frac{g^*}{x^*}\left(\frac{g^*}{x^*}-\frac{r}{K}x^*\right)\right) \\ &= d\frac{g^*}{x^*}\left(1-\left(\frac{k_1}{b}\frac{g^*}{y^*}+\frac{k_2}{b}\frac{g^*}{x^*}\right)+\frac{k_2}{b}\frac{r}{K}g^*\right) \end{aligned}$$

Now remark that $\frac{1}{b}(\frac{k_1}{y^*}+\frac{k_2}{x^*})g^*=\frac{k_1x^*+k_2y^*}{1+k_1x^*+k_2y^*}$ and therefore

$$P(0,\tau)+Q(0,\tau)>0,$$

for all $\tau\in I=[0,\tau^*)$.

The characteristic equation (5.5) at $\tau=0$ becomes

$$P(\lambda,0)+Q(\lambda,0)=0,$$

i.e.,

$$\lambda^2+(P_1(0)+Q_1(0))\lambda+P_0(0)+Q_0(0)=0, \tag{5.8}$$

where $P_0(0)+Q_0(0)>0$ since $P_0(\tau)+Q_0(\tau)>0$ for all $\tau\in[0,\tau^*)$.

Let us give an explicit structure for $P_1(0) + Q_1(0)$. From (5.6), (5.7)

$$P_1(0) + Q_1(0) = d - R + g'_{x^*} - ng'_{y^*},$$

where at interior equilibrium (and at $\tau = 0$)

$$R = \frac{g^*}{x^*} - \frac{r}{K}x^*, \quad n\frac{g^*}{y^*} = d.$$

Hence,

$$\begin{aligned} P_1(0) + Q_1(0) &= d - \frac{g^*}{x^*} + \frac{r}{K}x^* + \frac{g^*}{x^*} - \frac{k_1}{b}\frac{(g^*)^2}{x^*y^*} - n\frac{g^*}{y^*}\left(1 - \frac{k_2}{b}\frac{g^*}{x^*}\right) \\ &= d - \frac{g^*}{x^*} + \frac{r}{K}x^* + \frac{g^*}{x^*} - \frac{k_1}{b}\frac{(g^*)^2}{x^*y^*} - d + d\frac{k_2}{b}\frac{g^*}{x^*} \\ &= \frac{r}{K}x^* + \frac{d}{b}\frac{g^*}{x^*}\left(k_2 - k_1\frac{1}{d}\frac{g^*}{y^*}\right) \\ &= \frac{r}{K}x^* + \frac{d}{b}\frac{g^*}{x^*}\left(k_2 - \frac{k_1}{n}\right). \end{aligned}$$

Then, the roots of (5.8) determine the stability properties of the interior equilibrium at $\tau = 0$. Stability switches for increasing τ in $I = [0, \tau^*)$ may occur only with a pair of roots $\lambda = \pm i\omega(\tau)$, $\omega(\tau)$ real positive, that cross the imaginary axis.

To determine the stability switch delay values we proceed as follows (see [8, section 4]).

Assume $\lambda = \pm i\omega(\tau), \omega(\tau) > 0$ in (5.5); we have

$$\begin{cases} P(i\omega, \tau) = -\omega^2 + i\omega P_1(\tau) + P_0(\tau), \\ P_R(i\omega, \tau) = P_0(\tau) - \omega^2, \quad P_I(i\omega, \tau) = \omega P_1(\tau), \end{cases} \tag{5.9}$$

$$\begin{cases} Q(i\omega, \tau) = i\omega Q_1(\tau) + Q_0(\tau), \\ Q_R(i\omega, \tau) = Q_0(\tau), \quad Q_I(i\omega, \tau) = \omega Q_1(\tau). \end{cases} \tag{5.10}$$

The first step is that of looking for the positive roots $\omega(\tau) > 0$ of

$$F(\omega, \tau) = |P(i\omega, \tau)|^2 - |Q(i\omega, \tau)|^2 = 0 \tag{5.11}$$

in $I = [0, \tau^*)$. Since

$$\begin{aligned} F(\omega, \tau) &= (P_0(\tau) - \omega^2)^2 + \omega^2 P_1(\tau)^2 - [Q_0^2(\tau) + \omega^2 Q_1^2(\tau)], \\ &= P_0^2(\tau) + \omega^4 - 2P_0(\tau)\omega^2 + \omega^2 P_1^2(\tau) - Q_0^2(\tau) - \omega^2 Q_1^2(\tau), \\ &= \omega^4 + \omega^2(-2P_0(\tau) + P_1^2(\tau) - Q_1^2(\tau)) + P_0^2(\tau) - Q_0^2(\tau), \end{aligned}$$

hence we have

$$\begin{cases} F(\omega, \tau) = \omega^4 + b(\tau)\omega^2 + c(\tau) = 0, \\ b(\tau) = -2P_0(\tau) + P_1^2(\tau) - Q_1^2(\tau), \\ c(\tau) = P_0^2(\tau) - Q_0^2(\tau). \end{cases} \tag{5.12}$$

Depending on the sign of $b(\tau)$ and $c(\tau)$ the system (5.12) may have no positive real roots, or the root

$$\omega_+(\tau) = \left[\frac{1}{2}\{-b(\tau) + \sqrt{b(\tau)^2 - 4c(\tau)}\}\right]^{1/2}, \quad \tau \in I_+ \subseteq I,$$

or otherwise the root

$$\omega_-(\tau) = \left[\frac{1}{2}\{-b(\tau) - \sqrt{b(\tau)^2 - 4c(\tau)}\}\right]^{1/2}, \quad \tau \in I_- \subseteq I,$$

or, as a last case, both the roots $\omega_+(\tau)$ and $\omega_-(\tau)$. Note that if system (5.12) has no positive roots $\omega(\tau)$ in I, then no stability switches can occur.

According to characteristic equation (5.5), we can say that at $\tau = 0$, i.e., without stage structure, a necessary and sufficient condition for the asymptotic stability of positive equilibrium E is that

$$P_1(0) + Q_1(0) > 0,$$

whereas if $P_1(0) + Q_1(0) < 0$ the positive equilibrium is unstable.

Of course, from the structure of $P_1(0)+Q_1(0)$, a sufficient condition for asymptotic stability of E at $\tau = 0$ is that k_2 is sufficiently large to ensure that

$$k_2 - \frac{k_1}{n} > 0.$$

Stability switches for increasing τ in $I = [0, \tau^*)$ may occur only with a pair of roots $\lambda = \pm i\omega(\tau)$, $\omega(\tau)$ real positive, that cross the imaginary axis.

Now, we prove that E is asymptotically stable provided that k_2 is sufficiently large. We have the next theorem.

THEOREM 5.3. *The positive equilibrium E of* (1.7) *is asymptotically stable provided that system* (1.7) *is permanent and*

$$k_2 > \max\left\{\frac{k_1}{n},\ 2 \cdot \frac{bK(nbe^{-d_j\tau} - dk_1)}{r\left[bnKe^{-d_j\tau} - d(1 + k_1K)\right]}\right\}. \tag{5.13}$$

Define $M^K_{k_2}, M^n_{k_2}$ with

$$M^K_{k_2} = \sup_{K>0}\left\{\frac{nbe^{-d_j\tau}K}{1 + k_1K} > d + \delta_0 \,\middle|\, \frac{k_1}{n},\ 2 \cdot \frac{bK(nbe^{-d_j\tau} - dk_1)}{r\left[bnKe^{-d_j\tau} - d(1 + k_1K)\right]}\right\},$$

$$M^n_{k_2} = \sup_{n>0}\left\{\frac{nbe^{-d_j\tau}K}{1 + k_1K} > d + \delta_0 \,\middle|\, \frac{k_1}{n},\ 2 \cdot \frac{bK(nbe^{-d_j\tau} - dk_1)}{r\left[bnKe^{-d_j\tau} - d(1 + k_1K)\right]}\right\},$$

where δ_0 is some positive constant, thus $0 < M^K_{k_2}, M^n_{k_2} < \infty$. Using Theorem 3.2 and Theorem 5.3, we directly have the next corollary.

COROLLARY 5.4. *Assume $k_2 > M^K_{k_2}$, $\frac{nbe^{-d_j\tau}K}{1+k_1K} > d + \delta_0$; then the positive equilibrium E of* (1.7) *is asymptotically stable for all $K > 0$.*

COROLLARY 5.5. *Assume $k_2 > M^n_{k_2}$, $\frac{nbe^{-d_j\tau}K}{1+k_1K} > d + \delta_0$; then the positive equilibrium E of* (1.7) *is asymptotically stable for all $n > 0$.*

Proof of Theorem 5.3. To complete the proof, it suffices to prove that E has no stability switches as τ increases and that E is stable at $\tau = 0$. Hence we only need to consider the roots of (5.5) with $\tau = 0$, i.e., (5.8). Noting $P_0(0) + Q_0(0) > 0$ and

$$P_1(0) + Q_1(0) = \frac{r}{K}x^* + \frac{d}{b}\frac{g^*}{x^*}\left(k_2 - \frac{k_1}{n}\right) > \frac{r}{K}x^* > 0,$$

the roots of (5.8) must have negative real parts, proving E is stable at $\tau = 0$. Now we show E has no stability switches as τ increases in $I = [0, \tau^*)$. Thus we only need to prove that (5.12) has no positive roots $\omega(\tau)$ in I.

By (2.3), (5.6), (5.7), and (5.12) we have

$$\begin{aligned} b(\tau) &= -2(-R+g'_{x^*})d + (d-R+g'_{x^*})^2 - (-ne^{-d_j\tau}g'_{y^*})^2,\\ c(\tau) &= (-R+g'_{x^*})^2d^2 - (Rne^{-d_j\tau}g'_{y^*})^2,\\ R &= r - 2\tfrac{r}{K}x^* = \tfrac{g^*}{x^*} - \tfrac{r}{K}x^*. \end{aligned} \tag{5.14}$$

By Theorem 3.2, permanence of system (1.7) implies (2.2). Thus from (5.13) follows

$$k_2 > 2\cdot\frac{nbe^{-d_j\tau}-dk_1}{nre^{-d_j\tau}}. \tag{5.15}$$

Using (2.3), (5.2) and noting that $x^* < K$ and that B in (2.3) is negative under (5.15), then we have

$$x^* > \frac{1}{2}(-B+|B|) = -B = K\cdot\left(1-\frac{nbe^{-d_j\tau}-dk_1}{nrk_2e^{-d_j\tau}}\right) > K/2 > 0; \tag{5.16}$$

$$0 < y^* < \frac{K(nbe^{-d_j\tau}-dk_1)-d}{dk_2}; \tag{5.17}$$

$$0 < g* = \frac{dy^*}{ne^{-d_j\tau}} < \frac{K(nbe^{-d_j\tau}-dk_1)-d}{nk_2e^{-d_j\tau}}; \tag{5.18}$$

$$\begin{aligned} 0 < g'_{y^*} &= \frac{bx^*(1+k_1x^*)}{(1+k_1x^*+k_2y^*)^2} = \frac{d^2(1+k_1x^*)}{bn^2e^{-2d_j\tau}x^*} < \frac{d^2(1+k_1K)}{bn^2e^{-2d_j\tau}}\cdot\frac{1}{x^*}\\ &< \frac{d^2(1+k_1K)}{bn^2Ke^{-2d_j\tau}}\Big/\left(1-\frac{nbe^{-d_j\tau}-dk_1}{nrk_2e^{-d_j\tau}}\right); \end{aligned} \tag{5.19}$$

$$R = r - 2\frac{r}{K}x^* < -r + \frac{2(nbe^{-d_j\tau}-dk_1)}{nk_2e^{-d_j\tau}} < 0. \tag{5.20}$$

Thus by (5.14), we get

$$\begin{cases} b(\tau) = d^2+R^2+(g'_{x^*})^2 - 2Rg'_{x^*} - (ne^{-d_j\tau}g'_{y^*})^2\\ \qquad > R^2+d^2-(ne^{-d_j\tau}g'_{y^*})^2\\ \qquad = R^2+(d-ne^{-d_j\tau}g'_{y^*})(d+ne^{-d_j\tau}g'_{y^*}),\\ \\ c(\tau) = [d(-R+g'_{x^*})+Rne^{-d_j\tau}g'_{y^*}]\cdot[d(-R+g'_{x^*})-Rne^{-d_j\tau}g'_{y^*}]\\ \qquad = [dg'_{x^*}-R(d-ne^{-d_j\tau}g'_{y^*})]\cdot[d(-R+g'_{x^*})-Rne^{-d_j\tau}g'_{y^*}]. \end{cases} \tag{5.21}$$

By (5.19),

$$\begin{aligned} d-ne^{-d_j\tau}g'_{y^*} &> d - \frac{d^2(1+k_1K)}{bnKe^{-d_j\tau}}\Big/\left(1-\frac{nbe^{-d_j\tau}-dk_1}{nrk_2e^{-d_j\tau}}\right)\\ &= d\cdot\left[1-\frac{nbe^{-d_j\tau}-dk_1}{nrk_2e^{-d_j\tau}} - \frac{d(1+k_1K)}{bnKe^{-d_j\tau}}\right]\Big/\left(1-\frac{nbe^{-d_j\tau}-dk_1}{nrk_2e^{-d_j\tau}}\right)\\ &= d\cdot\left[\frac{bnKe^{-d_j\tau}-d(1+k_1K)}{bnKe^{-d_j\tau}} - \frac{nbe^{-d_j\tau}-dk_1}{nrk_2e^{-d_j\tau}}\right]\Big/\left(1-\frac{nbe^{-d_j\tau}-dk_1}{nrk_2e^{-d_j\tau}}\right)\\ &= d\cdot\frac{\frac{bnKe^{-d_j\tau}-d(1+k_1K)}{bnk_2Ke^{-d_j\tau}}}{1-\frac{nbe^{-d_j\tau}-dk_1}{nrk_2e^{-d_j\tau}}}\cdot\left(k_2-\frac{bK(nbe^{-d_j\tau}-dk_1)}{r\left[bnKe^{-d_j\tau}-d(1+k_1K)\right]}\right) > 0. \end{aligned}$$

Then we have $b(\tau),\ c(\tau) > 0$. Thus $F(\omega, \tau) \neq 0$ for all $\tau \in I = [0, \tau^*)$, i.e., there are no stability switches for $\tau \in I = [0, \tau^*)$. This proves Theorem 5.3. □

The second step to find the τ values of the stability switches requires that for each positive root $\omega(\tau)$ of (5.12) we define the angle $\theta(\tau) \in (0, 2\pi)$ as a solution of

$$\begin{cases} \sin\theta(\tau) = \dfrac{-(P_0(\tau) - \omega^2(\tau))\omega(\tau)Q_1(\tau) + \omega(\tau)P_1(\tau)Q_0(\tau)}{\omega^2(\tau)Q_1^2(\tau) + Q_0^2(\tau)}, \\ \cos\theta(\tau) = -\dfrac{(P_0(\tau) - \omega^2(\tau))Q_0(\tau) + \omega^2(\tau)P_1(\tau)Q_1(\tau)}{\omega^2(\tau)Q_1^2(\tau) + Q_0^2(\tau)} \end{cases} \tag{5.22}$$

for every $\tau \in I_\omega$, $I_\omega \subseteq I$, where I_ω is the subset of I in which the positive root $\omega(\tau)$ of (5.12) is defined (i.e., I_ω is I_+ or I_-).

The third step requires the definition for each $\omega(\tau)$ solution of (5.12) of the functions $I_\omega \mapsto \mathbf{R}$

$$S_n(\tau) := \tau - \frac{\theta(\tau) + n^2\pi}{\omega(\tau)},\ n \in \mathbf{N}_0, \tag{5.23}$$

that are continuous and differentiable in I_ω.

Still according to Beretta and Kuang [8, section 4], the following theorem holds true.

THEOREM 5.6. *The characteristic equation* (5.5) *has a pair of simple and conjugate pure imaginary roots* $\lambda = \pm i\omega(\tau^*)$, $\omega(\tau^*)$ *real positive, at* $\tau^* \in I_\omega$ *if* $S_n(\tau^*) = 0$ *for some* $n \in \mathbf{N}_0$.

If $\omega(\tau^*) = \omega_+(\tau^*)$, *this pair of simple conjugate pure imaginary roots crosses the imaginary axis from left to right (as* τ *increases) if* $\delta_+(\tau^*) > 0$ *and from right to left if* $\delta_+(\tau^*) < 0$, *where*

$$\delta_+(\tau^*) = \operatorname{sign}\left\{\frac{\mathrm{d}\ \mathrm{Re}\lambda}{\mathrm{d}\tau}\Big|_{\lambda=\mathrm{i}\omega_+(\tau^*)}\right\} = \operatorname{sign}\left\{\frac{\mathrm{dS_n}(\tau)}{\mathrm{d}\tau}\Big|_{\tau=\tau^*}\right\}. \tag{5.24}$$

If $\omega(\tau^*) = \omega_-(\tau^*)$, *this pair of simple conjugate pure imaginary roots crosses the imaginary axis from left to right if* $\delta_-(\tau^*) > 0$, *and from right to left if* $\delta_-(\tau^*) < 0$, *where*

$$\delta_-(\tau^*) = \operatorname{sign}\left\{\frac{\mathrm{d}\ \mathrm{Re}\lambda}{\mathrm{d}\tau}\Big|_{\lambda=\mathrm{i}\omega_-(\tau^*)}\right\} = -\operatorname{sign}\left\{\frac{\mathrm{dS_n}(\tau)}{\mathrm{d}\tau}\Big|_{\tau=\tau^*}\right\}. \tag{5.25}$$

Now, we show some numeric results. Figure 5.1 shows the solutions of model (1.4) with different predator maturation times τ. It seems that the interior equilibrium E is stable at $\tau = 0.8$ and $x(t), y(t)$ are unstable and periodically oscillated as $\tau = 6$; when $\tau = 10$, though still periodically oscillated, the oscillation amplitudes of $x(t), y(t)$ are smaller than those for $\tau = 6$, suggesting that E at $\tau = 10$ is less unstable than that at $\tau = 6$; when τ reaches 14, E becomes stable again. This behavior is expected as result of Theorem 5.7, which, for the same set of parameters as in Figure 5.1, states that interior equilibrium E remains asymptotically stable for τ from 0 up to $\tau_{0_1}^+ = 1.28$, is unstable with sustained oscillations for τ in the interval $(\tau_{0_1}^+ = 1.28,\ \tau_{0_2}^+ = 11.83)$, and returns asymptotically stable for $\tau > \tau_{0_2}^+ = 11.83$.

We try to have Figure 5.2 reflect the above changes of stability of positive equilibrium E as τ increases from 0.8 to 15. For Figure 5.2, each vertical black strip corresponds to the component of $x(t)$ and $y(t)$ in that $t \in [200 * \tau, 500 * \tau]$, respectively.

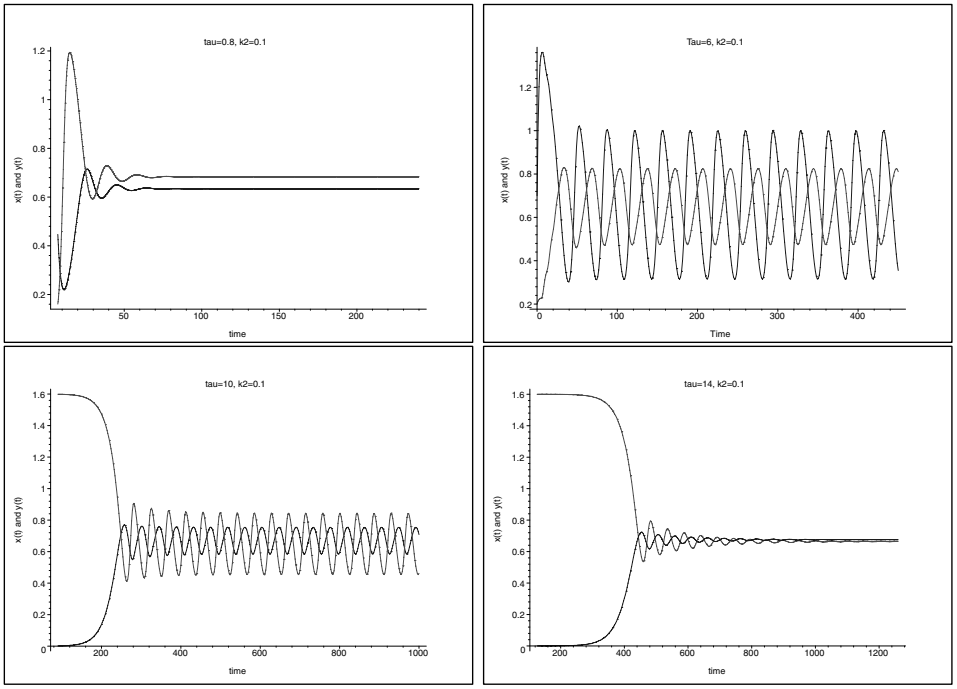


FIG. 5.1. *Solutions to system* (1.4) *with* $r = n = k_1 = 1$, $K = 1.6$, $b = 1.5$, $d = 0.5$, $k_2 = 0.1$, $d_j = 0.01$, $x(\theta) \equiv 0.7$, $y(\theta) \equiv 0.2$, $\theta \in [-\tau, 0]$.

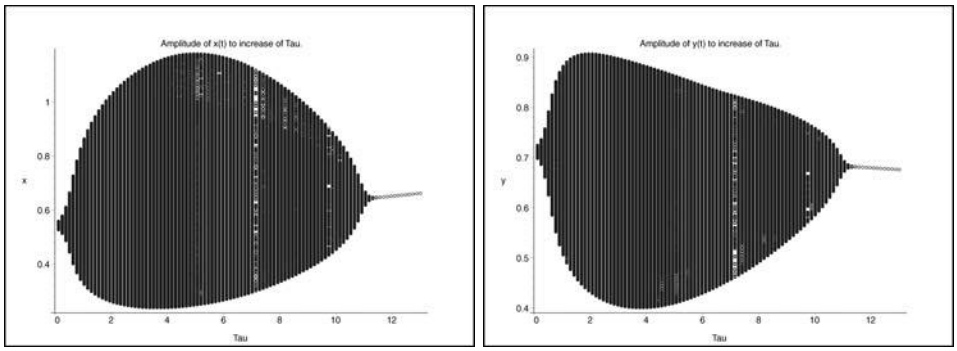


FIG. 5.2. *The ultimate oscillation interval of the solution to system* (1.4) *according to increase of predator maturation* τ, *where* $r = n = k_1 = 1$, $K = 1.6$, $b = 1.5$, $k_2 = 0.1$, $d = 0.5$, $d_j = 0.01$, $x(\theta) \equiv 0.7$, $y(\theta) \equiv 0.2$, $\theta \in [-\tau, 0]$.

From Figure 5.2, we see that if $\tau \in (0.2, 1)$ or $\tau > 12$, approximately, the vertical amplitudes of $x(t), y(t)$ are as small as a point, suggesting that E is asymptotically stable; if τ increases in the interval $(0.8, 6)$, approximately, the vertical amplitudes of $x(t), y(t)$ will becomes larger and larger, showing that E becomes more and more "unstable"; however, when τ smoothly increases in the interval $(6, 13)$, the points will become more concentrated and thus their amplitudes will become increasingly smaller until finally they gather into a point as $\tau > 11.83$, approximately. This shows that E becomes more stable as τ increases.

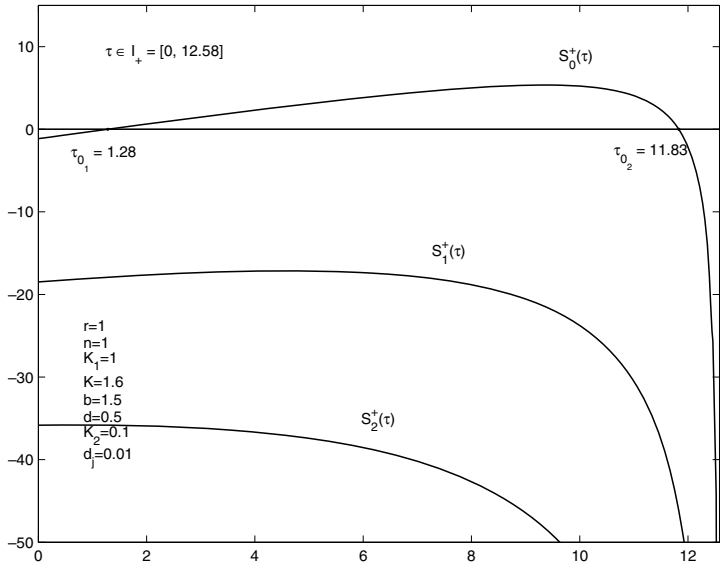


FIG. 5.3. *For the parameters choice* (5.26) *are depicted the curves* S_0^+, S_1^+, S_2^+ *in the interval of existence of* $\omega_+(\tau)$, *i.e.,* $\tau \in I_+ = [0, 12.58]$. *Only* S_0^+ *has two zeros* $\tau_{0_1} = 1.28$ *and* $\tau_{0_2}^+ = 11.83$.

Considering the case related to Figures 5.1 and 5.2, we have the following theorem.

THEOREM 5.7. *In system* (1.7), *let us choose the following parameters:*

$$r = n = k_1 = 1, \quad K = 1.6, \quad b = 1.5, \quad d = 0.5, \quad k_2 = 0.1, \quad d_j = 0.01. \tag{5.26}$$

Then system (1.7) *is asymptotically stable at* $\tau = 0$ *and remains asymptotically stable increasing* τ *up to the value* $\tau_{0_1}^+ = 1.28$, *which is a Hopf bifurcation value toward sustained oscillations around the positive equilibrium. The positive equilibrium remains unstable with sustained oscillations for* τ *up to the value* $\tau_{0_2}^+ = 11.83$, *at which there is a backward Hopf bifurcation value toward asymptotic stability. Whenever it exists the positive equilibrium remains asymptotically stable for* $\tau > \tau_{0_2}^+ = 11.83$.

Proof. For the proof, we have followed the algorithm presented previously in this section.

At $\tau = 0$ the roots of (5.8) are $\lambda = -0.0676 \pm i0.4581$ and therefore the positive equilibrium E is asymptotically stable. The equation (5.12) has only the positive root $\omega_+(\tau)$ in the interval $I_+ = [0, 12.58)$. According to the algorithm we have only the sequence $S_n^+(\tau)$, $\tau \in I_+$, $n \in \mathbf{N}_0$ of functions given by (5.23) (for which $S_n(\tau) > S_{n+1}(\tau)$ in I_+ for any $n \in \mathbf{N}_0$). By the algorithm presented for the characteristic equation (5.5), in Figure 5.3, we draw the curves S_0^+, S_1^+, S_2^+ in the interval of existence of $\omega_+(\tau)$. The $S_0^+(\tau)$ curve shows that for the parameter values (5.26), the function $S_0^+(\tau)$ has two zeros in I_+ the first at $\tau_{0_1} = 1.28$ and the second at $\tau_{0_2}^+ = 11.83$.

Thanks to Theorem 5.6, since in τ_{0_1} the slope of S_0^+ is positive, two pure imaginary roots $\lambda = \pm i\ \omega_+(\tau_{0_1})$ of (5.5) cross the imaginary axis entering in the right half complex plane and giving rise to two complex and conjugate roots with positive real part for $\tau > \tau_{0_1}$.

Hence the characteristic equation (5.5) has

(a) all roots with negative real parts if $\tau \in [0, \tau_{0_1})$;
(b) a pair of conjugate pure imaginary roots $\pm i\omega_+(\tau_{0_1})$, $\omega_+(\tau_{0_1}) > 0$, crossing the imaginary axis, and all other roots with negative real part if $\tau = \tau_{0_1}$;
(c) two roots with strictly positive real part if $\tau > \tau_{0_1}$ $(\tau < \tau_{0_2})$;
(d) and because of (b), all the roots λ $(\neq \pm i\ \omega_+(\tau_1))$ satisfy the condition $\lambda \neq i\ m\omega_+(\tau_1$, where m is any integer, if $\tau = \tau_{0_1}$.

Hence, at $\tau = \tau_{0_1}$ a Hopf bifurcation occurs (see [18, Chapter 11]).

Up to $\tau > \tau_{0_2}$ we have two complex and conjugate roots with positive real part giving rise to sustained oscillations. Since in τ_{0_2} the slope of S_0^+ is negative, two pure imaginary roots $\lambda = \pm i\omega_+(\tau_{0_2})$ of (5.5) cross the imaginary axis toward the left half complex plane and the total multiplicity of roots with positive real part returns to be zero for $\tau > \tau_{0_2}$. Similar to τ_{0_1} in τ_{0_2} we have another Hopf bifurcation toward asymptotic stability.

In conclusion, in $[0, \tau_{0_1})$ we have asymptotic stability of positive equilibrium E, in (τ_{0_1}, τ_{0_2}) sustained oscillations and in (τ_{0_2}, τ^*) asymptotic stability again. In τ_{0_1} there is a Hopf bifurcation toward sustained oscillations and in τ_{0_2} a Hopf bifurcation toward asymptotic stability. □

We observe that the outcomes of Theorem 5.7 are in agreement with the numeric simulations shown in Figure 5.1, where the solution of (1.7) is shown for the set of parameter values in (5.26). The agreement is the same for Figure 5.2.

For further analysis on the characteristic equation (5.5) it is worth noting that in the sequence $S_n,\ n \in \mathbf{N}_0$ (5.23), since $S_n(\tau) > S_{n+1}(\tau)$ for all $n \in \mathbf{N}$ and $\tau \in I_\omega$, the stability switches (Hopf bifurcations) occur only with the zeros of S_0. It is now interesting to study the role of predator interference coefficient k_2 on the stability of positive equilibrium E.

THEOREM 5.8. *In system* (1.7) *let us choose the following parameters:*

$$r = n = k_1 = 1, \quad K = 2.6, \quad b = 1.5,\ d = 0.5, \quad d_j = 0.01 \tag{5.27}$$

with varying k_2 *at the values* $k_2 = 0.2,\ 0.4,\ 0.6,\ 0.8$. *The positive equilibrium* E *undergoes stability switches from asymptotic stability to instability to asymptotic stability for increasing delay* τ *when* $k_2 = 0.2,\ 0.4,\ 0.6$ *and remains asymptotically stable for any* τ *when* $k_2 = 0.8$. *The delay instability interval has a decreasing width for increasing* k_2.

Proof. We have checked that parameter values (5.27), for each k_2 value, at $\tau = 0$ give rise to asymptotic stability of positive equilibrium E. By the algorithm presented for the characteristic equation (5.5), in Figure 5.4, we draw the curves S_0^+ versus τ for each k_2 value up to $k_2 = 0.7$. These curves S_0^+ correspond to the positive root $\omega_+(\tau)$ of (5.12) since for all k_2 values considered the root $\omega_-(\tau)$ of (5.12) is not feasible. For $k_2 > 0.7$ even the root $\omega_+(\tau)$ is not feasible, i.e., (5.12) does not have roots. Hence, no stability switches can occur for $k_2 > 0.7$, i.e., at $k_2 > 0.7$ the positive equilibrium E remains asymptotically stable for all $\tau \geq 0$ in its existence interval.

If we look at the curves S_0^+ in Figure 5.4, according to Theorem 5.6, we see that for each value $k_2 = 0.2,\ 0.4$, and 0.6 we have two stability switches, the first toward instability (since according to (5.24) at the first zero of $S_0{*}^+$ the slope of S_0^+ is positive) and the second toward asymptotic stability (since according to (5.24) at the second zero of S_0^+ the slope of S_0^+ is negative) with an unstable region in which sustained oscillations occur between the two stability switch delay values (Hopf bifurcations). Figure 5.4 shows that these delay instability regions have a decreasing

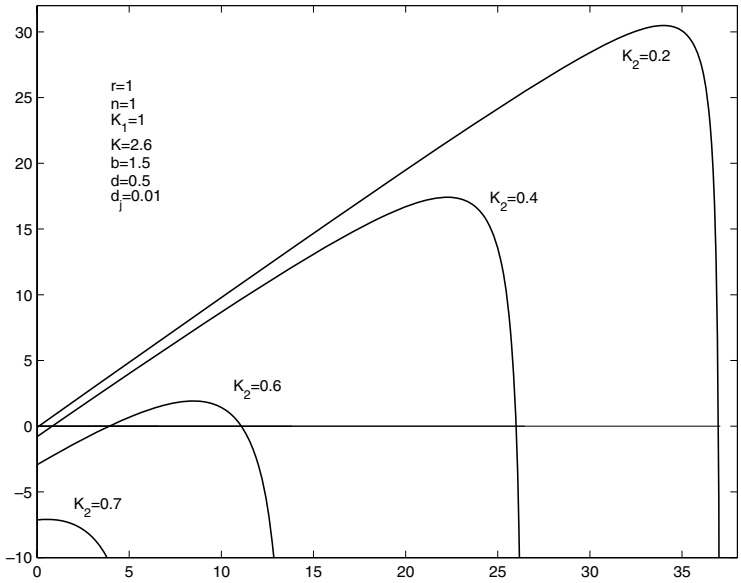


FIG. 5.4. *In the figure the curves* S_0^+ *are shown for the parameter values* (5.27) *and for the* k_2 *values* $k_2 = 0.2,\ 0.4,\ 0.6,\ 0.7$.

width for increasing values of k_2. At $k_2 = 0.7$ the curve S_0^+ has no zeros for $\tau \geq 0$ and therefore no stability switches can occur, i.e., the positive equilibrium E remains asymptotically stable for all τ in its existence interval. □

Figure 5.5, for the same parameters (5.27) of Theorem 5.8 and maturation time fixed at $\tau = 6$ show the behavior of $x(t), y(t)$ versus time at increasing values of k_2. At $k_2 = 0.2, 0.6$ we have sustained oscillations according to the fact that $\tau = 6$ falls within the instability intervals of S_0^+, whereas at $k_2 = 0.8$ S_0^+ is negative implying asymptotic stability of E.

For the parameter values (5.27) in Figure 5.6 we investigate the stabilizing role of predator interference coefficient k_2 on the positive equilibrium E by studying the zeros of the curves $S_0^+(\tau)$ for increasing values of k_2. The value of k_2, say, k_2^*, at which $S_0^+(\tau)$ has two coincident zeros (i.e., the delay instability region vanishes) is such that for $k_2 \geq k_2^*$ the positive equilibrium E does not undergo to stability switches and therefore E remains asymptotically stable in its existence interval $I = [0, \tau^*)$.

Figure 5.6 shows that $k_2^* = 0.6225$ when other parameters are fixed at values (5.27).

For the parameter values (5.27) and fixed maturation time delay $\tau = 6$ we try to have Figure 5.7 reflecting the above changes of stability of positive equilibrium E as k_2 increases from 0 to 0.9. For Figure 5.7, we see that if $k_2 \in (0, 0.64)$ the vertical amplitudes of $x(t),\ y(t)$ are oscillating showing that E is unstable. However, as k_2 smoothly increases in this interval, the points will become more concentrated and thus their amplitudes become smaller until finally, as that $k_2 > 0.64$ approximately, the vertical amplitudes of $x(t), y(t)$ will be small as a point, suggesting that E is asymptotically stable. This result is in good agreement with Figure 5.6.

The last case that we consider concerns Figure 5.8, which shows the interesting result that given enough large k_2 the predator and prey in system (1.4) will always coexist stably regardless how large K is.

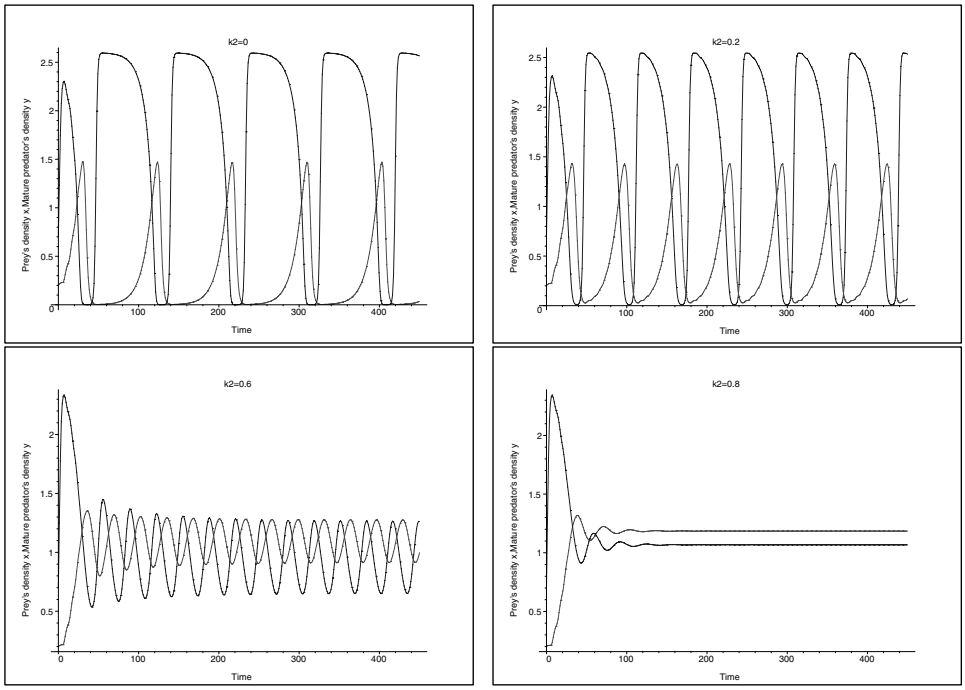


FIG. 5.5. *Solutions to system* (1.4) *with different predator interference coefficients* k_2*, here* $r = n = k_1 = 1$*,* $K = 2.6$*,* $b = 1.5$*,* $d = 0.5$*,* $\tau = 6$*,* $d_j = 0.01$*,* $x(\theta) \equiv 0.7$*,* $y(\theta) \equiv 0.2$*,* $\theta \in [-\tau, 0]$*.*

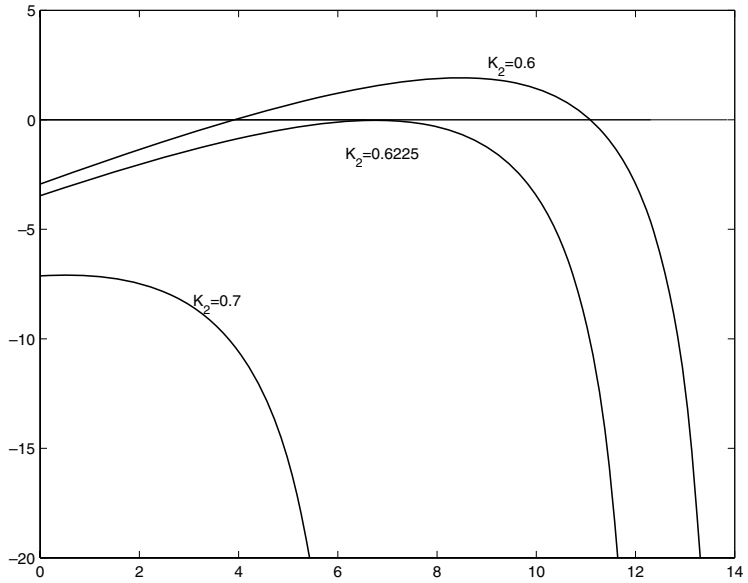


FIG. 5.6. *The zeros of the function* $S_0^+(\tau)$ *versus* τ *for increasing* k_2*. The figure shows that* $k_2^* = 0.6225$ *is the value of* k_2 *such that for* $k_2 \geq k_2^*$ *the positive equilibrium* E *remains asymptotically stable for* τ *in the interval* $[0, \tau^*)$*. The other parameter values are as in the set* (5.27)*.*

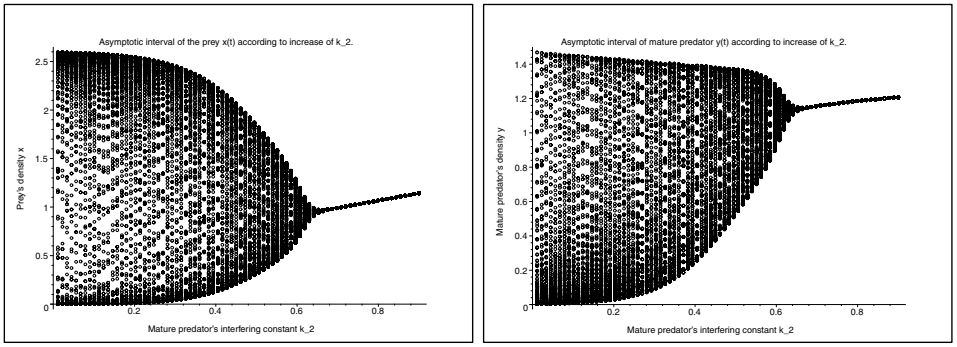


FIG. 5.7. *The ultimate oscillation interval of the solution to system* (1.4) *according to increase of predator interfering constant* k_2*, where* $r = n = k_1 = 1$*,* $K = 2.6$*,* $b = 1.5$*,* $d = 0.5$*,* $\tau = 6$*,* $d_j = 0.01$*,* $x(\theta) \equiv 0.7$*,* $y(\theta) \equiv 0.2$*,* $\theta \in [-\tau, 0]$.

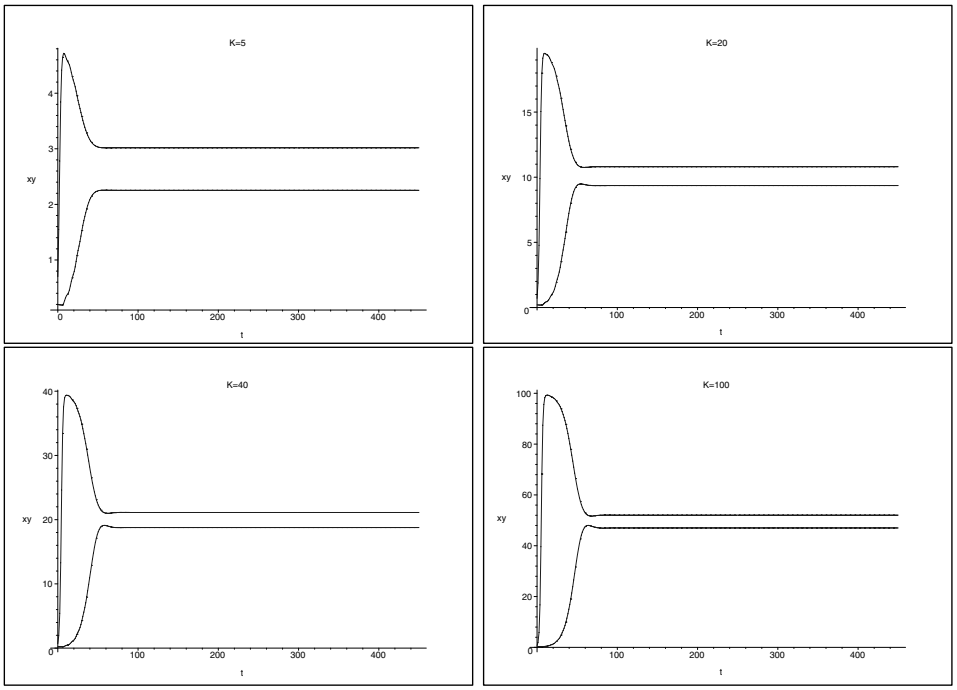


FIG. 5.8. *Solutions to system* (1.4) *with different carrying capacity* K*, here* $r = n = k_1 = 1$*,* $b = 1.5$*,* $d = 0.5$*,* $k_2 = 2$*,* $\tau = 6$*,* $d_j = 0.01$*,* $x(\theta) \equiv 0.7$*,* $y(\theta) \equiv 0.2$*,* $\theta \in [-\tau, 0]$.

The parameters are

$$r = n = k_1 = 1,\ b = 1.5,\ d = 0.5,\ k_2 = 2,\ d_j = 0.01 \tag{5.28}$$

with varying carrying capacity K at values $K = 5,\ 20,\ 40,\ 100$.

For each K we have checked that at $\tau = 0$ the roots of (5.8) have negative real part, i.e., at $\tau = 0$ the positive equilibrium is asymptotically stable.

Furthermore, (5.12) has no solutions in the existence interval $I = [0, \tau^*)$ of the positive equilibrium for each value of the considered carrying capacity. Hence, no sta-

bility switches can occur and the positive equilibrium E remains asymptotically stable in its existence interval for each considered value of carrying capacity, in agreement with the behavior shown by the solutions in Figure 5.8.

Figure 5.8 shows some interesting results: given enough large k_2, the predator and prey in system (1.4) will always coexist stably regardless how large K is. We have checked that the k_2 in Figure 5.8 does not satisfy (5.13); this shows that Theorem 5.3 has space to improve.

6. Discussion. In this paper, we study the stage-structured predator-prey model (1.4) of Beddington–DeAngelis-type functional response, which is an extension of both the ODE models studied by Cantrell and Cosner [9] and Hwang [23], [24] and the stage-structured predator-prey model of Holling II type functional response studied by Gourley and Kuang [15].

We give the conditions which are both necessary and sufficient for the permanence and extinction of system (1.4). Our results suggest that the predator coexists with prey permanently iff (2.2) holds true, i.e., predator's recruitment rate at the peak of prey abundance is larger than its death rate; and that the predator goes extinct iff $\frac{nbe^{-d_j\tau}K}{1+k_1K} \le d$ holds true, i.e., the predator's possible highest recruitment rate is less than or equal to its death rate. These results generalize the corresponding results in [9] and improve those in [15]. Comparing to the corresponding Theorem 3.1 in [9] for the ODE system (1.2), we find that there is an extra term $e^{-d_j\tau}$ in our permanence and extinction criteria, i.e., the surviving probability of each immature predator to become mature, which exists because of the stage structure. We can get that $d_j\tau$ has a negative effect on the persistence of the predator in that a proper increase of $d_j\tau$ (which is defined as the "degree of stage structure" by Liu et al. [28]) can directly destroy (2.2) and thus drive the predator into extinction, regardless how large $\frac{nbK}{1+k_1K}$ was.

On the other hand, even if (2.2) holds true, the proper increase of $d_j\tau$ may still cause the extinction of predators provided that demographic stochasticity were present: since y^* becomes smaller and smaller as $d_j\tau$ increases until $y^* = 0$ at $\frac{nbe^{-d_j\tau}K}{1+k_1K} = d$, then the predators will be at risk of stochastic extinction as the $d_j\tau$ holds the predator population to a sufficiently low level.

Therefore, the high death rate d_j or the long maturation of the juvenile predator τ may be responsible for the extinction of predators. These conclusions are analogous to those obtained for stage-structured predator-prey models of Holling I [35] and Holling II [15] type and similar to those for the stage-structured competitive system [28], [30], [31], [32], [2], [41].

We also find the stability switches of the interior equilibrium E due to the increase of τ: as τ increases, we see that oscillatory dynamics may appear and further increase of τ will return the oscillatory dynamics to the steady state form, implying that a large delay can be stabilizing.

Interesting results from this paper are the effects by the degree of predator interference k_2. First, we get that the permanence and extinction criteria for the stage structured BD model are independent of k_2. This shows that k_2 does not affect either the extinction or the permanence of the community, provided that the demographic stochasticity were ignored. Otherwise, as Cantrell and Cosner [9] argued for the non-stage-structured system (1.2), since $y^* \to 0$ as $k_2 \to \infty$, an increase of k_2 can lower y^* and may drive the predators into the risk of stochastic extinction when the predator population is at a sufficiently low level.

Second, k_2 can stabilize system (1.4):

(i) When E is unstable, having k_2 increasing from zero can reduce the oscillation amplitudes of solutions. This indicates that the interior equilibrium for the BD model is usually more stable than that for the corresponding H2 model.

(ii) Having a sufficiently large k_2 can directly drive E into the global attractiveness and the asymptotic stability.

Third, sufficiently large k_2 guarantees the robustness of the system against the increase of the carrying capacity K and the birth rate n of the adult predators. For the H2 model, Hsu, Hubbell, and Waltman [22, Lemma 4.5] showed that both the large carrying capacity K and the birth rate n of the predator can destabilize the positive equilibrium E and lead to the existence of periodic oscillation. Similar results can also be found in [25]. Hsu, Hubbell, and Waltman [21] even showed that an increase of K can cause E to be increasingly "unstable" by enlarging the amplitude of the limit cycle. However, we get completely different results for the BD model: Corollary 5.4 (or Corollary 5.5) shows that a sufficiently large k_2 can drive E stable regardless how large K (or n) is.

System (1.3) in [33] considers a gestation delay but not maturation delay of the BD model, while our model (1.4) accounts for maturation but not for gestation. Thus both (1.3) and (1.4) are the special cases of the following general model with both maturation and gestation effects:

(6.1)
$$\begin{cases} x'(t) = rx(t)\left(1 - \dfrac{x(t)}{K}\right) - \dfrac{bx(t)y(t)}{1 + k_1 x(t) + k_2 y(t)}, \\ y'(t) = \dfrac{nbe^{-d_j\tau}x(t-\tau-\tau_1)y(t-\tau)}{1 + k_1 x(t-\tau-\tau_1) + k_2 y(t-\tau-\tau_1)} - dy(t), \\ y_j'(t) = \dfrac{nbx(t-\tau_1)y(t)}{1+k_1 x(t-\tau_1)+k_2 y(t-\tau_1)} - \dfrac{nbe^{-d_j\tau}x(t-\tau-\tau_1)y(t-\tau)}{1+k_1 x(t-\tau-\tau_1)+k_2 y(t-\tau-\tau_1)} - d_j y_j(t), \\ x(\theta),\ y(\theta) \geq 0 \text{ are continuous on } -\tau \leq \theta \leq 0, \text{ and } x(0), y(0), y_j(0) > 0, \end{cases}$$

where τ, τ_1 are the maturation and the gestation delay, respectively. Then we have when $\tau = 0$, (6.1) becomes (1.3), and when $\tau_1 = 0$, (6.1) becomes (1.4). Thus system (6.1) unifies (1.3) and (1.4). It will be interesting for us to consider system (6.1); we leave this as our future work.

Acknowledgments. We are very grateful to two anonymous referees for their careful reading and very valuable comments, which led to an improvement of our original manuscript. We would like to thank Professor Horst Thieme, Dr. Dashun Xu and Dr. Yi Wang for the helpful discussions and Dr. Sanyi Tang, Dr. Yanni Xiao, and Dr. Margherita Carletti for their warm help on the numeric simulations. The first author acknowledges the financial support of Istituto di Biomatematica at Universitá di Urbino.

REFERENCES

[1] W. Aiello and H. I. Freedman, *A time-delay model of single-species growth with stage structure,* Math. Biosci., 101 (1990), pp. 139–153.

[2] J. Al-Omari and S. Gourley, *Stability and traveling fronts in Lotka-Volterra competition models with stage structure,* SIAM J. Appl. Math., 63 (2003), pp. 2063–2086.

[3] R. Arditi and L. R. Ginzburg, *Coupling in predator-prey dynamics: Ratio-dependence,* J. Theoret. Biol., 1989 (139), pp. 311–326.

[4] P. A. Abrams and C. J. Walters, *Invulnerable prey and the statics and dynamics of predator-prey interaction,* Ecology, 1996 (77), pp. 1125–1133.
[5] P. A. Abrams and L. R. Ginzburg, *The nature of predation: Prey dependent, ratio dependent or neither?* TREE, 2000 (15), pp. 337–341.
[6] J. R. Beddington, *Mutual interference between parasites or predators and its effect on searching efficiency,* J. Animal Ecol., 44 (1975), pp. 331–340.
[7] E. Beretta and Y. Kuang, *Global analysis in some delayed ratio-dependent predator-prey systems,* Nonlinear Anal., 32 (1998), pp. 381–408.
[8] E. Beretta and Y. Kuang, *Geometric stability switch criteria in delay differential systems with delay dependent parameters,* SIAM J. Math. Anal., 33 (2002), pp. 1144–1165.
[9] R. S. Cantrell and C. Cosner, *On the dynamics of predator-prey models with the Beddington–DeAngelis functional response,* J. Math. Anal. Appl., 257 (2001), pp. 206–222.
[10] R. S. Cantrell, C. Cosner, and S. G. Ruan, *Intraspecific interference and consumer-resource dynamics*, Dyn. Syst. Ser. B, 4 (2004), pp. 527–546.
[11] C. Cosner, D. L. DeAngelis, J. S. Ault, and D. B. Olson, *Effects of spatial grouping on the functional response of predators,* Theoret. Popul. Biol., 56 (1999), pp. 65–75.
[12] P. H. Crowley and E. K. Martin, *Functional responses and interference within and between year classes of a dragonfly population,* J. North American Benthological Soc., 8 (1989), pp. 211–221.
[13] D. L. DeAngelis, R. A. Goldstein, and R. Neill, *A model for trophic interaction,* Ecology, 56 (1975), pp. 881–892.
[14] M. Fan and Y. Kuang, *Dynamics of a nonautonomous predator prey system with the Beddington–DeAngelis functional response,* J. Math. Anal. Appl., 295 (2004), pp. 15–39.
[15] S. A. Gourley and Y. Kuang, *A stage structured predator-prey model and its dependence on through-stage delay and death rate*, J. Math. Biol., 49 (2004), pp. 188–200.
[16] J. K. Hale and P. Waltman, *Persistence in infinite-dimensional systems,* SIAM J. Math. Anal., 20 (1989), pp. 388–395.
[17] M. P. Hassell and C. C. Varley, *New inductive population model for insect parasites and its bearing on biological control,* Nature, 223 (1969), pp. 1133–1137.
[18] J. K. Hale and S. V. Lunel, *Introduction to Functional Differential Equations,* Appl. Math. Sci. 99, Springer-Verlag, New York, 1993.
[19] C. S. Holling, *The components of predation as revealed by a study of small mammal predation of the European pine sawfly,* Canad. Entomologist, 91 (1959), pp. 293–320.
[20] C. S. Holling, *Some characteristics of simple types of predation and parasitism,* Canad. Entomologist, 91 (1959), pp. 385–395.
[21] S. B. Hsu, S. P. Hubbell, and P. Waltman, *A contribution to the theory of competing predators,* Ecological Monogr., 35 (1978), pp. 617–625.
[22] S. B. Hsu, S. P. Hubbell, and P. Waltman, *Competing predators,* SIAM J. Appl. Math., 35 (1978), pp. 617–625.
[23] Z. W. Hwang, *Global analysis of the predator-prey system with Beddington–DeAngelis functional response,* J. Math. Anal. Appl., 281 (2003), pp. 395–401.
[24] Z. W. Hwang, *Uniqueness of limit cycles of the predator-prey system with Beddington–DeAngelis functional response,* J. Math. Anal. Appl., 290 (2004), pp. 113–122
[25] Y. Kuang, *Global stability of Gause-type predator-prey systems,* J. Math. Biol., 28 (1990), pp. 463–474.
[26] Y. Kuang, *Delay Differential Equations with Applications in Population Dynamics*, Academic Press, New York, 1993.
[27] Y. Kuang and E. Beretta, *Global qualitative analysis of a ratio-dependent predator-prey system*, J. Math. Biol., 36 (1998), pp. 389–406.
[28] S. Liu, L. Chen, G. Luo, and Y. Jiang, *Asymptotic behavior of competitive Lotka-Volterra system with stage structure,* J. Math. Anal. Appl., 271 (2002), pp. 124–138.
[29] S. Liu, L. Chen, and R. Agarwal, *Recent progress on stage-structured population dynamics*, Math. Comput. Modelling, 36 (2002), pp. 1319–1360.
[30] S. Liu, L. Chen, and G. Luo, *Extinction and permanence in competitive stage structured system with time delays*, Nonlinear Anal., 51 (2002), pp. 1347–1361.
[31] S. Liu, L. Chen, and Z. Liu, *Extinction and permanence in nonautonomous competitive system with stage structure,* J. Math. Anal. Appl., 274 (2002), pp. 667–684.
[32] S. Liu, M. Kouche, and N. Tatar, *Permanence and global asymptotic stability in a stage structured system with distributed delays,* J. Math. Anal. Appl., 301 (2005), pp. 187–207.
[33] Z. Liu and R. Yuan, *Stability and bifurcation in a delayed predator-prey system with Beddington–DeAngelis functional response*, J. Math. Anal. Appl., 296 (2004), pp. 521–537.
[34] A. Martin and S. Ruan, *Predator-prey models with delay and prey harvesting,* J. Math. Biol., 43 (2001), pp. 247–267.

[35] L. Ou, G. Luo, Y. Jiang, and Y. Li, *The asymptotic behaviors of a stage-structured autonomous predator-prey system with time delay,* J. Math. Anal. Appl., 283 (2003), pp. 534–548.
[36] Z. P. Qiu, J. Yu, and Y. Zou, *The asymptotic behavior of a chemostat model with the Beddington–DeAngelis functional response,* Math. Biosci., 187 (2004), pp. 175–187.
[37] J. D. Reeve, *Predation and bark beetle longterm dynamics,* Oecologia, 112 (1997), pp. 48–54.
[38] G. Ruxton, W. S. C. Gurney, and A. DeRoos, *Interference and generation cycles,* Theoret. Population Biol. 42 (1992), pp. 235–253.
[39] G. T. Skalski and J. F. Gilliam, *Functional responses with predator interference: Viable alternatives to the Holling type* II *model,* Ecology, 82 (2001), pp. 3083–3092.
[40] H. R. Thieme, *Persistence under relaxed point-dissipativity (with application to an endemic model),* SIAM J. Math. Anal., 24 (1993), pp. 407–435.
[41] W. Wang, G. Mulone, F. Salemi, and V. Salone, *Permanence and stability of a stage-structured predator-prey model,* J. Math. Anal. Appl., 262 (2001), pp. 499–528.

SIAM J. Appl. Math.
Vol. 66, No. 4, pp. 1130–1149

ON THE RELATIONSHIP BETWEEN SUPPLEMENTAL BALANCES IN TWO THEORIES FOR PURE INTERFACE MOTION*

ELIOT FRIED†

Abstract. A matched asymptotic analysis is used to exhibit the connection between supplemental balance equations arising in sharp-interface and phase-field theories for transformations between two rigid phases distinguished only by their constant free-energy densities. The analysis exposes the relationship between the forces and balances arising in the two theories.

Key words. configurational forces, phase transformations, sharp-interface theory, phase-field theory

AMS subject classifications. 74N20, 80A22, 82C26, 45M05, 58J37, 35L65

DOI. 10.1137/050632890

1. Introduction.

The purpose of this paper is to examine the question of whether the configurational force balance can always be viewed merely as a rephrasing of the standard force balance. We address this question in the setting of a transformation between two rigid phases distinguished only by their constant free-energy densities. Because the phases are rigid, standard forces are extraneous and their balance is satisfied trivially. Nevertheless, configurational forces are essential and their balance is not implied by the standard force balance.

Standard forces are associated with the motion of material particles. Configurational forces first arose in the works of Peach and Koehler [29] and Eshelby [8, 9] on lattice defects and of Herring [25] on the sintering of powders. Beginning with appropriate energy functionals, those works derive configurational forces variationally by considering rearrangements of the relevant defects. The distinction between configurational forces and standard forces is evident from the derivations: whereas standard forces arise from variations in the placement of material particles, configurational forces arise from variations in the arrangement, relative to material particles, of nonmaterial defects. These derivations also show that, when deformation is taken into account, the necessary conditions for equilibrium in a defective medium include not only the Euler–Lagrange equations imposing the balance of standard forces at and away from defects but also an additional Euler–Lagrange equation valid at defects and involving configurational forces. Although the variations leading to these conditions are performed independently, the ensuing equilibrium conditions are generally coupled.

Aside from giving conditions for the description of equilibrium, the variational approach provides guidance on how configurational forces should enter the description of dissipative processes involving defect generation and evolution. Indeed, the conventional generalization of a variationally based theory for a defective medium involves replacing the relevant Euler–Lagrange equation with a gradient-flow equation requiring that the time-rate of the kinematical entity describing the configuration of the defect be proportional to the associated configurational force, with the sign of

*Received by the editors June 1, 2005; accepted for publication (in revised form) December 15, 2005; published electronically March 17, 2006.
http://www.siam.org/journals/siap/66-4/63289.html
†Department of Mechanical and Aerospace Engineering, Washington University in St. Louis, St. Louis, MO 63130-4899 (efried@me.wustl.edu).

the constant of proportionality assigned to rule out spurious growth of the underlying energy functional. In that setting, the rate term can be viewed as a configurational drag force that accounts for energy dissipation associated with the motion of defects relative to the underlying material.

The structure of classical theories of continua allows for a clear distinction to be drawn between basic laws and constitutive equations. Whereas the basic laws hold for large classes of materials, constitutive equations distinguish between different types of materials. However, because Euler–Lagrange and gradient-flow arguments rest on the provision of constitutive equations, the physical status of any supplemental equations they engender is unclear. Do such equations represent additional balances, above and beyond that involving standard forces, or do they simply represent additional constitutive information?

Commencing with a series of papers (Gurtin [18]; Angenent and Gurtin [2]; Gurtin and Struthers [23]) concerning phase transformations, Gurtin advocates the first of the alternative interpretations stated above. Those papers take a Gibbsian approach: phase interfaces are modeled as sharp nonmaterial surfaces across which the properties of the bulk phases may suffer discontinuities; to account for localized interactions between the phases, those surfaces are endowed with excess fields. Briefly, Gurtin's innovation centers on his treatment of configurational forces as primitive objects that expend power in conjunction with the motion of defects (relative to the underlying material) and are subject to a configurational balance distinct from and supplemental to that involving standard forces. The question of whether configurational forces are necessary for the description of three-dimensional bodies containing lower-dimensional defects has been addressed by Podio-Guidugli [31], who shows that when only standard forces are taken into account, the reasonable requirement that the power expended on a migrating referential control volume be invariant under changes of the tangential, and thus extrinsic, component of the velocity that describes the motion of the volume has a consequence that is generally untenable: the standard stress must be a pressure. To avoid that, it is necessary and sufficient to account for power expenditures above and beyond those associated with standard forces. That can be accomplished with the introduction of configurational forces. Gurtin's approach has been applied to the description of defect structures other than interfaces, including cracks (Gurtin and Podio-Guidugli [21, 22]; Gurtin and Shvartsman [24]; Kalpakides and Dascalu [26]), edges and junctions (Simha and Bhattacharya [32]), dislocations (Cermelli and Gurtin [7]), plasticity (Cermelli, Fried, and Sellers [4]), liquid-crystalline disclinations (Cermelli and Fried [5]), epitaxy (Gurtin and Jabbour [17]; Fried and Gurtin [15, 16], nematic-isotropic transformations in liquid crystals (Cermelli, Fried, and Gurtin [6]), and fluid-fluid phase transformations (Anderson, Cermelli, Fried, Gurtin and McFadden [1]). A comprehensive treatment of configurational forces and their applications is given in Gurtin's book [20].

Within Gurtin's framework, the distinction between the balances for configurational and standard forces is not merely an efficiency in capturing singularities. To illustrate this point, consider a setting involving a sharp interface $\mathcal{S}$ separating two phases, say α and β. For simplicity, neglect deformation, heat transport, and mass transport. Suppose that the free-energy density of phase $\gamma = \alpha, \beta$ is a constant, say Ψ_γ. Consider the problem of developing a theory that accounts for dependence of the interfacial free-energy density on the interfacial orientation and for dissipation associated with the growth of one phase at the expense of another. As shown by Gurtin [19, 20], such a theory involves a single equation governing the evolution of $\mathcal{S}$. Writing $\mathbf{n}$ for the unit orientation of $\mathcal{S}$, directed from the region occupied by phase-α

into the region occupied by phase-β, and $V_{\mathcal{S}}$ for the (scalar) normal velocity of $\mathcal{S}$ in the direction of $\mathbf{n}$, that equation is

$$\hat{b}_{\mathcal{S}}(\mathbf{n}, V_{\mathcal{S}})V_{\mathcal{S}} = \left\{\hat{\psi}_{\mathcal{S}}(\mathbf{n})\mathbf{P} + \frac{\partial^2 \hat{\psi}_{\mathcal{S}}(\mathbf{n})}{\partial \mathbf{n}^2}\right\} \cdot \mathbf{L} + [\![\Psi]\!], \tag{1.1}$$

where $\mathbf{P} = \mathbf{1} - \mathbf{n} \otimes \mathbf{n}$ is the interfacial projector, $\mathbf{L} = -\nabla_{\mathcal{S}}\mathbf{n}$ is the interfacial curvature tensor, $\hat{\psi}_{\mathcal{S}}$ is the free-energy per unit interfacial area, $\hat{b}_{\mathcal{S}}$ is the nonnegative kinetic modulus, and $[\![\Psi]\!] = \Psi_\beta - \Psi_\alpha$. Generally, the dependence of $\hat{\psi}_{\mathcal{S}}$ on $\mathbf{n}$ renders certain interfacial orientations more energetically favorable than others. Similarly, the dependence of $\hat{b}_{\mathcal{S}}$ on $\mathbf{n}$ allows for growth at different rates along different orientations. Further, the dependence of $\hat{b}$ on $V_{\mathcal{S}}$ allows for nonlinear growth kinetics. The nonnegativity of $\hat{b}_{\mathcal{S}}$ ensures satisfaction of the second law. If $\hat{\psi}_{\mathcal{S}}(\mathbf{n}) = \psi_{\mathcal{S}}$ and $\hat{b}_{\mathcal{S}}(\mathbf{n}, V_{\mathcal{S}}) = b_{\mathcal{S}}$ with $\psi_{\mathcal{S}}$ and $b_{\mathcal{S}}$ constant, the evolution equation (1.1) then reduces to

$$b_{\mathcal{S}}V_{\mathcal{S}} = \psi_{\mathcal{S}}K_{\mathcal{S}} + [\![\Psi]\!], \tag{1.2}$$

with $K_{\mathcal{S}} = \operatorname{tr}\mathbf{L} = -\operatorname{div}_{\mathcal{S}}\mathbf{n}$ (twice) the mean curvature. When $\Psi_\alpha = \Psi_\beta$, which would be the case for an interface separating two grains of a crystal, (1.2) reduces to $b_{\mathcal{S}}V_{\mathcal{S}} = \psi_{\mathcal{S}}K_{\mathcal{S}}$, the two-dimensional specialization of which was first proposed by Mullins [28] as a model for grain-boundary evolution. The two-dimensional version of (1.1), with $b_{\mathcal{S}}$ independent of $V_{\mathcal{S}}$, was proposed by Uhuwa [34]. The general equation (1.1) was first given by Gurtin [18]. A formulation of (1.1) using a variational definition of the curvature term is given by Taylor, Cahn, and Handwerker [33], who provide background and extensive references.

Within Gurtin's theory, the evolution equation (1.1) arises from the normal component of the interfacial configurational force balance

$$\operatorname{div}_{\mathcal{S}}\mathbf{C} + \mathbf{f} + [\![\mathbf{C}]\!]\mathbf{n} = \mathbf{0}, \tag{1.3}$$

in conjunction with the representations

$$\mathbf{C}_\gamma = \Psi_\gamma \mathbf{1} \qquad \text{and} \qquad \mathbf{C} = \psi_{\mathcal{S}}\mathbf{P} - \mathbf{n} \otimes \mathbf{c} \tag{1.4}$$

for the bulk and interfacial configurational stresses and constitutive relations

$$\psi_{\mathcal{S}} = \hat{\psi}_{\mathcal{S}}(\mathbf{n}), \qquad \mathbf{c} = -\frac{\partial \hat{\psi}_{\mathcal{S}}(\mathbf{n})}{\partial \mathbf{n}}, \qquad \text{and} \qquad \mathbf{f} \cdot \mathbf{n} = -\hat{b}_{\mathcal{S}}(\mathbf{n}, V_{\mathcal{S}})V_{\mathcal{S}}, \tag{1.5}$$

with $\hat{b}_{\mathcal{S}} \geq 0$, that determine the interfacial free-energy density $\psi_{\mathcal{S}}$, the interfacial configurational shear $\mathbf{c}$, and the normal component $\mathbf{f} \cdot \mathbf{n}$ of the internal interfacial configurational body force density $\mathbf{f}$.

Since the theory described above neglects deformation, the phases are rigid. Standard stresses are therefore indeterminate both in bulk and on the interface. The standard force balance is therefore extraneous. It is therefore difficult to conceive of how the interfacial configurational force balance (1.3) or (its consequence) the interfacial evolution equation (1.1) could be an expression of standard force balance. Nevertheless, because there do exist circumstances in which the configurational force balance is a derived consequence of the standard force balance, some researchers contend that the configurational force balance can *always* be viewed as a rephrasing of the standard force balance.

To eliminate that confusion, we use an alternative method to derive the interfacial evolution equation (1.1). Our approach involves considering a theory in which the phases are described by a phase field φ. In that theory, an interface is not a surface but is, rather, a transition layer across which φ varies smoothly. The thickness of layers is constitutively determined. We consider a version of the phase-field theory that, due to a special choice of constitutive equations and a special scaling, permits us to control the thickness of transition layers. We then investigate the ramifications of shrinking that thickness. The phase-field theory allows for two approaches to deriving sharp-interface equations. We refer to those approaches as "direct" and "indirect." While the direct and indirect approaches yield the same analytical results, they afford different insights. We illustrate the indirect approach in the simple case where the desired interfacial evolution equation is (1.2). In the indirect approach, which follows the work of Caginalp [3], that equation arises as a solvability condition imposed by the Fredholm alternative on the inner expansion of φ. This result renders problematic any attempt to interpret (1.2) as an expression of standard force balance, but otherwise leaves open the question of just what physical law underlies (1.2). The direct approach, which involves the configurational force balance of the phase-field theory, sheds light on that question. The smoothness of the phase field makes consideration of configurational forces or their balance unnecessary. Nevertheless, a configurational force balance can be derived within the phase-field theory, and considerations based on that balance prove to be useful. Specifically, we work with the component of the configurational force balance normal to time-dependent level sets of φ. In the direct approach, (1.1) arises by expanding and integrating that equation over a layer while simultaneously shrinking the thickness of that layer to zero. This shows clearly that the interfacial evolution equation (1.1) of the sharp-interface theory is an expression of configurational force balance and, bearing in mind that deformation is neglected, verifies that (1.1) is unrelated to the standard force balance.

The paper is organized as follows. We begin, in section 2, with a brief overview of the phase-field theory. In so doing, we present both the standard variational derivation, which yields the governing equation for φ as a gradient-flow equation, and the less conventional continuum-mechanical derivation due to Fried and Gurtin [12]. Next, in section 3, we derive the configurational force balance germane to the phase-field theory. Here, again, we consider two approaches. In the first approach, we mimick Maugin's [27] approach to the study of configurational forces associated with material inhomogeneities in elastic solids. Specifically, we multiply the evolution equation for φ by $\nabla\varphi$ and arrive at the desired result in a few simple steps. The second approach employs Gurtin's [19, 20] general framework for configurational forces. In section 4, we consider time-dependent level sets of φ and obtain the component of the evolution equation for φ normal to those sets. In section 5, we discuss the role of standard forces. In section 6, we specialize the constitutive equations of the phase-field theory to yield an unscaled version of the theory that leads to the simple sharp-interface equation (1.2). Section 7 is concerned with scaling. In section 8, we discuss expansions. In section 9, we obtain asymptotic results for the regions occupied by the bulk phases. In section 10, we obtain asymptotic results for a generic transition layer. In so doing, we first take the indirect approach and then take the direct approach. In section 11, we generalize the constitutive assumptions imposed in section 4 and derive (1.1) using only the direct approach. Finally, in section 12, we conclude with a brief discussion.

2. Phase-field theory. We present a simple theory for transformations between two phases as described by a dimensionless scalar-valued *phase field* φ. Intuitively, at an instant when both phases are present, we expect φ to vary smoothly between distinct values associated with each of the phases. Further, letting L denote a suitable characteristic length, the product $L|\nabla\varphi|$ should exhibit large values in any zone connecting the two phases. Otherwise, at an instant when only one phase is present, φ should be essentially uniform. Thus, φ can be thought of as a regularized characteristic function for one of the phases, and phase transformations are embodied in the evolution of the phase distribution, as described by φ.

2.1. Variational approach. The conventional approach to developing an equation governing the evolution of φ is variational. Assuming that the free-energy density, say ψ, is determined constitutively as a function $\hat{\psi}$ depending on φ and, to account for energetic contributions from the zones connecting the two phases, on $\nabla\varphi$, the total free energy of the body $\mathcal{B}$ is then given by the functional

$$\mathcal{F}(\varphi) = \int_{\mathcal{B}} \hat{\psi}(\varphi, \nabla\varphi)\, dv. \tag{2.1}$$

The evolution equation for φ then has the form of a gradient-flow equation,

$$\beta(\varphi, \nabla\varphi, \dot{\varphi})\dot{\varphi} = -\frac{\delta\mathcal{F}(\varphi)}{\delta\varphi}, \tag{2.2}$$

with $\beta \geq 0$ a constitutively determined *kinetic coefficient* and $\delta\mathcal{F}(\varphi)/\delta\varphi$ defined via the first variation of $\mathcal{F}$, viz.,

$$\frac{\delta\mathcal{F}(\varphi)}{\delta\varphi} = \frac{\partial\hat{\psi}(\varphi, \nabla\varphi)}{\partial\varphi} - \operatorname{div}\left\{\frac{\partial\hat{\psi}(\varphi, \nabla\varphi)}{\partial(\nabla\varphi)}\right\}. \tag{2.3}$$

Tacit to the foregoing discussion is an understanding that, to encompass the existence of two energetically viable phases, the restriction $\hat{\psi}(\cdot, \mathbf{0})$ of $\hat{\psi}$ to homogeneous values of φ should be a double-well potential.

2.2. Alternative approach. An alternative to the variational approach shown is provided by Fried and Gurtin [12]. That alternative hinges on distinguishing between kinematical ingredients, laws of balance and imbalance, and constitutive equations. The phase field φ is the sole kinematical variable of the theory. Under the recognition that power expenditures should accompany temporal variations of any kinematical descriptor and that such expenditures must involve conjugate forces, a vector-valued *microstress* $\boldsymbol{\xi}$ and a scalar-valued *internal microforce density* π are introduced. The basic laws of the theory then consist of the *balance of microforces* and the *imbalance of free energy*, which require that for each body part $\mathcal{P}$, with boundary $\partial\mathcal{P}$ and outward unit normal $\boldsymbol{\nu}$,

$$\int_{\partial\mathcal{P}} \boldsymbol{\xi}\cdot\boldsymbol{\nu}\, da + \int_{\mathcal{P}} \pi\, dv = 0 \tag{2.4}$$

and

$$\overline{\int_{\mathcal{P}} \psi\, dv}^{\cdot} \leq \int_{\partial\mathcal{P}} (\boldsymbol{\xi}\cdot\boldsymbol{\nu})\dot{\varphi}\, da. \tag{2.5}$$

The local equivalents of the global laws are the field equation

$$\operatorname{div}\boldsymbol{\xi} + \pi = 0 \tag{2.6}$$

and the free-energy inequality

$$\dot{\psi} + \pi\dot{\varphi} - \boldsymbol{\xi}\cdot\nabla\dot{\varphi} \leq 0. \tag{2.7}$$

Assuming that ψ, $\boldsymbol{\xi}$, and π are determined constitutively by smooth functions of φ, $\nabla\varphi$, and $\dot{\varphi}$, and requiring those functions to be consistent with (2.7) in all processes, then gives

$$\psi = \hat{\psi}(\varphi, \nabla\varphi), \quad \boldsymbol{\xi} = \frac{\partial\hat{\psi}(\varphi,\nabla\varphi)}{\partial(\nabla\varphi)}, \quad \text{and} \quad \pi = -\frac{\partial\hat{\psi}(\varphi,\varphi}{\partial\varphi} - \beta(\varphi,\nabla\varphi,\dot{\varphi})\dot{\varphi} \tag{2.8}$$

with $\beta \geq 0$. Finally, using (2.8) in the local microforce balance (2.6) yields

$$\beta(\varphi,\nabla\varphi,\dot{\varphi})\dot{\varphi} = \operatorname{div}\left\{\frac{\partial\hat{\psi}(\varphi,\nabla\varphi)}{\partial(\nabla\varphi)}\right\} - \frac{\partial\hat{\psi}(\varphi,\nabla\varphi)}{\partial\varphi}, \tag{2.9}$$

which is identical to the evolution equation arising from (2.2) and (2.3).

In view of (2.8), the free-energy inequality (2.7) yields an expression

$$\delta = -\beta(\varphi,\nabla\varphi,\dot{\varphi})\dot{\varphi}^2 \tag{2.10}$$

for the rate at which energy is dissipated per unit volume.

3. Configurational forces and their balance.

3.1. Formal approach. Consider the evolution equation (2.9). Multiplying each term of that equation by $\nabla\varphi$ and performing a few simple manipulations, one is led to the identity

$$\operatorname{div}\left\{\hat{\psi}(\varphi,\nabla\varphi)\mathbf{1} - \nabla\varphi\otimes\frac{\partial\hat{\psi}(\varphi,\nabla\varphi)}{\partial(\nabla\varphi)}\right\} + \beta(\varphi,\nabla\varphi,\dot{\varphi})\dot{\varphi}\nabla\varphi = \mathbf{0}. \tag{3.1}$$

The tensor

$$\hat{\psi}(\varphi,\nabla\varphi)\mathbf{1} - \nabla\varphi\otimes\frac{\partial\hat{\psi}(\varphi,\nabla\varphi)}{\partial(\nabla\varphi)} \tag{3.2}$$

appearing in (3.1) is immediately recognizable as the *configurational stress tensor* relevant to the present context (Eshelby [10]). Further, the vector

$$\beta(\varphi,\nabla\varphi,\dot{\varphi})\dot{\varphi}\nabla\varphi \tag{3.3}$$

represents a *configurational body force density*. Thus, the derived identity (3.1) can be viewed as a *configurational force balance* associated with the evolution equation (2.9). In the absence of defects, which would be associated with irregularities of φ, (3.1) is equivalent to (2.9) whenever $\nabla\varphi$ is nontrivial and, thus, superfluous.

Within the context of the phase-field theory, the configurational force balance (3.1) is a consequence of microforce balance (2.6) and the thermodynamically derived constitutive equations (2.8). In particular, neither standard forces nor their balance enter the derivation. Hence, (3.1) is unrelated to standard force balance.

3.2. Alternative approach. Like the variational derivation of (2.9), the above derivation of (3.1) is predicated on the provision of constitutive equations. An alternative derivation that is free from that restriction is due to Gurtin [19, 20]. That approach treats configurational forces as basic entities that are associated with the integrity of the material structure of a body and expend power in connection with the transfer of material and the evolution of defects. Specifically, a *configurational stress tensor* $\mathbf{C}$ and a *configurational body force density* $\mathbf{f}$ are introduced. These fields are required to satisfy the *configurational force balance*

$$\int_{\partial\mathcal{P}} \mathbf{C}\boldsymbol{\nu}\, da + \int_{\mathcal{P}} \mathbf{f}\, dv = \mathbf{0} \tag{3.4}$$

for each part $\mathcal{P}$ of $\mathcal{B}$, which is equivalent to the local configurational force balance

$$\operatorname{div}\mathbf{C} + \mathbf{f} = \mathbf{0}. \tag{3.5}$$

To characterize the manner in which configurational forces expend power, a means of capturing the kinematics associated with the transfer of material is needed. Gurtin [19, 20] accomplishes this with the aid of migrating control volumes. The evolution of a migrating control volume $\mathcal{R}$ can be generically described by a time-dependent field $\mathbf{q}$ defined over $\partial\mathcal{R}$, and the configurational traction $\mathbf{C}\boldsymbol{\nu}_{\partial\mathcal{R}}$ is assumed to be power conjugate to $\mathbf{q}$. Further, to properly reckon the power expended by the microtraction $\boldsymbol{\xi}\cdot\boldsymbol{\nu}_{\partial\mathcal{R}}$ on $\mathcal{R}$, it is necessary to consider the convected time-rate $\dot{\varphi}+\nabla\varphi\cdot\mathbf{q}$ of φ following the motion of $\partial\mathcal{R}$. The net power expended on $\mathcal{R}$ by external agencies can then be expressed as

$$\int_{\partial\mathcal{R}} \{(\mathbf{C}+\nabla\varphi\otimes\boldsymbol{\xi})\boldsymbol{\nu}_{\partial\mathcal{R}}\cdot\mathbf{q} + (\boldsymbol{\xi}\cdot\boldsymbol{\nu}_{\partial\mathcal{R}})\dot{\varphi}\}\, da. \tag{3.6}$$

Since the intrinsic motion of $\partial\mathcal{R}$ involves only the normal component $\mathbf{q}\cdot\boldsymbol{\nu}_{\partial\mathcal{R}}$ of $\mathbf{q}$, the net power should be invariant with respect to the choice of the tangential component of $\mathbf{q}$. This invariance implies that $\mathbf{C}+\nabla\varphi\otimes\boldsymbol{\xi}=\alpha\mathbf{1}$ and, thus, that

$$\int_{\partial\mathcal{R}} \{(\mathbf{C}+\nabla\varphi\otimes\boldsymbol{\xi})\boldsymbol{\nu}_{\partial\mathcal{R}}\cdot\mathbf{q} + (\boldsymbol{\xi}\cdot\boldsymbol{\nu}_{\partial\mathcal{R}})\dot{\varphi}\}\, da = \int_{\partial\mathcal{R}} (\alpha\mathbf{q}\cdot\boldsymbol{\nu}_{\partial\mathcal{R}} + (\boldsymbol{\xi}\cdot\boldsymbol{\nu}_{\partial\mathcal{R}})\dot{\varphi})\, da. \tag{3.7}$$

In view of the foregoing discussion, the free-energy imbalance for a migrating control volume $\mathcal{R}$ can then be expressed in the form

$$\dot{\int_{\mathcal{R}} \psi\, dv} \le \int_{\partial\mathcal{R}} (\alpha\mathbf{q}\cdot\boldsymbol{\nu}_{\partial\mathcal{R}} + (\boldsymbol{\xi}\cdot\boldsymbol{\nu}_{\partial\mathcal{R}})\dot{\varphi})\, da, \tag{3.8}$$

from which it follows that

$$\int_{\mathcal{R}} \dot{\psi}\, dv \le \int_{\partial\mathcal{R}} (\boldsymbol{\xi}\cdot\boldsymbol{\nu}_{\partial\mathcal{R}})\dot{\varphi}\, da + \int_{\partial\mathcal{R}} (\alpha-\psi)\mathbf{q}\cdot\boldsymbol{\nu}_{\partial\mathcal{R}}\, da. \tag{3.9}$$

Since it is always possible to find another control volume, say $\mathcal{R}'$, which coincides with $\mathcal{R}$ at a given instant but with normal velocity $\mathbf{q}'\cdot\boldsymbol{\nu}_{\partial\mathcal{R}}$ different from $\mathbf{q}\cdot\boldsymbol{\nu}_{\partial\mathcal{R}}$, it

therefore follows that $\alpha = \psi$ and that the configurational stress tensor must be of the form

$$\mathbf{C} = \psi\mathbf{1} - \nabla\varphi \otimes \boldsymbol{\xi}. \tag{3.10}$$

Finally, using (3.10) in the local configurational force balance (3.5) yields

$$\operatorname{div}(\psi\mathbf{1} - \nabla\varphi \otimes \boldsymbol{\xi}) + \mathbf{f} = \mathbf{0}. \tag{3.11}$$

In the absence of lower-dimensional defect structures, this equation *determines* the configurational body force density $\mathbf{f} = -\operatorname{div}(\psi\mathbf{1} - \nabla\varphi \otimes \boldsymbol{\xi})$. The balance (3.11) is independent of any particular constitutive assumptions. Only when one invokes (2.8) does it reduce to (3.1), in which case $\mathbf{f} = \beta(\varphi, \nabla\varphi, \dot{\varphi})\dot{\varphi}\nabla\varphi$.

4. Uniformity surfaces. Normal configurational force balance. In the phase-field theory, an interface is a diffuse transition layer, and each value that φ takes within such a layer can be thought of as representing a particular state of the material. For this reason the time-dependent level sets

$$\{\mathbf{x} : \varphi(\mathbf{x}, t) = \text{constant}\} \tag{4.1}$$

are important. We refer to such sets as *uniformity surfaces*.

Within transition layers, $\nabla\varphi$ should be nontrivial; this being the case, we see that

$$\mathbf{n} = \frac{\nabla\varphi}{|\nabla\varphi|} \tag{4.2}$$

and

$$V = -\frac{\dot{\varphi}}{|\nabla\varphi|} \tag{4.3}$$

represent a unit normal field and a corresponding (scalar) normal velocity field for uniformity surfaces, that

$$\mathbf{P} = \mathbf{1} - \mathbf{n} \otimes \mathbf{n} \tag{4.4}$$

projects vector fields onto their components tangent to uniformity surfaces, and that

$$\mathbf{L} = -(\nabla\mathbf{n})\mathbf{P} \qquad \text{and} \qquad K = \operatorname{tr}\mathbf{L} = -\operatorname{div}\mathbf{n} \tag{4.5}$$

are the curvature tensor and (twice) the mean curvature of uniformity surfaces.

From (4.2), $|\nabla\varphi|\nabla\mathbf{n} = \mathbf{P}\nabla\nabla\varphi$, and it follows that

$$\mathbf{L} = -\frac{1}{|\nabla\varphi|}\mathbf{P}(\nabla\nabla\varphi)\mathbf{P} \tag{4.6}$$

and

$$K = -\frac{1}{|\nabla\varphi|}(\Delta\varphi - \mathbf{n} \cdot (\nabla\nabla\varphi)\mathbf{n}). \tag{4.7}$$

Assuming that $\nabla\varphi \neq \mathbf{0}$, we may calculate the component of the configurational force balance (3.11) in the direction $\mathbf{n}$ normal to uniformity surfaces. Bearing in mind (4.2) and (4.6), that calculation yields an identity

$$\operatorname{div}(\psi\mathbf{n} - |\nabla\varphi|\boldsymbol{\xi}) + \psi K + \mathbf{f} \cdot \mathbf{n} = 0 \tag{4.8}$$

that we refer to as the *normal configurational force balance for uniformity surfaces*. In combination with the constitutive equations (2.8), the auxiliary consequence $\mathbf{f} = \beta(\varphi, \nabla\varphi, \dot{\varphi})\dot{\varphi}\nabla\varphi$ of (2.8), and (4.3), the balance (4.8) provides an evolution equation,

$$|\nabla\varphi|^2\beta(\varphi, \nabla\varphi, \dot{\varphi})V = \hat{\psi}(\varphi, \nabla\varphi)K + \operatorname{div}\left\{\hat{\psi}(\varphi, \nabla\varphi)\mathbf{n} - |\nabla\varphi|\frac{\partial\hat{\psi}(\varphi, \nabla\varphi)}{\partial(\nabla\varphi)}\right\}, \tag{4.9}$$

for φ. The evolution equation (4.9) is valid and equivalent to the evolution equation (2.9), provided that $\nabla\varphi \neq \mathbf{0}$. Otherwise, if $\nabla\varphi = \mathbf{0}$, we cannot impose (4.9), which was derived based on the assumption that $\nabla\varphi \neq \mathbf{0}$.

5. The role of standard forces. In the preceding discussion, we have ignored all mention of standard forces. If we allow for a *standard stress tensor* $\mathbf{S}$ and a *standard body force density* $\mathbf{b}$, the balances for forces and moments require that

$$\int_{\partial\mathcal{P}} \mathbf{S}\boldsymbol{\nu}\, da + \int_{\mathcal{P}} \mathbf{b}\, dv = \mathbf{0} \tag{5.1}$$

and

$$\int_{\partial\mathcal{P}} (\mathbf{x} - \mathbf{0}) \times \mathbf{S}\boldsymbol{\nu}\, da + \int_{\mathcal{P}} (\mathbf{x} - \mathbf{0}) \times \mathbf{b}\, dv = \mathbf{0} \tag{5.2}$$

for each part $\mathcal{P}$ of $\mathcal{B}$, or, equivalently, that

$$\operatorname{div}\mathbf{S} + \mathbf{b} = \mathbf{0} \qquad \text{and} \qquad \mathbf{S} = \mathbf{S}^{\top}. \tag{5.3}$$

Since the material is rigid, $\mathbf{S}$ is constitutively indeterminate and the field equation $(5.3)_1$ is satisfied trivially—that is, given an externally supplied body force density $\mathbf{b}$, $\mathbf{S}$ is any symmetric tensor field consistent with $(5.3)_1$.

With standard forces taken into account, it is reasonable to expect some modification of the configurational force system. We therefore write $\tilde{\mathbf{C}}$ and $\tilde{\mathbf{f}}$ for the configurational stress tensor and body force density and assume that these obey the configurational force balance

$$\int_{\partial\mathcal{P}} \tilde{\mathbf{C}}\boldsymbol{\nu}\, da + \int_{\mathcal{P}} \tilde{\mathbf{f}}\, dv = \mathbf{0} \tag{5.4}$$

for each part $\mathcal{P}$ of $\mathcal{B}$, or, equivalently,

$$\operatorname{div}\tilde{\mathbf{C}} + \tilde{\mathbf{f}} = \mathbf{0}. \tag{5.5}$$

Since the body is rigid, the standard force balance (5.1) ensures that the external power expended on any part $\mathcal{P}$ vanishes. However, for a control volume $\mathcal{R}$ whose boundary $\partial\mathcal{R}$ migrates with velocity $\mathbf{q}$, the power expenditure of the traction $\mathbf{S}\boldsymbol{\nu}_{\partial\mathcal{R}}$ distributed over $\partial\mathcal{R}$ is not generally trivial. Thus, when the body is rigid, standard stress is taken into account, the net power expended on a migrating control volume $\mathcal{R}$ is given by

$$\int_{\partial\mathcal{R}} \{(\tilde{\mathbf{C}} + \nabla\varphi \otimes \boldsymbol{\xi} + \mathbf{S})\boldsymbol{\nu}_{\partial\mathcal{R}} \cdot \mathbf{q} + (\boldsymbol{\xi} \cdot \boldsymbol{\nu}_{\partial\mathcal{R}})\dot{\varphi}\}\, da, \tag{5.6}$$

and the requirement that (5.6) be invariant with respect to the choice of the tangential component of the velocity $\mathbf{q}$ yields $\tilde{\mathbf{C}} + \nabla\varphi \otimes \boldsymbol{\xi} + \mathbf{S} = \alpha\mathbf{1}$. Thus, the free-energy

imbalance is unchanged from (3.8) and we find that, when the body is rigid and standard stress is taken into account, the configurational stress tensor admits the representation

$$\tilde{\mathbf{C}} = \psi\mathbf{1} - \nabla\varphi \otimes \boldsymbol{\xi} - \mathbf{S}. \tag{5.7}$$

Comparing (5.7) with (3.10), we see that $\tilde{\mathbf{C}} = \mathbf{C} - \mathbf{S}$. Further, in view of the standard force balance $(5.3)_1$ and the configurational force balance (5.5), it follows that $\tilde{\mathbf{f}} = \mathbf{f} - \mathbf{b}$. On the other hand, granted that $\tilde{\mathbf{C}} = \mathbf{C} - \mathbf{S}$ and $\tilde{\mathbf{f}} = \mathbf{f} - \mathbf{b}$, $(5.3)_1$ implies that

$$\operatorname{div}\tilde{\mathbf{C}} + \tilde{\mathbf{f}} = \operatorname{div}\mathbf{C} + \mathbf{f} = \mathbf{0}, \tag{5.8}$$

and we may conclude that, for a rigid body, standard forces have no impact on the configurational force balance and may be neglected without loss of generality.

6. Specialization. For simplicity, we suppose that the constitutive relation determining the free-energy density has the simple form

$$\psi = f(\varphi) + \Psi_\alpha(1 - g(\varphi)) + \Psi_\beta g(\varphi) + \frac{1}{2}\lambda|\nabla\varphi|^2, \tag{6.1}$$

where f is a double-well potential with equal minima at $\varphi = \varphi_\alpha$ and $\varphi = \varphi_\beta$, with $\varphi_\alpha < \varphi_\beta$, viz.,

$$0 = f(\varphi_\alpha) = f(\varphi_\beta) < f(\varphi) \qquad \text{for all } \varphi \neq \varphi_\alpha, \varphi_\beta; \tag{6.2}$$

g vanishes for $\varphi \leq \varphi_\alpha$, is equal to unity for $\varphi \geq \varphi_\beta$, and increases monotonically between $\varphi = \varphi_\alpha$ and $\varphi = \varphi_\beta$, viz.,

$$g(\varphi) = \begin{cases} 0 & 0 \leq \varphi_\alpha, \\ 1 & \varphi \geq \varphi_\beta \end{cases} \tag{6.3}$$

and

$$g'(\varphi) > 0 \qquad \text{for all } \varphi \in (\varphi_\alpha, \varphi_\beta); \tag{6.4}$$

λ is constant and strictly positive, viz.,

$$\lambda > 0; \tag{6.5}$$

and Ψ_α and Ψ_β are the constant energy densities of the bulk phases α and β. For a body $\mathcal{B}$, where the average of φ lies between φ_α and φ_β, the double-well structure of f lends energetic preference to distributions of φ consisting of regions with $\varphi = \varphi_\alpha$ and regions with $\varphi = \varphi_\beta$. On the other hand, the gradient term $\frac{1}{2}\lambda|\nabla\varphi|^2$ penalizes sharp transitions between such regions and in so doing facilitates the existence of equilibria in which φ is smooth and $\mathcal{B}$ contains interfacial layers separating regions with φ close to φ_α from regions with φ close to φ_β. Because the gradient term depends only on the magnitude $|\nabla\varphi|$ of $\nabla\varphi$, interfacial layers of all orientations are of equal energetic cost.

As a further simplification, we assume that the kinetic modulus β is constant and strictly positive, viz.,

$$\beta(\varphi, \nabla\varphi, \dot{\varphi}) = B > 0. \tag{6.6}$$

With the choice of (6.6), (2.10) specializes to $\delta = -B\dot{\varphi}^2$; thus, the rate at which energy is dissipated by the growth of either phase at the expense of another is quadratic in $\dot{\varphi}$ and is insensitive to layer orientation.

In view of the specializations (6.1) and (6.6), the evolution equation (2.9) becomes

$$B\dot{\varphi} = \lambda\Delta\varphi - f'(\varphi) - [\![\Psi]\!]g'(\varphi), \tag{6.7}$$

with $[\![\Psi]\!] = \Psi_\beta - \Psi_\alpha$, and the normal configurational force balance (4.9) for uniformity surfaces becomes

$$\begin{aligned} |\nabla\varphi|^2 BV = &\left\{ f(\varphi) + \Psi_\alpha(1-g(\varphi)) + \Psi_\beta g(\varphi) + \frac{1}{2}\lambda|\nabla\varphi|^2 \right\} K \\ &+ \operatorname{div}\left\{ (f(\varphi) + \Psi_\alpha(1-g(\varphi)) + \Psi_\beta g(\varphi) - \frac{1}{2}\lambda|\nabla\varphi|^2)\mathbf{n} \right\}. \end{aligned} \tag{6.8}$$

Since

$$\big\{\Psi_\alpha(1-g(\varphi)) + \Psi_\beta g(\varphi)\big\}K + \operatorname{div}\big\{\Psi_\alpha(1-g(\varphi)) + \Psi_\beta g(\varphi)\big\} = [\![\Psi]\!]\mathbf{n}\cdot\nabla g(\varphi), \tag{6.9}$$

(6.8) can be rewritten somewhat more concisely as

$$|\nabla\varphi|^2 BV = \left\{ f(\varphi) + \frac{1}{2}\lambda|\nabla\varphi|^2 \right\} K + \operatorname{div}\left\{ (f(\varphi) - \frac{1}{2}\lambda|\nabla\varphi|^2)\mathbf{n} \right\} + [\![\Psi]\!]\mathbf{n}\cdot\nabla g(\varphi), \tag{6.10}$$

and we will use this in lieu of (6.8). We emphasize that (6.8) (and, thus, (6.10)) is meaningful only if $\nabla\varphi \neq \mathbf{0}$ and that, if (6.10) is meaningful, it is equivalent to (6.7).

7. Scaling. We introduce characteristic measures

$$\mu = \frac{1}{2}(\Psi_\alpha + \Psi_\beta) \qquad \text{and} \qquad \nu = \max_{\varphi\in(\varphi_\alpha,\varphi_\beta)} f(\varphi) \tag{7.1}$$

for the free energies, per unit volume, of the bulk phases and interfacial transition layers, and assume that these yield a small dimensionless parameter

$$0 < \epsilon = \frac{\mu}{\nu} \ll 1. \tag{7.2}$$

Then, letting L denote a characteristic length and T a characteristic time, and labeling the dimensional (unscaled) fields with asterisks, we introduce the dimensionless independent and dependent variables

$$\mathbf{x} = \frac{\mathbf{x}^*}{L}, \qquad t = \frac{t^*}{T}, \qquad \varphi_\epsilon(\mathbf{x},t) = \varphi^*(\mathbf{x}^*,t^*), \tag{7.3}$$

and constitutive quantities

$$f(\varphi_\epsilon) = \frac{f^*(\varphi^*)}{\nu}, \quad \Psi_\alpha = \frac{\Psi_\alpha^*}{\mu}, \quad \Psi_\beta = \frac{\Psi_\beta^*}{\mu}, \quad \epsilon\lambda = \frac{\lambda^*}{\mu L^2}, \quad \epsilon B = \frac{B^*}{\mu T}, \tag{7.4}$$

where the dependence of the fields on the parameter ϵ has been made explicit and the quantities without asterisks in (7.4) are assumed to be of $O(1)$ in ϵ.

With this scaling, the dimensionless free-energy density is given by

$$\psi_\epsilon = \frac{\psi}{\mu} = \epsilon^{-1} f(\varphi_\epsilon) + \Psi_\alpha(1 - g(\varphi_\epsilon)) + \Psi_\beta g(\varphi_\epsilon) + \frac{1}{2}\epsilon\lambda|\nabla\varphi_\epsilon|^2, \tag{7.5}$$

and the governing evolution equation for φ_ϵ becomes

$$\epsilon B\dot{\varphi}_\epsilon = \epsilon\lambda\Delta\varphi_\epsilon - \epsilon^{-1} f'(\varphi_\epsilon) - [\![\Psi]\!] g'(\varphi_\epsilon). \tag{7.6}$$

Further, the normal configurational force balance (6.10) reads

$$\begin{aligned}\epsilon|\nabla\varphi_\epsilon|^2 BV_\epsilon = {} & \left\{\epsilon^{-1} f(\varphi_\epsilon) + \frac{1}{2}\epsilon\lambda|\nabla\varphi_\epsilon|^2\right\}K_\epsilon \\ & + \operatorname{div}\left\{\left\{\epsilon^{-1} f(\varphi_\epsilon) - \frac{1}{2}\epsilon\lambda|\nabla\varphi_\epsilon|^2\right\}\mathbf{n}_\epsilon\right\} + [\![\Psi]\!]\mathbf{n}_\epsilon \cdot \nabla g(\varphi_\epsilon),\end{aligned} \tag{7.7}$$

with (cf. (4.2), (4.3), and (4.7))

$$\mathbf{n}_\epsilon = \frac{\nabla\varphi_\epsilon}{|\nabla\varphi_\epsilon|}, \quad V_\epsilon = -\frac{\dot{\varphi}_\epsilon}{|\nabla\varphi_\epsilon|}, \quad \text{and} \quad K_\epsilon = -\frac{1}{|\nabla\varphi_\epsilon|}(\Delta\varphi_\epsilon - \mathbf{n}_\epsilon \cdot (\nabla\nabla\varphi_\epsilon)\mathbf{n}_\epsilon). \tag{7.8}$$

8. Expansions. Hereafter, we focus on a fixed part $\mathcal{P}$ of $\mathcal{B}$ that, over some time interval, consists of three evolving subregions $\mathcal{P}_\epsilon^\alpha$, $\mathcal{S}_\epsilon$, and $\mathcal{P}_\epsilon^\beta$. At each time t, $\mathcal{S}_\epsilon(t)$ is a transition layer consisting of points $\mathbf{x}$ in $\mathcal{B}$ with $\varphi_\alpha < \varphi_\epsilon(\mathbf{x}, t) < \varphi_\beta$, while $\mathcal{P}_\epsilon^\alpha(t)$ and $\mathcal{P}_\epsilon^\beta(t)$ consist of points $\mathbf{x}$ with $\varphi_\epsilon(\mathbf{x}, t) \approx \varphi_\alpha$ and $\varphi_\epsilon(\mathbf{x}, t) \approx \varphi_\beta$, respectively. We assume that the limit

$$\mathcal{S} = \lim_{\epsilon \to 0+} \mathcal{S}_\epsilon \tag{8.1}$$

exists, with $\mathcal{S}(t)$ a smoothly evolving surface and with

$$\mathcal{P} = \mathcal{P}^\alpha(t) \cup \mathcal{S}(t) \cup \mathcal{P}^\beta(t), \tag{8.2}$$

with $\mathcal{P}^\gamma(t) = \lim_{\epsilon \to 0} \mathcal{P}_\epsilon^\gamma(t)$ for $\gamma = \alpha, \beta$.

We write $\ell(\mathbf{x}, t)$ for the *signed distance* between a point $\mathbf{x}$ in $\mathcal{P}$ and the surface $\mathcal{S}(t)$, with $\ell(\mathbf{x}, t) < 0$ in $\mathcal{P}_\epsilon^\alpha(t)$ and $\ell(\mathbf{x}, t) > 0$ in $\mathcal{P}_\epsilon^\beta(t)$. Then

$$\mathbf{n}(\mathbf{x}, t) = \nabla\ell(\mathbf{x}, t) \qquad \text{and} \qquad V_{\mathcal{S}}(\mathbf{x}, t) = -\dot{\ell}(\mathbf{x}, t) \tag{8.3}$$

represent a unit normal field and corresponding scalar normal velocity field for $\mathcal{S}(t)$. We also assume that $\ell(\mathbf{x}, t)$ is smooth within $\mathcal{S}_\epsilon(t)$ and that given any $\mathbf{x}$ on $\mathcal{S}_\epsilon(t)$, there is a unique $\mathbf{z}$ on $\mathcal{S}(t)$ with $\mathbf{z} = \mathbf{x} - \ell(\mathbf{x}, t)\mathbf{n}(\mathbf{x}, t)$. The mapping $\mathbf{x} \mapsto \big(\ell(\mathbf{x}, t), \mathbf{z}(\mathbf{x}, t)\big)$ is then one-to-one on $\mathcal{S}_\epsilon(t)$; further, $\mathbf{n}(\mathbf{x}, t)$ and $V_{\mathcal{S}}(\mathbf{x}, t)$ are well defined and independent of $\ell(\mathbf{x}, t)$ at each $\mathbf{x}$ in $\mathcal{S}_\epsilon(t)$: $\mathbf{n}(\mathbf{x}, t) = \mathbf{n}(\mathbf{z}, t)$, $V_{\mathcal{S}}(\mathbf{x}, t) = V_{\mathcal{S}}(\mathbf{z}, t)$. Thus, writing $\nabla_{\mathcal{S}}$ and $\operatorname{div}_{\mathcal{S}}$ for the surface gradient and surface divergence on $\mathcal{S}$, the curvature tensor $\mathbf{L}$ and the total curvature $K_{\mathcal{S}}$ for $\mathcal{S}$,

$$\mathbf{L} = -\nabla_{\mathcal{S}}\mathbf{n} \qquad \text{and} \qquad K_{\mathcal{S}} = \operatorname{tr}\mathbf{L} = -\operatorname{div}_{\mathcal{S}}\mathbf{n} \tag{8.4}$$

are also independent of ℓ: $\mathbf{L}(\mathbf{x}, t) = \mathbf{L}(\mathbf{z}, t)$, $K_{\mathcal{S}}(\mathbf{x}, t) = K_{\mathcal{S}}(\mathbf{z}, t)$.

Within $\mathcal{S}_\epsilon(t)$, we stretch the coordinate normal to $\mathcal{S}(t)$ by letting

$$r(\mathbf{x}, t) = \epsilon^{-1}\ell(\mathbf{x}, t) \tag{8.5}$$

and, in accord with this, we assume that the thickness $h_\epsilon(t)$ of $\mathcal{S}_\epsilon(t)$ tends to zero with ϵ, but at a slightly slower rate, viz.,

$$\lim_{\epsilon\to 0} h_\epsilon = 0, \quad \lim_{\epsilon\to 0}(\epsilon^{-1}h_\epsilon) = +\infty, \quad \lim_{\epsilon\to 0}(\epsilon^{-1}h_\epsilon^2) = 0. \tag{8.6}$$

For the phase field φ_ϵ, we introduce an *outer expansion*

$$\varphi_\epsilon(\mathbf{x},t) = \varphi_0^{\text{out}}(\mathbf{x},t) + \epsilon\varphi_1^{\text{out}}(\mathbf{x},t) + O(\epsilon^2), \tag{8.7}$$

assumed valid within the regions $\mathcal{P}_\epsilon^\alpha$ and $\mathcal{P}_\epsilon^\beta$, and an *inner expansion*

$$\varphi_\epsilon(\mathbf{x},t) = \varphi_0^{\text{in}}\big(r(\mathbf{x},t),\mathbf{z}(\mathbf{x},t),t\big) + \epsilon\varphi_1^{\text{in}}\big(r(\mathbf{x},t),\mathbf{z}(\mathbf{x},t),t\big) + O(\epsilon^2), \tag{8.8}$$

assumed valid within the layer; here, $\varphi_0^{\text{out}}(\mathbf{x},t)$, $\varphi_1^{\text{out}}(\mathbf{x},t)$ and $\varphi_0^{\text{in}}(r,\mathbf{z},t)$, $\varphi_1^{\text{in}}(r,\mathbf{z},t)$ are smooth, bounded functions of their arguments. We further assume that these expansions are twice formally differentiable in their arguments in the sense that $\nabla\varphi_\epsilon = \nabla\varphi_0^{\text{out}} + \epsilon\nabla\varphi_1^{\text{out}} + O(\epsilon^2)$ for the outer expansion and, on letting $\acute{\varphi}_\epsilon$ denote the partial derivative of φ_ϵ with respect to r, $\acute{\varphi}_\epsilon = \acute{\varphi}_0^{\text{in}} + \epsilon\acute{\varphi}_1^{\text{in}} + O(\epsilon^2)$ for the inner expansion, and so forth.

Hence, we do not presume that $\mathcal{S}_\epsilon(t)$ is disjoint from $\mathcal{P}_\epsilon^\alpha(t)$ and $\mathcal{P}_\epsilon^\beta(t)$: the regions $\mathcal{S}_\epsilon(t)\cap\big(\mathcal{P}_\epsilon^\alpha(t)\cup\mathcal{P}_\epsilon^\beta(t)\big)$ of overlap represent sets where the outer and inner expansions agree. In particular, we have the matching condition

$$(\varphi_0^{\text{out}})^\pm(\mathbf{x},t) = \lim_{\ell(\mathbf{x},t)\to 0\pm}\varphi_0^{\text{out}}(\mathbf{x},t) = \lim_{r\to\pm\infty}\varphi_0^{\text{in}}(r,\mathbf{z},t) = (\varphi_0^{\text{in}})^\pm(r,\mathbf{z},t) \tag{8.9}$$

relating the $O(1)$ terms of the inner and outer expansions for φ_ϵ within the overlap region.

In terms of the variables $(r,\mathbf{z})$, the derivative with respect to $\mathbf{z}$ holding r fixed may be identified with the gradient $\nabla_{\mathcal{S}}$ on $\mathcal{S}$. Let

$$\mathbf{P} = \mathbf{1} - \mathbf{n}\otimes\mathbf{n}. \tag{8.10}$$

Then, since $\mathbf{z}(\mathbf{x},t) = \mathbf{x} - \ell(\mathbf{x},t)\mathbf{n}(\mathbf{x},t)$, it follows that

$$\nabla\mathbf{z} = \mathbf{P} + \ell\mathbf{M}_\epsilon, \tag{8.11}$$

with

$$\mathbf{M}_\epsilon = -\nabla\mathbf{n}. \tag{8.12}$$

To determine the dependence of $\mathbf{M}_\epsilon$ on ϵ, note that, since $|\ell| \le h_\epsilon = o(1)$ and $\acute{\ell} = \epsilon$, differentiating both sides of the relation $\mathbf{n}(\mathbf{x},t) = \mathbf{n}(\mathbf{z}(\mathbf{x},t),t)$ with respect to $\mathbf{x}$ yields

$$\mathbf{M}_\epsilon = (\mathbf{1}-\ell\mathbf{L})^{-1}\mathbf{L} = \mathbf{L} + o(1). \tag{8.13}$$

Thus, for Φ and $\mathbf{v}$ scalar- and vector-valued fields, we find that

$$\begin{cases} \nabla\Phi = \epsilon^{-1}\acute{\Phi}\mathbf{n} + (\mathbf{P}+\ell\mathbf{M}_\epsilon)\nabla_{\mathcal{S}}\Phi = \epsilon^{-1}\acute{\Phi}\mathbf{n} + \big(1+o(1)\big)\nabla_{\mathcal{S}}\Phi, \\ \nabla\mathbf{v} = \epsilon^{-1}\acute{\mathbf{v}}\otimes\mathbf{n} + (\nabla_{\mathcal{S}}\mathbf{v})(\mathbf{P}+\ell\mathbf{M}_\epsilon) = \epsilon^{-1}\acute{\mathbf{v}}\otimes\mathbf{n} + \big(1+o(1)\big)\nabla_{\mathcal{S}}\mathbf{v}, \end{cases} \tag{8.14}$$

so that

$$\begin{aligned} \nabla\nabla\Phi = {}& \epsilon^{-2}\acute{\acute{\Phi}}\mathbf{n}\otimes\mathbf{n} + \epsilon^{-1}\big(1+o(1)\big)\big(\nabla_{\mathcal{S}}\acute{\Phi}\otimes\mathbf{n} + \mathbf{n}\otimes\nabla_{\mathcal{S}}\acute{\Phi} - \acute{\Phi}\mathbf{L}\big) \\ & + (\nabla_{\mathcal{S}}\nabla_{\mathcal{S}}\Phi)O(1) + O(1)\nabla_{\mathcal{S}}\Phi, \end{aligned} \tag{8.15}$$

with the $O(1)$ and $o(1)$ estimates in (8.14) and (8.15) of appropriate tensorial order and independent of Φ and $\mathbf{v}$.

As a further consequence of the relation $\mathbf{z}(\mathbf{x},t) = \mathbf{x} - \ell(\mathbf{x},t)\mathbf{n}(\mathbf{x},t)$, it follows that

$$\dot{\mathbf{z}} = V_{\mathcal{S}}\mathbf{n} + \ell\mathbf{v}_\epsilon, \tag{8.16}$$

with

$$\mathbf{v}_\epsilon = -\dot{\mathbf{n}}. \tag{8.17}$$

To determine the dependence of $\mathbf{v}_\epsilon$ on ϵ, note that $\dot{\mathbf{n}} = \nabla\dot{\ell} = -\nabla V_{\mathcal{S}}$. Thus, since $\nabla V_{\mathcal{S}} = (\mathbf{P} + \ell\mathbf{M}_\epsilon)\nabla_{\mathcal{S}} V_{\mathcal{S}}$ and $\mathring{\mathbf{n}} = -\nabla_{\mathcal{S}} V_{\mathcal{S}}$, with $\mathring{\mathbf{n}}$ the time-rate of $\mathbf{n}$ following the normal trajectories of $\mathcal{S}$,

$$\mathbf{v}_\epsilon = (\mathbf{P} + \ell\mathbf{M}_\epsilon)\nabla_{\mathcal{S}} V_{\mathcal{S}} = \nabla_{\mathcal{S}} V_{\mathcal{S}} + o(1). \tag{8.18}$$

Thus, for Φ a scalar field,

$$\dot{\Phi} = -\epsilon^{-1} V_{\mathcal{S}}\acute{\Phi} + \ell\nabla_{\mathcal{S}}\Phi \cdot (\mathbf{P} + \ell\mathbf{L})\mathring{\mathbf{n}} + \Phi_t = -\epsilon^{-1} V_{\mathcal{S}}\acute{\Phi} + \Phi_t + o(1), \tag{8.19}$$

where Φ_t denotes the partial time-rate of Φ holding r and $\mathbf{z}$ fixed.

9. Bulk regions. Using the outer expansion (8.7) of φ_ϵ in the scaled evolution equation (7.6) and neglecting terms of $O(1)$ and smaller in ϵ, we find that $f'(\varphi_0^{\text{out}}) = 0$ so that, since f is a double-well potential with equal minima at φ_α and φ_β,

$$\varphi_0^{\text{out}} = \begin{cases} \varphi_\alpha & \text{on } \mathcal{P}_\epsilon^\alpha, \\ \varphi_\beta & \text{on } \mathcal{P}_\epsilon^\beta. \end{cases} \tag{9.1}$$

Further

$$f(\varphi_\epsilon) = o(\epsilon) \quad \text{and} \quad f'(\varphi_\epsilon) = o(1) \quad \text{on} \quad \mathcal{P}_\epsilon^\alpha \cup \mathcal{P}_\epsilon^\beta, \tag{9.2}$$

and

$$\dot{\varphi}_\epsilon, \nabla\varphi_\epsilon, \nabla\dot{\varphi}_\epsilon = O(\epsilon) \quad \text{on} \quad \mathcal{P}_\epsilon^\alpha \cup \mathcal{P}_\epsilon^\beta. \tag{9.3}$$

Thus, it follows that

$$\psi_\epsilon = \begin{cases} \Psi_\alpha + O(\epsilon) & \text{on } \mathcal{P}_\epsilon^\alpha, \\ \Psi_\beta + O(\epsilon) & \text{on } \mathcal{P}_\epsilon^\beta. \end{cases} \tag{9.4}$$

10. Transition layer.

10.1. Basic estimates. Applying $(8.14)_1$, $(8.14)_2$, and (8.19) to the inner expansion (8.8) of φ_ϵ, we find that

$$\begin{cases} \nabla\varphi_\epsilon = \epsilon^{-1}\acute{\varphi}_0^{\text{in}}\mathbf{n} + \nabla_{\mathcal{S}}\varphi_0^{\text{in}} + \acute{\varphi}_1^{\text{in}}\mathbf{n} + O(\epsilon), \\ |\nabla\varphi_\epsilon| = \epsilon^{-1}\acute{\varphi}_0^{\text{in}} + \acute{\varphi}_1^{\text{in}} + O(\epsilon), \\ \nabla\nabla\varphi_\epsilon = \epsilon^{-2}\ddot{\varphi}_0^{\text{in}}\mathbf{n}\otimes\mathbf{n} \\ \qquad\qquad + \epsilon^{-1}(\nabla_{\mathcal{S}}\acute{\varphi}_1^{\text{in}}\otimes\mathbf{n} + \mathbf{n}\otimes\nabla_{\mathcal{S}}\acute{\varphi}_1^{\text{in}} - \acute{\varphi}_0^{\text{in}}\mathbf{L} + \ddot{\varphi}_1^{\text{in}}\mathbf{n}\otimes\mathbf{n}) + O(1), \\ \Delta\varphi_\epsilon = \epsilon^{-2}\ddot{\varphi}_0^{\text{in}} - \epsilon^{-1}(K_{\mathcal{S}}\acute{\varphi}_0^{\text{in}} - \ddot{\varphi}_1^{\text{in}}) + O(1), \\ \dot{\varphi}_\epsilon = -\epsilon^{-1} V_{\mathcal{S}}\acute{\varphi}_0^{\text{in}} + O(1), \end{cases} \tag{10.1}$$

and, applying these estimates to (7.8), that

$$\mathbf{n}_\epsilon = \mathbf{n} + O(\epsilon), \qquad V_\epsilon = V_{\mathcal{S}} + O(\epsilon), \qquad \text{and} \qquad K_\epsilon = K_{\mathcal{S}} + o(1). \tag{10.2}$$

10.2. Equipartition of free-energy density and its consequences. Using the inner expansion of φ_ϵ and the estimates (10.1) in the scaled evolution equation (7.6) and neglecting terms of $O(1)$ and smaller, we find that φ_0^{in} must satisfy the ordinary differential equation

$$\lambda \ddot{\varphi}_0^{\text{in}} = f'(\varphi_0^{\text{in}}). \tag{10.3}$$

Further, in view of the matching condition (8.9) and the result (9.1) concerning φ_0^{out}, φ_0^{in} must satisfy

$$\varphi_0^{\text{in}} \to \begin{cases} \varphi_\alpha & \text{as } r \to -\infty, \\ \varphi_\beta & \text{as } r \to +\infty, \end{cases} \tag{10.4}$$

along with

$$\dot{\varphi}_0^{\text{in}} \to 0 \qquad \text{and} \qquad \ddot{\varphi}_0^{\text{in}} \to 0 \qquad \text{as } r \to \pm\infty. \tag{10.5}$$

Since f is a double-well potential with equal minima at φ_α and φ_β, the boundary-value problem formed by (10.3) and (10.4) possesses a unique solution φ_0^{in} that increases monotonically from the value φ_α at $r = -\infty$ to the value φ_β at $r = +\infty$. Further, φ_0^{in} must be independent of $\mathbf{z}$.

Granted that the boundary conditions $(10.4)_{2,3}$ hold, the differential equation $(10.3)_1$ possesses a first integral

$$\frac{1}{2}\lambda|\dot{\varphi}_0^{\text{in}}|^2 = f(\varphi_0^{\text{in}}), \tag{10.6}$$

which we interpret as an expression of the equipartition of the free-energy density within the layer (upto the most significant order in ϵ), between the double-well potential f and the gradient energy density $\frac{1}{2}\lambda|\nabla\varphi_\epsilon|^2$. Since f and f' vanish at $\varphi = \varphi_\alpha$ and $\varphi = \varphi_\beta$, φ_0^{in} must decay according to $\dot{\varphi}_0^{\text{in}}(r,\cdot) = O(e^{-c|r|})$ as $|r| \to \infty$, with $c > 0$ independent of r. Hence, $\dot{\varphi}_0^{\text{in}}$ is, as a function of r, square-integrable on $(-\infty, +\infty)$. Thus, by (10.6), (10.3), and (10.4),

$$\int_{-\infty}^{+\infty} \sqrt{\lambda}|\dot{\varphi}_0^{\text{in}}(r,\cdot)|^2\, dr = \int_{\varphi_\alpha}^{\varphi_\beta} \sqrt{2f(\varphi)}\, d\varphi. \tag{10.7}$$

For convenience, we introduce

$$\psi_{\mathcal{S}} = \sqrt{\lambda} \int_{\varphi_\alpha}^{\varphi_\beta} \sqrt{2f(\varphi)}\, d\varphi \tag{10.8}$$

and note that, if rewritten in terms of dimensional quantities, $\psi_{\mathcal{S}}$ would carry dimensions of free-energy per unit area. Granted that (10.8) holds, it follows from (10.7) that

$$\int_{-\infty}^{+\infty} \lambda|\dot{\varphi}_0^{\text{in}}(r)|^2\, dr = \psi_{\mathcal{S}}. \tag{10.9}$$

10.3. Interfacial evolution equation. Indirect approach. At $O(1)$, the scaled evolution equation (7.6) yields the linear but inhomogeneous equation

$$\lambda \ddot{\varphi}_1^{\text{in}} - f''(\varphi_0^{\text{in}})\varphi_1^{\text{in}} = -\rho, \tag{10.10}$$

with

$$\rho = BV_{\mathcal{S}}\varphi_0^{\prime\,\mathrm{in}} - \lambda K_{\mathcal{S}}\varphi_0^{\prime\prime\,\mathrm{in}} - [\![\Psi]\!]g'(\varphi_0^{\mathrm{in}}). \tag{10.11}$$

On differentiating (10.3) with respect to r, it follows that $\varphi_0^{\prime\,\mathrm{in}}$ must satisfy the homogeneous equation $\lambda\varphi_0^{\prime\prime\prime\,\mathrm{in}} - f''(\varphi_0^{\mathrm{in}})\varphi_0^{\prime\,\mathrm{in}} = 0$. Thus, by the Fredholm alternative, ρ and $\varphi_0^{\prime\,\mathrm{in}}$ must be orthogonal:

$$\int_{-\infty}^{+\infty} \rho\varphi_0^{\prime\,\mathrm{in}}\, dr = 0. \tag{10.12}$$

Evaluating the integral on the left side of (10.12), using (10.9) and the boundary conditions (10.4), and recalling from (6.3) that g vanishes at $\varphi = \varphi_\alpha$ and is equal to unity at $\varphi = \varphi_\beta$, we find that

$$b_{\mathcal{S}}V_{\mathcal{S}} = \psi_{\mathcal{S}}K_{\mathcal{S}} + [\![\Psi]\!], \tag{10.13}$$

where we have introduced

$$b_{\mathcal{S}} = \frac{B}{\sqrt{\lambda}}\int_{\varphi_\alpha}^{\varphi_\beta} \sqrt{2f(\varphi)}\, d\varphi = \frac{B\psi_{\mathcal{S}}}{\lambda}. \tag{10.14}$$

We note that, if rewritten in terms of dimensional quantities, $b_{\mathcal{S}}$ would carry dimensions of mass per unit time per unit area and would, therefore, represent interfacial reciprocal mobility.

On performing a suitable redimensionalization, we find that (10.13) is precisely the interfacial evolution equation (1.2) governing the evolution of a sharp phase interface endowed with a constant interfacial free energy per unit area $\psi_{\mathcal{S}}$ and reciprocal mobility $b_{\mathcal{S}}$ that separates bulk phases α and β with constant free-energy densities Ψ_α and Ψ_β.

10.4. Interfacial evolution equation. Direct approach. Within the layer, $\nabla\varphi_\epsilon$ is generally nontrivial. Thus, it is permissible to work with the scaled normal configurational force balance (7.7) for uniformity surfaces instead of the scaled evolution equation (7.6). Using the inner expansion of φ_ϵ and the estimates (10.1) and (10.2) in (7.7), and neglecting terms of $O(\epsilon^{-1})$ and smaller, we arrive once again at (10.3) and, bearing in mind (8.9), (9.1), and the properties of f, all the conclusions of section 9 follow. Next, at $O(\epsilon^{-1})$, (7.7) yields, in view of the result (10.6) concerning the partition of free-energy density,

$$BV_{\mathcal{S}}|\varphi_0^{\prime\,\mathrm{in}}|^2 = \lambda K_{\mathcal{S}}|\varphi_0^{\prime\,\mathrm{in}}|^2 + \overline{\left\{f'(\varphi_0^{\mathrm{in}})\varphi_1^{\mathrm{in}} - \lambda\varphi_0^{\prime\,\mathrm{in}}\varphi_1^{\prime\,\mathrm{in}}\right\}} + [\![\Psi]\!]\,\overline{g(\varphi_0^{\mathrm{in}})}^{\,\prime}. \tag{10.15}$$

By integrating (10.15) over r from $r = -\infty$ to $r = +\infty$ and utilizing the definitions (10.9) and (10.14), the boundary and far-field conditions (10.4) and (10.5), and properties of f and g, once again we obtain the evolution equation (10.13) obtained in the previous section by the Fredholm alternative. In this sense, (10.13) can be viewed as a consequence of the normal component of the configurational force balance for uniformity surfaces, obtained in passing to the limit $\epsilon \to 0$ a limit that corresponds to collapsing the transition layer into a surface.

We remarked earlier that, when φ_ϵ is regular, the configurational force balance contains no information beyond that already contained in the evolution equation for φ_ϵ. However, passing to the limit $\epsilon \to 0$ generates surfaces across which φ_ϵ is discontinuous and $\nabla\varphi_\epsilon$ and $\dot{\varphi}_\epsilon$ (as well as other associated derivatives) are undefined. The asymptotic analysis performed here shows that, at such a defect, the normal configurational force balance for uniformity surfaces yields directly information that arises only indirectly—as a solvability condition imposed by the Fredholm alternative—from the evolution equation for φ_ϵ. In this sense, we view the asymptotically derived interfacial evolution equation as a statement of normal configurational force balance for the interface.

Our asymptotic derivations of the evolution equation (1.2) are predicated on (10.8) and (10.14). We interpret (10.8) and (10.14) as *constitutive connections* between the theories at hand, connections that guarantee that the phase-field theory corresponds asymptotically to the sharp-interface theory.

11. Generalization. To obtain the general evolution equation (1.1) from the phase-field theory, we first modify the constitutive equations (6.1) and (6.6), determining the free-energy density and the kinetic modulus to be

$$\psi = f(\varphi) + \Psi_\alpha(1 - g(\varphi)) + \Psi_\beta g(\varphi) + \frac{1}{2}\lambda(\mathbf{n})|\nabla\varphi|^2 \tag{11.1}$$

and

$$\beta(\varphi, \nabla\varphi, \dot{\varphi}) = B(\mathbf{n}, V) > 0, \tag{11.2}$$

with $\mathbf{n}$ and V as defined in (4.2) and (4.3).

Scaling as in section 7, we arrive at the evolution equation

$$\epsilon B(\mathbf{n}_\epsilon, V_\epsilon)\dot{\varphi}_\epsilon = \epsilon \mathrm{div}\left\{|\nabla\varphi_\epsilon|\left(\lambda(\mathbf{n}_\epsilon)\mathbf{n}_\epsilon + \frac{1}{2}\frac{\partial\lambda(\mathbf{n}_\epsilon)}{\partial\mathbf{n}_\epsilon}\right)\right\} - \epsilon^{-1}f'(\varphi_\epsilon) - [\![\Psi]\!]g'(\varphi_\epsilon) \tag{11.3}$$

and the normal configurational force balance for uniformity surfaces

$$\begin{aligned}\epsilon|\nabla\varphi_\epsilon|^2 B(\mathbf{n}_\epsilon, V_\epsilon)V_\epsilon &= \left\{\epsilon^{-1}f(\varphi_\epsilon) + \frac{1}{2}\epsilon\lambda(\mathbf{n}_\epsilon)|\nabla\varphi_\epsilon|^2\right\}K_\epsilon \\ &\quad + \mathrm{div}\left\{\left\{\epsilon^{-1}f(\varphi_\epsilon) - \frac{1}{2}\epsilon\lambda(\mathbf{n}_\epsilon)|\nabla\varphi_\epsilon|^2\right\}\mathbf{n}_\epsilon - \frac{1}{2}\epsilon|\nabla\varphi_\epsilon|^2\frac{\partial\lambda(\mathbf{n}_\epsilon)}{\partial\mathbf{n}_\epsilon}\right\} + [\![\Psi]\!]\mathbf{n}\cdot\nabla g(\varphi_\epsilon),\end{aligned} \tag{11.4}$$

which generalize (7.6) and (7.7).

The results for the bulk regions are unchanged from those presented in section 9. To study the layer, we follow the approach taken in section 10.4. Specifically, at $O(\epsilon^{-2})$, (11.4) yields

$$\overline{\left\{\tfrac{1}{2}\lambda(\mathbf{n})|\acute{\varphi}_0^{\mathrm{in}}|^2 - f(\varphi_0^{\mathrm{in}})\right\}} = 0. \tag{11.5}$$

Further, the matching conditions (8.9) and the bulk results (9.1) yield, as before, the far-field conditions (10.4) and (10.5). Combining (11.5), (10.4), and (10.5), we arrive at the first integral

$$\frac{1}{2}\lambda(\mathbf{n})|\acute{\varphi}_0^{\mathrm{in}}|^2 = f(\varphi_0^{\mathrm{in}}) \tag{11.6}$$

and find, in view of the properties of f, that once again, as a function of r, $\varphi_0^{\rm in}$ must be square-integrable on $(-\infty,+\infty)$. This leads to a generalization

$$\int_{-\infty}^{+\infty}\sqrt{\lambda(\mathbf{n})}|\varphi_0^{\rm in}(r,\cdot,\cdot)|^2\,dr=\int_{\varphi_\alpha}^{\varphi_\beta}\sqrt{2f(\varphi)}\,d\varphi \tag{11.7}$$

of (10.7).

We next introduce analogues

$$\hat{\psi}_{\mathcal{S}}(\mathbf{n})=\sqrt{\lambda(\mathbf{n})}\int_{\varphi_\alpha}^{\varphi_\beta}\sqrt{2f(\varphi)}\,d\varphi \tag{11.8}$$

and

$$b_{\mathcal{S}}(\mathbf{n},V_{\mathcal{S}})=\frac{B(\mathbf{n},V_{\mathcal{S}})}{\sqrt{\lambda(\mathbf{n})}}\int_{\varphi_\alpha}^{\varphi_\beta}\sqrt{2f(\varphi)}\,d\varphi=\frac{B(\mathbf{n},V_{\mathcal{S}})\hat{\psi}_{\mathcal{S}}(\mathbf{n})}{\lambda(\mathbf{n})} \tag{11.9}$$

of the constitutive connections (10.8) and (10.14). Direct consequences of (11.7) and (11.8) are the identities

$$\begin{cases}\displaystyle\int_{-\infty}^{+\infty}\lambda(\mathbf{n})|\varphi_0^{\rm in}(r,\cdot,\cdot)|^2\,dr=\hat{\psi}_{\mathcal{S}}(\mathbf{n}),\\[2ex]\displaystyle\int_{-\infty}^{+\infty}\frac{\partial\lambda(\mathbf{n})}{\partial\mathbf{n}}|\varphi_0^{\rm in}(r,\cdot,\cdot)|^2\,dr=2\frac{\partial\hat{\psi}_{\mathcal{S}}(\mathbf{n})}{\partial\mathbf{n}}.\end{cases} \tag{11.10}$$

Finally, proceeding as in section 10.4, (11.4) yields at $O(\epsilon^{-1})$

$$\begin{aligned}|\varphi_0^{\rm in}|^2B(\mathbf{n},V_{\mathcal{S}})V_{\mathcal{S}}&=|\varphi_0^{\rm in}|^2\lambda(\mathbf{n})K_{\mathcal{S}}-\operatorname{div}_{\mathcal{S}}\left\{\frac{1}{2}|\varphi_0^{\rm in}|^2\frac{\partial\lambda(\mathbf{n})}{\partial\mathbf{n}}\right\}\\&\quad+\overline{\left\{f'(\varphi_0^{\rm in})\varphi_1^{\rm in}-\varphi_0^{\rm in}\varphi_1^{\rm in}\lambda(\mathbf{n})-\frac{1}{2}\varphi_0^{\rm in}\nabla\varphi_0^{\rm in}\cdot\frac{\partial\lambda(\mathbf{n})}{\partial\mathbf{n}}\right\}}+[\![\Psi]\!]\overline{g(\varphi_0^{\rm in})};\end{aligned} \tag{11.11}$$

by integrating (11.11) over r from $r=-\infty$ to $r=+\infty$ and utilizing the constitutive connections (11.8) and (11.9), the boundary and far-field conditions (10.4) and (10.5), and properties of f and g, we obtain

$$\hat{b}_{\mathcal{S}}(\mathbf{n},V_{\mathcal{S}})V_{\mathcal{S}}=\left\{\hat{\psi}_{\mathcal{S}}(\mathbf{n})\mathbf{P}+\frac{\partial^2\hat{\psi}_{\mathcal{S}}(\mathbf{n})}{\partial\mathbf{n}^2}\right\}\cdot\mathbf{L}+[\![\Psi]\!], \tag{11.12}$$

which, on performing a suitable redimensionalization, is precisely the general interfacial evolution equation (1.1) of the sharp-interface theory.

12. Discussion. Our results are predicated on the provision of constitutive equations within the phase-field theory and, moreover, upon stipulated connections (11.8) and (11.9) between those constitutive equations and the constitutive equations of the sharp-interface theory. However, because the framework of the phase-field theory is a dynamical one that allows for dissipation, the results of our analysis are more broadly applicable than any alternative based on variational methods.

By restricting our attention to a setting where the standard force balance is irrelevant, we leave open the possibility that, once the constraint of rigidity is relaxed

and the standard stress is no longer indeterminate, the standard force balance might somehow give rise to the interfacial configurational force balance and, thus, to a law governing the evolution of the interface. However, such an outcome would be at odds with the implications of variationally based descriptions. Indeed, in considering phase interfaces within the context of the theory of finite elastostatics, Podio-Guidugli [30] shows that, while both the bulk configurational force balance and the tangential component of the interfacial configurational force balance are implied consequences of standard force balance, the normal component of the interfacial configurational force balance is independent. Moreover, asymptotic analyses of phase-field theories that account for deformation (Fried and Grach [11]; Fried and Gurtin [14]) demonstrate that the supplemental evolution equation of the sharp-interface theory arises not from the deformational force balance but, as in the simple theory considered here, either "indirectly" from the evolution equation for the phase field or "directly" from the associated configurational force balance.

REFERENCES

[1] D. R. Anderson, P. Cermelli, E. Fried, M. E. Gurtin, and G. B. McFadden, *General dynamical sharp-interface conditions for phase transformations in viscous heat-conducting fluids*, J. Fluid Mech., submitted.

[2] S. Angenent and M. E. Gurtin, *Multiphase thermodynamics with interfacial structure. 2. Evolution of an isothermal interface*, Arch. Ration. Mech. Anal., 108 (1989), pp. 323–391.

[3] G. Caginalp, *An analysis of a phase field model of a free boundary*, Arch. Ration. Mech. Anal., 92 (1986), pp. 205–245.

[4] P. Cermelli, E. Fried, and S. Sellers, *Configurational stress, yield, and flow in rate-independent plasticity*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 457 (2001), pp. 1447–1467.

[5] P. Cermelli and E. Fried, *The evolution equation for a disclination in a nematic liquid crystal*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 458 (2002), pp. 1–20; *Erratum*, 458 (2002), pp. 3090–3091.

[6] P. Cermelli, E. Fried, and M. E. Gurtin, *Sharp-interface nematic-isotropic phase transitions without flow*, Arch. Ration. Mech. Anal., 174 (2004), pp. 151–178.

[7] P. Cermelli and M. E. Gurtin, *The motion of screw dislocations in crystalline materials undergoing antiplane shear: Glide, cross-slip, fine cross-slip*, Arch. Ration. Mech. Anal., 148 (1999), pp. 3–52.

[8] J. D. Eshelby, *The force on an elastic singularity*, Philos. Trans. Roy. Soc. Lond. Ser. A, 244 (1951), pp. 84–112.

[9] J. D. Eshelby, *The continuum theory of lattice defects*, in Solid State Physics. Advances in Research and Applications. Vol. 3, F. Seitz and D. Turnbull, eds., Academic Press, New York, 1956, pp. 79–144.

[10] J. D. Eshelby, *The energy-momentum tensor of complex continua*, in Continuum Models of Discrete Systems, E. Kröner and K. H. Anthony, eds., University of Waterloo Press, Waterloo, 1980, pp. 651–688.

[11] E. Fried and G. Grach, *An order-parameter based theory as a regularization of a sharp-interface theory for solid-solid phase transitions*, Arch. Ration. Mech. Anal., 138 (1997), pp. 355–404.

[12] E. Fried and M. E. Gurtin, *Continuum theory of thermally induced phase transitions based on an order parameter*, Phys. D, 68 (1993), pp. 326–343.

[13] E. Fried and M. E. Gurtin, *A phase-field theory for solidification based on a general anisotropic sharp-interface theory with interfacial energy and entropy*, Phys. D, 91 (1996), pp. 143–181.

[14] E. Fried and M. E. Gurtin, *Coherent solid-state phase transitions with atomic diffusion: A thermomechanical treatment*, J. Statist. Phys., 95 (1999), pp. 1361–1427.

[15] E. Fried and M. E. Gurtin, *The role of configurational force balance in the nonequilibrium epitaxy of films*, J. Mech. Phys. Solids, 51 (2003), pp. 487–517.

[16] E. Fried and M. E. Gurtin, *A unified treatment of evolving interfaces accounting for deformation and atomic transport with emphasis on grain-boundaries and epitaxy*, Adv. Appl. Mech., 40 (2004), pp. 1–177.

[17] M. E. Gurtin and M. E. Jabbour, *Interface evolution in three dimensions with curvature-dependent energy and surface diffusion: Interface-controlled evolution, phase transitions, epitaxial growth of elastic films*, Arch. Ration. Mech. Anal., 163 (2002), pp. 171–208.
[18] M. E. Gurtin, *Multiphase thermomechanics with interfacial structure 1. Heat conduction and the capillary balance law*, Arch. Ration. Mech. Anal., 104 (1988), pp. 195–221.
[19] M. E. Gurtin, *The nature of configurational forces*, Arch. Ration. Mech. Anal., 131 (1995), pp. 67–100.
[20] M. E. Gurtin, *Configurational Forces as Basic Concepts of Continuum Physics*. Springer, New York, 2000.
[21] M. E. Gurtin and P. Podio-Guidugli, *Configurational forces and the basic laws for crack propagation*, J. Mech. Phys. Solids, 44 (1996), pp. 905–927.
[22] M. E. Gurtin and P. Podio-Guidugli, *Configurational forces and a constitutive theory for crack propagation that allows for curving and kinking*, J. Mech. Phys. Solids, 46 (1998), pp. 1343–1378.
[23] M. E. Gurtin and A. Struthers, *Multiphase thermomechanics with interfacial structure 3. Evolving phase boundaries in the presence of bulk deformation*, Arch. Ration. Mech. Anal., 112 (1990), pp. 97–160.
[24] M. E. Gurtin and M. Shvartsman, *Configurational forces and the dynamics of planar cracks in three-dimensional bodies*, J. Elasticity, 48 (1997), pp. 167–191.
[25] C. Herring, *Surface tension as a motivation for sintering*, in The Physics of Powder Metallurgy, W. E. Kingston, ed., McGraw-Hill, New York, 1951, pp. 143–179.
[26] V. K. Kalpakides and C. Dascalu, *On the configurational force balance in thermomechanics*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 458 (2002), pp. 3023–3039.
[27] G. A. Maugin, *Material Inhomogeneities in Elasticity*. Chapman and Hall, London, 1993.
[28] W. W. Mullins, *Two-dimensional motion of idealized grain boundaries*, J. Appl. Phys., 27 (1956), pp. 900–904.
[29] M. O. Peach and J. S. Koehler, *The forces exerted on dislocations and the stress fields produced by them*, Phys. Rev., 80 (1950), pp. 436–439.
[30] P. Podio-Guidugli, *Configurational balances via variational arguments*, Interfaces Free Bound., 3 (2001), pp. 223–232.
[31] P. Podio-Guidugli, *Configurational forces: Are they needed?*, Mech. Res. Comm., 29 (2002), pp. 513–519.
[32] N. K. Simha and K. Bhattacharya, *Kinetics of phase boundaries with edges*, J. Mech. Phys. Solids, 46 (1998), pp. 2323–2359.
[33] J. Taylor, J. W. Cahn, and C. A. Handwerker, *Geometrical models of crystal growth*, Acta Metallurgica, 40 (1992), pp. 1443–1474.
[34] M. Uhuwa, *Asymptotic growth shapes developed from two-dimensional nuclei*, J. Crystal Growth, 80 (1987), pp. 84–90.

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1150–1162

ON THE FUNDAMENTAL DIAGRAM OF TRAFFIC FLOW*

FLORIAN SIEBEL† AND WOLFRAM MAUSER†

Abstract. We present a new fluid-dynamical model of traffic flow. This model generalizes the model of Aw and Rascle [*SIAM J. Appl. Math.*, 60 (2000), pp. 916–938] and Greenberg [*SIAM J. Appl. Math.*, 62 (2001), pp. 729–745] by prescribing a more general source term to the velocity equation. This source term can be physically motivated by experimental data, when taking into account relaxation and reaction time. In particular, the new model has a (linearly) unstable regime as observed in traffic dynamics. We develop a numerical code that solves the corresponding system of balance laws. Applying our code to a wide variety of initial data, we find the observed inverse-λ shape of the fundamental diagram of traffic flow.

Key words. macroscopic traffic model, instability, system of balance laws, high-resolution shock-capturing schemes

AMS subject classifications. 35L65, 90B20, 70K50

DOI. 10.1137/050627113

1. Introduction.

After two-equation models of traffic flow were seriously criticized by Daganzo [5], the main focus of the traffic community shifted toward microscopic models of traffic flow. However, the criticism has been overcome; see, e.g., [26, 11]. By replacing the space derivate in old two-equation models by the convective derivative, Aw and Rascle [2] and Greenberg [8] deduced a two-equation model which solves all inconsistencies of the earlier models, as they showed with a detailed mathematical analysis and numerical simulations. In particular, in their model (which we call the ARG model), no information travels faster than the vehicle velocity; i.e., in general drivers do not react to the traffic situation behind them. Moreover, the velocity is always nonnegative. In the ARG model, traffic flow is described by the following system of balance laws determining the density $\rho = \rho(t,x)$ and velocity $v = v(t,x)$ of cars:

$$\frac{\partial \rho}{\partial t} + \frac{\partial (\rho v)}{\partial x} = 0, \tag{1.1}$$

$$\frac{\partial (\rho (v - u(\rho)))}{\partial t} + \frac{\partial (\rho v (v - u(\rho)))}{\partial x} = \frac{\rho (u(\rho) - v)}{T}. \tag{1.2}$$

As usual, (t,x) denote the time and space variables. $u(\rho)$ denotes the *equilibrium velocity*, which fulfills the following conditions:

$$u'(\rho) < 0 \quad \text{for } 0 \le \rho \le \rho_m, \tag{1.3}$$

$$\frac{d^2(\rho u(\rho))}{d\rho^2} < 0 \quad \text{for } 0 \le \rho \le \rho_m, \tag{1.4}$$

with the maximum vehicle density ρ_m. $T > 0$ is an additional parameter, the *relaxation time*. In the formal limit $T \to 0$ the ARG model reduces to the classic Lighthill–Whitham–Richards model [16, 19, 25]. For smooth solutions, the ARG model can be

*Received by the editors March 18, 2005; accepted for publication (in revised form) December 1, 2005; published electronically March 24, 2006.
http://www.siam.org/journals/siap/66-4/62711.html
†Department of Earth and Environmental Sciences, University of Munich, Luisenstraße 37, D-80333 Munich, Germany (f.siebel@iggf.geo.uni-muenchen.de, w.mauser@iggf.geo.uni-muenchen.de).

rewritten as

$$\frac{\partial \rho}{\partial t} + v\frac{\partial \rho}{\partial x} + \rho\frac{\partial v}{\partial x} = 0, \tag{1.5}$$

$$\frac{\partial v}{\partial t} + (v + \rho u'(\rho))\frac{\partial v}{\partial x} = \frac{u(\rho) - v}{T}. \tag{1.6}$$

In our opinion the ARG model still has a drawback; i.e., it cannot explain the growth of structures and the general behavior for congested traffic, as observed in traffic dynamics (see, e.g., [20, 10, 12]). To see this, we consider a linear stability analysis around the equilibrium solution $\rho(t,x) = \rho_0$, $v(t,x) = u(\rho_0)$, i.e.,

$$\rho(t,x) = \rho_0 + \tilde{\rho}\exp(ikx + \omega(k)t), \tag{1.7}$$

$$v(t,x) = u(\rho_0) + \tilde{v}\exp(ikx + \omega(k)t). \tag{1.8}$$

Substituting this ansatz into system (1.5)–(1.6) we obtain

$$\begin{pmatrix} \omega + iku & ik\rho_0 \\ -\frac{u'}{T} & \omega + \frac{1}{T} + ik(u + \rho_0 u') \end{pmatrix} \begin{pmatrix} \tilde{\rho} \\ \tilde{v} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{1.9}$$

Nontrivial solutions of this linear system exist if and only if

$$(\omega + iku)\Big(\omega + \frac{1}{T} + ik(u + \rho_0 u')\Big) + ik\frac{\rho_0 u'}{T} = 0 \tag{1.10}$$

or, equivalently, for

$$\omega_1 = -ik(u + \rho_0 u'), \tag{1.11}$$

$$\omega_2 = -\frac{1}{T} - iku. \tag{1.12}$$

For the stability properties, the real parts of the above solutions are important; i.e.,

$$Re(\omega_1) = 0, \tag{1.13}$$

$$Re(\omega_2) = -\frac{1}{T}. \tag{1.14}$$

For $T > 0$ both real parts are nonpositive, which means that the ARG model is linearly stable, and the velocity v relaxes to the equilibrium velocity u in the entire region $0 \le \rho \le \rho_m$. This is clearly in contrast to observations, where a wide range of states in the fundamental diagram, the relation between vehicle flux and the density, are observed for congested traffic flow. To correct this defect, Greenberg, Klar, and Rascle developed an extended model with two equilibrium velocities [9]. In the present paper, we propose an alternative model, which takes into account the reaction times of drivers (as well as mechanical restrictions).

We give a physical argument for our new model and define it in section 2. Section 3 presents the methods used for numerically solving the model equations. Section 4 describes tests to validate our numerical algorithm, before we finally discuss the numerical results on the fundamental diagram obtained with our model in section 5.

2. A heuristic derivation of the new model. Before we turn to the new model, let us first give a simple derivation of the ARG model. Note that the model was mathematically derived from a car-following theory in [1]. Suppose that in the

reference frame of individual drivers, drivers adjust their speed v in such a way that they asymptotically approach the equilibrium velocity u; i.e.,

$$\frac{d(v-u)}{dt} = \frac{u-v}{T}. \tag{2.1}$$

Here, $T = \text{const} > 0$ is the relaxation time. In comparison to optimal velocity models (see, e.g., [3]) the equilibrium velocity term on the left has been added which vanishes for $u = \text{const}$. It is easy to verify that the analytical solution of the ordinary differential equation (2.1) reads

$$v(t) = u(t) + (v(0) - u(0)) \exp\left(-\frac{t}{T}\right). \tag{2.2}$$

In the coordinate system of the road, (2.1) translates to

$$\frac{\partial(v-u)}{\partial t} + v\frac{\partial(v-u)}{\partial x} = \frac{u-v}{T}. \tag{2.3}$$

Moreover, since

$$-\left(\frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x}\right) = -u'\left(\frac{\partial \rho}{\partial t} + v\frac{\partial \rho}{\partial x}\right) = \rho u'\frac{\partial v}{\partial x}, \tag{2.4}$$

where we have used the continuity equation (1.5) for the last equality, we recover the velocity equation of the ARG model (1.6). From this derivation, it is obvious that drivers instantaneously react to the current traffic situation.

We therefore tried to generalize (2.1) and took the reaction time of drivers $\tau > 0$ into account:

$$\frac{dv}{dt}(t,x) - \frac{du}{dt}(\rho(t-\tau, x-v\tau)) = \frac{u(\rho(t-\tau,x-v\tau)) - v(t-\tau,x-v\tau)}{T}. \tag{2.5}$$

Using a Taylor series expansion in τ and keeping only terms up to order 0 in τ and T, i.e.,

$$\frac{du}{dt}(\rho(t-\tau,x-v\tau)) = \frac{\partial u(\rho(t,x))}{\partial t} + v\frac{\partial u(\rho(t,x))}{\partial x} + O^1(\tau,T), \tag{2.6}$$

$$u(\rho(t-\tau,x-v\tau)) = u(\rho(t,x)) - \tau u'(\rho(t,x))\left(\frac{\partial \rho}{\partial t} + v\frac{\partial \rho}{\partial x}\right) + O^2(\tau,T)$$

$$= u + \rho u'\frac{\partial v}{\partial x}\tau + O^2(\tau,T), \tag{2.7}$$

$$v(t-\tau,x-v\tau) = v(t,x) - \tau\left(\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x}\right) + O^2(\tau,T), \tag{2.8}$$

we find

$$\frac{\partial v}{\partial t} + (v + \rho u'(\rho))\frac{\partial v}{\partial x} = \frac{u(\rho) - v}{T - \tau}. \tag{2.9}$$

This equation is identical to the velocity equation of the ARG model (1.6), except that the relaxation time T has been replaced by $T - \tau$. In particular it follows from the stability analysis of the ARG model that for $\tau > T$ the new system is (linearly) unstable.

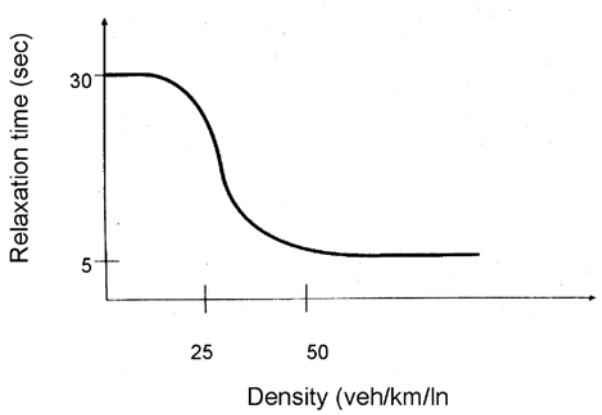


FIG. 1. *Dependence of the relaxation time $\tilde{T}$ on the vehicle density per lane. Reprinted with permission from* [14].

Before we look at the experimental data on the relaxation and reaction times, we remark that it is tempting to include an anticipation length in the model, as, e.g., in [23, 17]. This approach has not been followed here for two reasons: First, the ARG model already includes anticipatory elements, as noted by [8]. Second, including the anticipation length in the above derivation yields a system that does not guarantee that the maximum speed at which information travels is bounded from above by the velocity of cars, and is therefore unrealistic.

For the reaction time τ, which includes the human perception time as well as the time it takes to realize the reaction (hence τ is also called perception-reaction time), typical values [13] are of the order

$$\tau \approx 0.5 \,..\, 2 \text{ s}. \tag{2.10}$$

Figure 1 shows experimental results for the relaxation time $\tilde{T}$ taken from [14]. However, these values have to be interpreted with care and cannot be translated directly to our model context. To see this, we note that the relaxation time $\tilde{T}$ is determined for the ansatz $v(t+\tilde{T}, x) = u(\rho(t, x+\Delta x))$. In this expression, $x+\Delta x$ denotes the anticipated location, where according to [14], $\Delta x = -\tilde{T} c_0^2(\rho) u'(\rho)^{-1} \rho^{-1}$, with an anticipation coefficient $c_0^2(\rho)$ corresponding to the standard deviation of the vehicular speed distribution, which for simplicity is often set to a constant value. In this approach, the drivers have fully adjusted their velocities to the equilibrium velocity u after the relaxation time $\tilde{T}$. Here, according to (2.1) and (2.2), the equilibrium velocity in general will never be reached exactly. Instead, if we require that $|v(t) - u(t)| < |v(0) - u(0)|/1000$, we find that $t > 6.908\, T \approx \tilde{T}$. Hence it seems reasonable to set

$$\tilde{T} \approx 5 \,..\, 10\, T. \tag{2.11}$$

With typical values $\tau = 1$ s and $\tilde{T} = 7.5\, T$, we indeed find that $T - \tau < 0$ for about $\rho > 40$ [1/km/lane]. It was also pointed out in [14] that for large densities the relaxation time $\tilde{T}$ increases, which we interpret as $T - \tau \geq 0$ for $\rho \approx \rho_m$. We remark that the precise form in which the reaction time and anticipation effects are included in the traffic equations differs for different traffic models; see, e.g., [17, 6, 18].

One could try to repeat the derivation leading to (2.9) for a general relaxation time $T = T(\rho, v)$. Note that the above derivation is valid only for a constant relaxation time. Moreover, it involves only the leading term of a Taylor series expansion. We therefore decided to generalize the velocity equation of the ARG model in the following way:

$$\frac{\partial v}{\partial t} + (v + \rho u'(\rho))\frac{\partial v}{\partial x} = \beta(\rho, v)(u(\rho) - v). \tag{2.12}$$

Note that we do not require $v \leq u$ as in Greenberg [8]. From the experimental data and the argument put forward before (note that the sign of β determines whether the traffic flow is linearly stable or not), we require

$$\beta(\rho, v) < 0 \quad \text{for } 0 < \rho_1 < \rho < \rho_2 \leq \rho_m,\ v = u(\rho), \tag{2.13}$$

$$\lim_{\rho \to 0, \rho_m} \beta(\rho, v) \geq 0, \tag{2.14}$$

$$\lim_{v \to 0, u_m = u(0)} \beta(\rho, v) \geq 0. \tag{2.15}$$

The last condition is necessary, as the system would otherwise be driven to negative or arbitrarily large vehicle velocities, which is clearly unrealistic. Throughout this paper we use a functional form

$$\beta(\rho, v) = \begin{cases} \frac{a_c}{u - v} & \text{if } \tilde{\beta}(\rho, v)(u - v) - a_c \geq 0, \\ \frac{d_c}{u - v} & \text{if } \tilde{\beta}(\rho, v)(u - v) - d_c \leq 0, \\ \tilde{\beta}(\rho, v) & \text{else,} \end{cases} \tag{2.16}$$

where the function $\tilde{\beta}(\rho, v)$ is defined as

$$\tilde{\beta} = \frac{1}{\hat{T}}\left(1 + \alpha\frac{|u - v|}{u_m} + \frac{1}{\rho_1 \rho_2}(-(\rho_1 + \rho_2)\rho + \rho^2)\right). \tag{2.17}$$

For $v = u(\rho)$, the function $\tilde{\beta}$ reduces to a parabola with zeros ρ_1 and ρ_2 in accordance with conditions (2.13) and (2.14). The term involving the velocity is added in order to fulfill condition (2.15). In the following we specify the free functions and parameters for a two-lane highway. For the choice of the velocity-density relation of Cremer [4],

$$u(\rho) = u_m\left(1 - \left(\frac{\rho}{\rho_m}\right)^{n_1}\right)^{n_2}, \tag{2.18}$$

and the parameters $\rho_m = 300$ [1/km], $u_m = 140$ km/h, $n_1 = 0.35$, $n_2 = 1$ (note that with these parameters, the equilibrium velocity of Cremer (2.18) fulfills the conditions (1.3) and (1.4)), $\hat{T} = 1$ s, $\alpha = 12$, $\rho_1 = 70$ [1/km], and $\rho_2 = 270$ [1/km], the function $\tilde{\beta}(\rho, v)$ already fulfills all requirements (2.13)–(2.15). However, due to mechanical restrictions, the maximum acceleration a_c and deceleration d_c give stronger limitations, i.e.,

$$\frac{dv}{dt} \leq a_c \qquad \text{and} \qquad \frac{dv}{dt} \geq d_c \tag{2.19}$$

with typical values $a_c = 2$ m/s^2 and $d_c = -5$ m/s^2. Since the resulting system is not strictly hyperbolic for equality in (2.19), which is problematic for a numerical solution, we prescribe the limitations on

$$\frac{d(v - u)}{dt} \leq a_c \qquad \text{and} \qquad \frac{d(v - u)}{dt} \geq d_c, \tag{2.20}$$

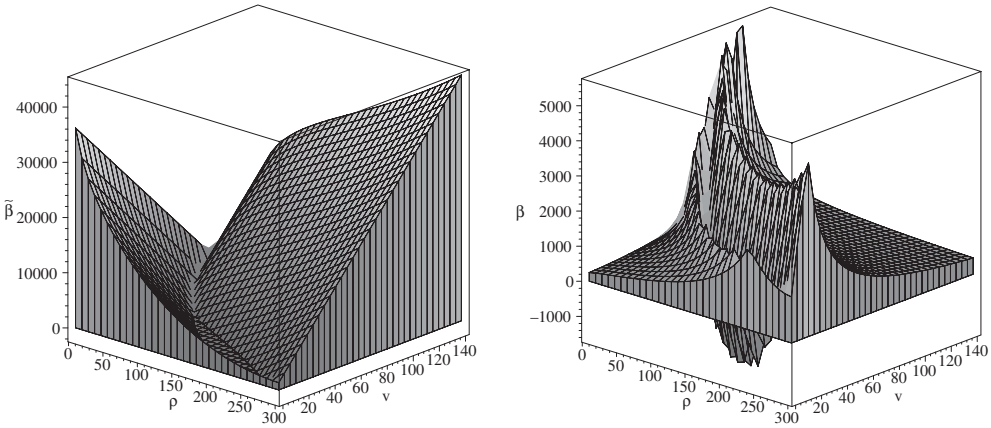


FIG. 2. *Functions $\tilde{\beta}(\rho, v)$ (left panel) and $\beta(\rho, v)$ (right panel) defined in* (2.17) *and* (2.16). *We used units* $[\rho] = 1/\text{km}, [v] = \text{km/h}$, *and* $[\tilde{\beta}] = [\beta] = 1/\text{h}$. *The function $\tilde{\beta}(\rho, v)$ reaches very large values leading to unrealistic accelerations and decelerations according to* (2.12). *For the function $\beta(\rho, v)$ mechanical restrictions were taken into account, which change the functional shape but not the sign.*

which then leads to the functional form (2.16). We plot the functions $\tilde{\beta}(\rho, v)$ and $\beta(\rho, v)$ for the mentioned parameter values in Figure 2. We stress that the above functions describe reality only qualitatively. For realistic simulations of traffic flow, experimental data are required to determine $\beta(\rho, v)$.

In the new model, traffic flow is described by the system of balance laws

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho v)}{\partial x} = 0, \tag{2.21}$$

$$\frac{\partial(\rho(v - u(\rho)))}{\partial t} + \frac{\partial(\rho v(v - u(\rho)))}{\partial x} = \beta\rho(u - v) \tag{2.22}$$

or, equivalently, for smooth solutions by (1.5) and (2.12). As the corresponding system of the ARG model, the new system is strictly hyperbolic for $0 < \rho \leq \rho_m$.

3. The numerical implementation. Writing traffic flow as a system of balance laws in (2.21) and (2.22) is very adequate for numerical purposes, as it allows the application of well-established hydrodynamic methods for the numerical solution. We use a high-resolution shock-capturing scheme with an approximate Riemann solver for the numerical solution (see, e.g., [15]).

We rewrite (2.21) and (2.22) in the form

$$\frac{\partial U}{\partial t} + \frac{\partial F(U)}{\partial x} = S(U), \tag{3.1}$$

where

$$U = \begin{pmatrix} \rho \\ \rho(v - u) \end{pmatrix} = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}, \tag{3.2}$$

$$F(U) = \begin{pmatrix} \rho v \\ \rho v(v - u) \end{pmatrix} = \begin{pmatrix} U_2 + U_1 u(U_1) \\ \frac{U_2^2}{U_1} + U_2 u(U_1) \end{pmatrix}, \tag{3.3}$$

$$S(U) = \begin{pmatrix} 0 \\ \beta\rho(u - v) \end{pmatrix}. \tag{3.4}$$

We use the second-order reconstruction scheme of van Leer [24] to reconstruct quantities at cell interfaces. At cell i with cell center at the location $x_i = x_0 + i\Delta x$ the update in time from t^n to t^{n+1} is performed according to a conservative algorithm

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x}(\hat{F}_{i+\frac{1}{2}} - \hat{F}_{i-\frac{1}{2}}) + \Delta t S_i, \tag{3.5}$$

where $U_i^n = U(t^n, x_i)$ and $\Delta t = t^{n+1} - t^n$. To obtain a higher order of convergence, we use the third-order scheme of Shu and Osher [21]. The numerical fluxes $\hat{F}$ are determined according to the flux formula of Marquina [7], which reads

$$\hat{F} = \frac{1}{2}(F^R + F^L - \Delta q). \tag{3.6}$$

Here, the superscripts R and L denote the reconstructed values on the right and left of a cell interface. The numerical viscosity term takes the form

$$\Delta q = \mathbf{R}^R |\Lambda|_{\max} \mathbf{L}^R U^R - \mathbf{R}^L |\Lambda|_{max} \mathbf{L}^L U^L. \tag{3.7}$$

The matrix $|\Lambda|_{\max}$ involves the characteristic speeds

$$|\Lambda|_{\max} = \begin{pmatrix} \max(|\lambda_1^R|, |\lambda_1^L|) & 0 \\ 0 & \max(|\lambda_2^R|, |\lambda_2^L|) \end{pmatrix}, \tag{3.8}$$

where the characteristic speeds read explicitly as

$$\lambda_1 = v + \rho u', \tag{3.9}$$

$$\lambda_2 = v. \tag{3.10}$$

$\mathbf{R}$ and $\mathbf{L}$ are the matrices of the right and left eigenvectors of the matrix

$$\frac{\partial F}{\partial U} = \begin{pmatrix} u + \rho u' & 1 \\ -(v-u)^2 + \rho(v-u)u' & 2v - u \end{pmatrix}. \tag{3.11}$$

Explicitly,

$$\mathbf{R} = \begin{pmatrix} 1 & 1 \\ v - u & v - u - \rho u' \end{pmatrix}, \tag{3.12}$$

$$\mathbf{L} = \frac{1}{\rho u'} \begin{pmatrix} u - v + \rho u' & 1 \\ v - u & -1 \end{pmatrix}. \tag{3.13}$$

4. Code tests. We checked that our numerical algorithm is convergent. Moreover, the density equation (2.21) is a strict conservation law. Prescribing periodic boundary conditions as in section 5, the total number of cars included in the numerical domain Ω should therefore be constant; i.e.,

$$\int_\Omega \rho \, dx = \text{const}. \tag{4.1}$$

We checked that our numerical code fulfills (4.1) up to machine precision (see also the corresponding results for a network simulation based on the Lighthill–Whitham theory in [22]).

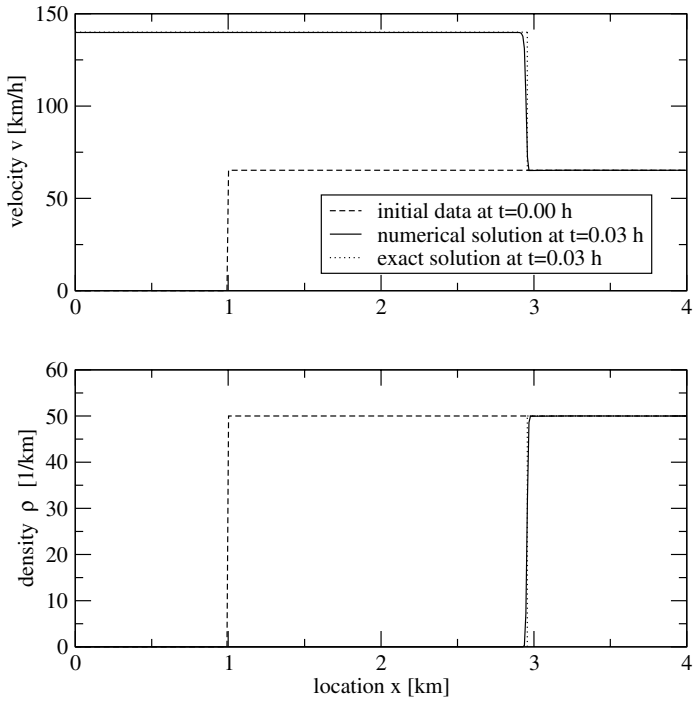


FIG. 3. *Numerical solution for a Riemann problem of Aw and Rascle* [2]. *The numerical solution at time* $t = 0.03$ h *(solid line) for the initial data (dashed line) reproduces the exact solution (dotted line).*

Finally, Aw and Rascle presented in their paper the exact solution of a Riemann problem, for which old two-equation models fail to describe the correct behavior (see [2, Figure 5.4]). This Riemann problem consists of the following initial data:

$$\rho = \begin{cases} 0 & \text{if } x < 1 \text{ km}, \\ \rho_+ & \text{if } x \geq 1 \text{ km}, \end{cases} \tag{4.2}$$

$$v = \begin{cases} 0 & \text{if } x < 1 \text{ km}, \\ v_+ & \text{if } x \geq 1 \text{ km}. \end{cases} \tag{4.3}$$

The exact solution to the homogeneous system consists of the constant state on the right (ρ_+, v_+) moving to the right with velocity v_+, leaving behind a vacuum. If we choose $v_+ = u(\rho_+)$, this exact solution will carry over to our inhomogeneous system. Figure 3 displays our numerical solution for a choice $\rho_+ = 50$ [1/km]. For numerical reasons, we prescribe a density $\rho = 10^{-6}$ [1/km] for $x < 1$ km. Note that our numerical algorithm resolves the steep gradient within only a few grid cells, at the same time reproducing the correct velocity at which the constant state moves to the right. Moreover, the velocity relaxes to the equilibrium velocity behind the constant state (ρ_+, v_+).

5. Results on the fundamental diagram. For the results presented in this section we restrict the calculation to a 7 km long section of a (two-lane) highway with periodic boundary conditions. On this section of the highway, we start our simulations

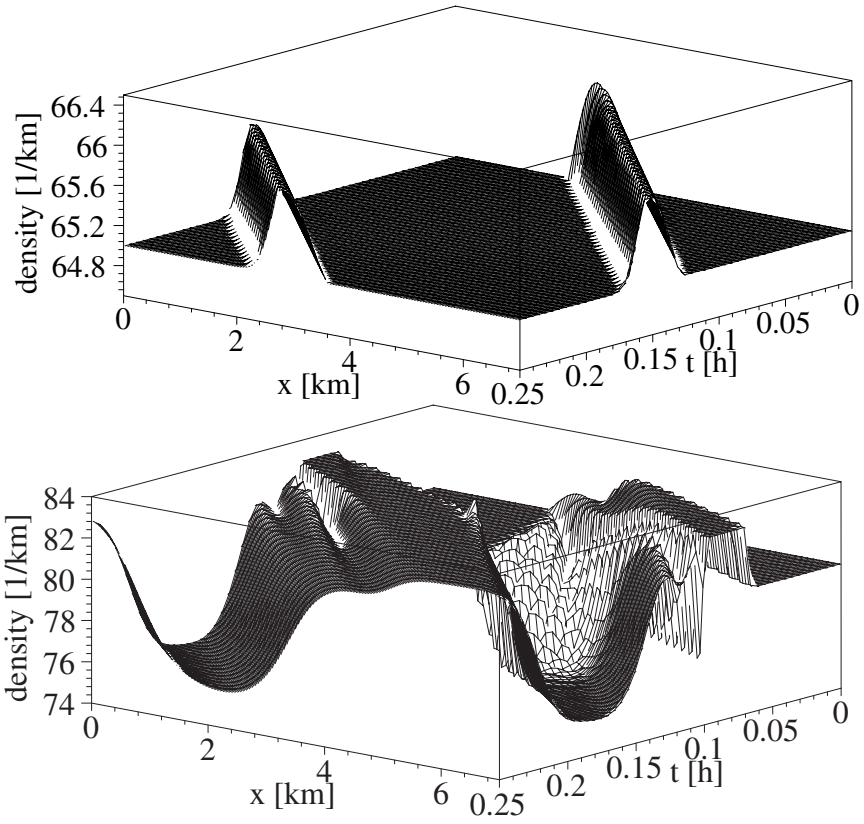


FIG. 4. *Time evolution of the density for stable and unstable initial data. We prescribe an equilibrium density* $\rho_0 = 65$ [1/km] *and* $\rho_0 = 80$ [1/km]*, respectively, and in addition a sinusoidal density perturbation. For the unstable data, the initial perturbation located at* $x = 2.5$ [km] *is amplified and finally two clusters form.*

with constant equilibrium data $\rho = \rho_0$, $v = u(\rho_0)$ and, in addition, between kilometers 2 and 3 a sinusoidal density perturbation

$$\Delta\rho = \sin(\pi x) \quad \text{for } 2 < x < 3 \text{ km}. \tag{5.1}$$

For all numerical results presented we used a resolution of 50 m. Figure 4 shows the evolution of these data for parameters $\rho_0 = 65$[1/km] and $\rho_0 = 80$[1/km]. Whereas the amplitude of the perturbation is gradually damped with time for the stable initial data $\rho_0 = 65$[1/km], the amplitude of the perturbation increases for the unstable initial data $\rho_0 = 80$ [1/km]. Moreover, the perturbation travels with a larger velocity downstream in the first case. For the unstable situation, two clusters are forming. We plot the corresponding time evolutions of the velocity in Figure 5.

To obtain a more general picture we varied the initial density in the entire density regime and analyzed the resulting flow-density relation as a function of time. More precisely, we used initial values for the equilibrium data $\rho_0 = 2, 4, \ldots, 298$ and read off the resulting values for the density ρ and the flux function ρv at five equidistantly distributed cross sections of the highway. Figure 6 shows the results for evolution times $t = 0.00$ h (initial data), $t = 0.05$ h, $t = 0.10$ h, $t = 0.15$ h, $t = 0.20$ h, and $t = 0.25$ h. For the initial data, the flow-density curve closely corresponds to the

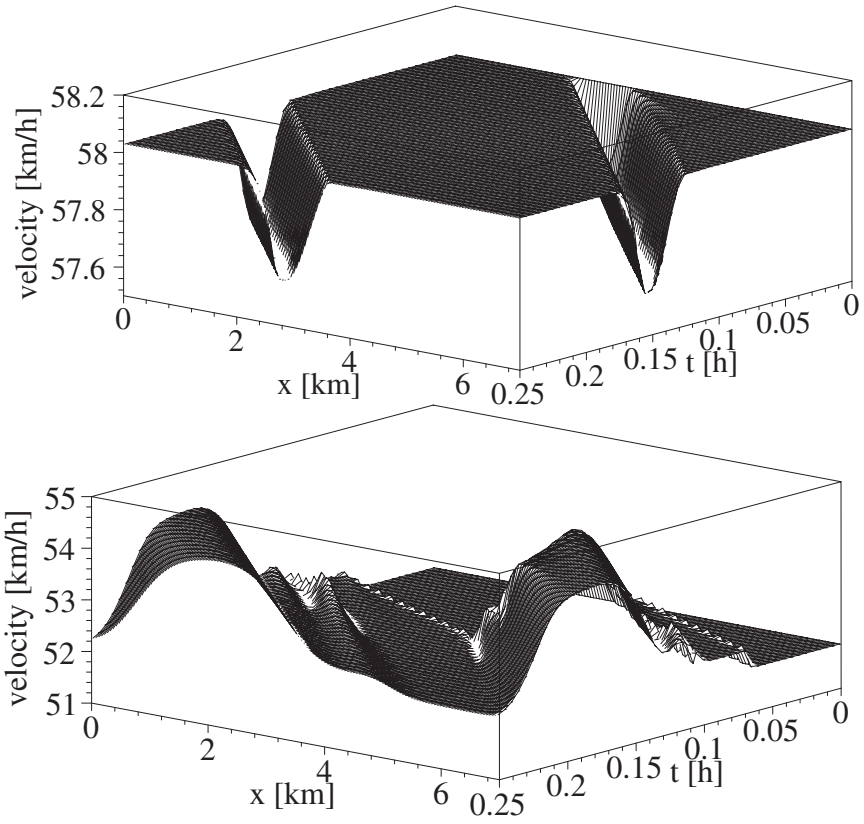


FIG. 5. *Time evolution of the velocity for stable ($\rho_0 = 65$ [1/km]) and unstable initial data ($\rho_0 = 80$ [1/km]) with initial density perturbation.*

equilibrium flow density, the initial perturbation (5.1) being negligible for the visual output. After an evolution time $t = 0.05$ h, the equilibrium flow-density curve is still visible, but in the unstable regime for densities $70 < \rho < 270$ [1/km] two new flow-density curves start to appear. In the evolution further in time, the equilibrium density curve vanishes. Instead the two new branches produce an inverse-λ shape.

6. Conclusion and outlook. We generalized the traffic model of Aw, Rascle, and Greenberg by prescribing a more general source term to the velocity equation and developed a new numerical code to solve the resulting system of balance laws. In total, our (numerical) results show the following:

- The new model can explain the large variance of the measured values of the fundamental diagram in the congested regime, which corresponds to fluctuations between two branches in the unstable density regime. Moreover, due to the stability properties, the model predicts oscillations in the relative velocity of cars in the congested regime, as they are found in experimental data. At the same time, it reproduces the small variance of velocities for free traffic flow and can explain the appearance of wide traffic jams.
- Macroscopic traffic models have often used an equilibrium velocity $u(\rho)$, for which $\frac{d(\rho u(\rho))}{d\rho^2} > 0$ in the congested regime, in order to account for the values of traffic flow at the maximum (the tip of the inverted λ). According to

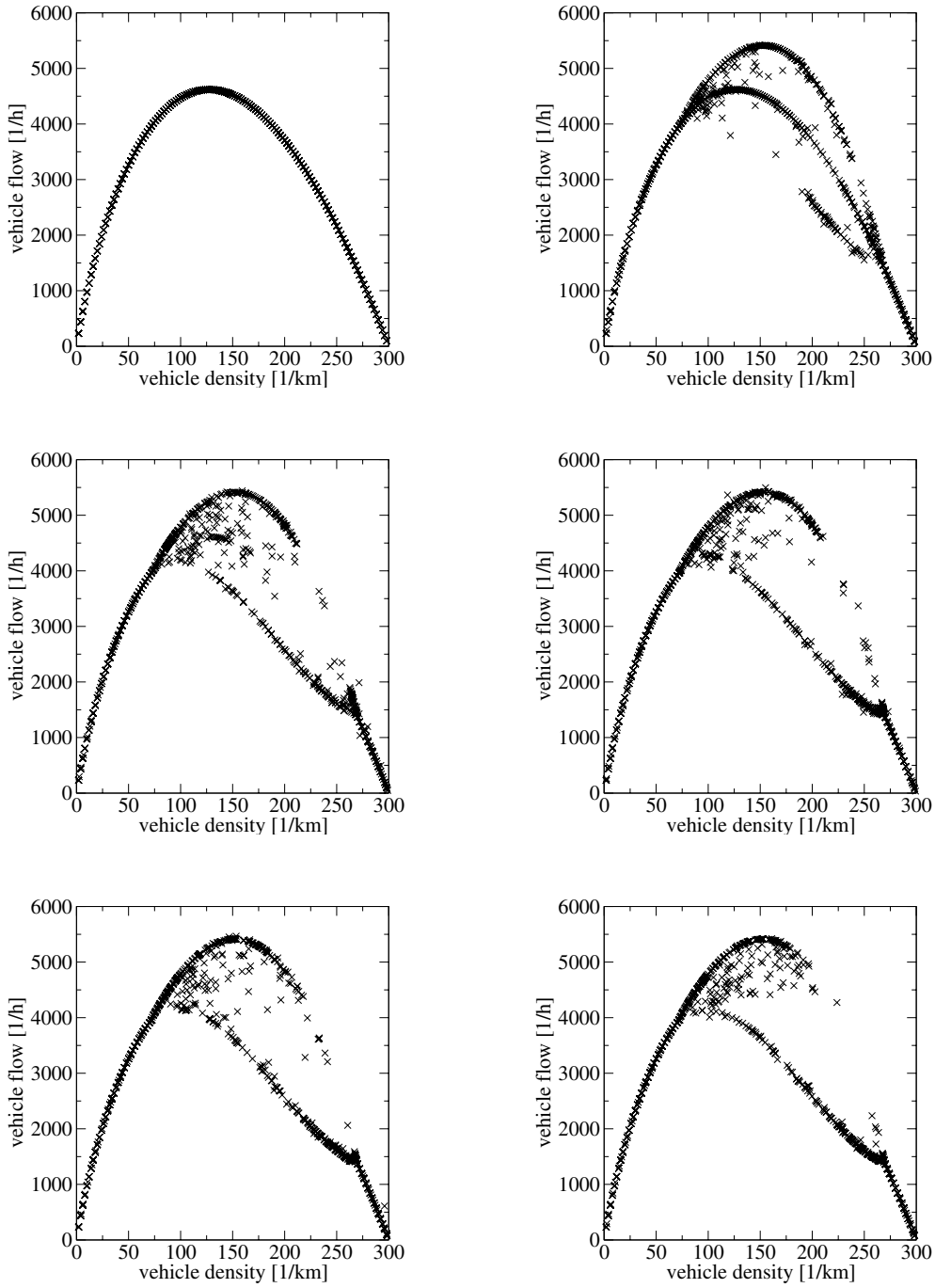


FIG. 6. *Fundamental diagram for the initial data ($t = 0.00$ h) (top left), for $t = 0.05$ h (top right), for $t = 0.10$ h (middle left), for $t = 0.15$ h (middle right), for $t = 0.20$ h (bottom left), and for $t = 0.25$ h (bottom right). At intermediate densities, our traffic model is linearly unstable; the representative points in the fundamental diagram are shifted toward two branches, which gives the visual impression of an inverted λ.*

our study, this is not necessary, as the high values for the fluxes can be explained with overcritical solutions and an equilibrium velocity function with $\frac{d(\rho u(\rho))}{d\rho^2} < 0$ everywhere.

The new model, which is a deterministic and effective one-lane model, has the capacity to reproduce many features observed in traffic dynamics. In the present work, the form of the function β in Figure 2 was motivated by a physical argument, but the quantitative details were determined rather ad hoc. However, we found that the fundamental diagram in the unstable region (e.g., the tip of the inverted λ) depends on the particular form of β. Hence one should try to determine the function β from experimental data of the fundamental diagram. In our opinion, the presented algorithm is adequate for use in network simulations of traffic flow.

REFERENCES

[1] A. Aw, A. Klar, T. Materne, and M. Rascle, *Derivation of continuum traffic flow models from microscopic follow-the-leader models*, SIAM J. Appl. Math., 63 (2002), pp. 259–278.

[2] A. Aw and M. Rascle, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.

[3] M. Bando, K. Hasebe, A. Nakayama, A. Shibata, and Y. Sugiyama, *Dynamical model of traffic congestion and numerical simulation*, Phys. Rev. E (3), 51 (1995), pp. 1035–1042.

[4] M. Cremer, *Der Verkehrsfluß auf Schnellstraßen (Traffic Flow on Freeways)*, Springer, Berlin, 1979.

[5] C.F. Daganzo, *Requiem for second-order fluid approximations of traffic flow*, Transportation Res. B, 29 (1995), pp. 277–286.

[6] L.C. Davis, *Multilane simulations of traffic phases*, Phys. Rev. E (3), 69 (2004), p. 016108.

[7] R. Donat and A. Marquina, *Capturing shock reflections: An improved flux formula*, J. Comput. Phys., 146 (1996), pp. 42–58.

[8] J.M. Greenberg, *Extensions and amplifications of a traffic model of Aw and Rascle*, SIAM J. Appl. Math., 62 (2001), pp. 729–745.

[9] J.M. Greenberg, A. Klar, and M. Rascle, *Congestion on multilane highways*, SIAM J. Appl. Math., 63 (2003), pp. 818–833.

[10] D. Helbing, *Traffic and related self-driven many-particle systems*, Rev. Modern Phys., 73 (2001), pp. 1067–1141.

[11] R. Jiang, Q.-S. Wu, and Q.-S. Zhu, *A new continuum model for traffic flow and numerical tests*, Transportation Res. B, 36 (2002), pp. 405–419.

[12] B.S. Kerner, *The Physics of Traffic*, Springer, Berlin, 2004.

[13] R.J. Koppa, *Human factors*, in Monograph on Traffic Flow Theory, U.S. Department of Transportation, Federal Highway Administration, Washington, DC, 1997, pp. (3)1–32; available online at http://www.tfhrc.gov/its/tft/tft.htm.

[14] R. Kühne and P. Michalopoulos, *Continuum flow models*, in Monograph on Traffic Flow Theory, U.S. Department of Transportation, Federal Highway Administration, Washington, DC, 1997, pp. (5)1–51; available online at http://www.tfhrc.gov/its/tft/tft.htm.

[15] R.J. LeVeque, *Numerical Methods for Conservation Laws*, 2nd ed., Birkhäuser-Verlag, Basel, 1992.

[16] M.J. Lighthill and G.B. Whitham, *On kinematic waves.* II. *A theory of traffic flow on long crowded roads*, Proc. Roy. Soc. London Ser. A, 229 (1955), pp. 317–345.

[17] P. Nelson, *Synchronized traffic flow from a modified Lighthill-Whitman model*, Phys. Rev. E (3), 61 (2000), p. R6052.

[18] G. Orosz, R.E. Wilson, and B. Krauskopf, *Global bifurcation investigation of an optimal velocity traffic model with driver reaction time*, Phys. Rev. E (3), 70 (2004), p. 026207.

[19] P.I. Richards, *Shock waves on the highway*, Oper. Res., 4 (1956), pp. 42–51.

[20] M. Schönhof and D. Helbing, *Empirical Features of Congested Traffic States and Their Implications for Traffic Modelling*, http://arxiv.org/pdf/cond-mat/0408138 (6 August 2004).

[21] C.-W. Shu and S. Osher, *Efficient implementation of essentially nonoscillatory shock-capturing schemes.* II, J. Comput. Phys., 83 (1989), pp. 32–78.

[22] F. Siebel and W. Mauser, *Simulating vehicular traffic in a network using dynamic routing*, Math. Comput. Model. Dynam. Syst., to appear.

[23] M. Treiber, A. Hennecke, and D. Helbing, *Derivation, properties, and simulation of a gas-kinetic-based, nonlocal traffic model*, Phys. Rev. E (3), 59 (1999), pp. 239–253.
[24] B.J. van Leer, *Towards the ultimate conservative difference scheme.* IV. *A new approach to numerical convection*, J. Comput. Phys., 23 (1977), pp. 276–299.
[25] G.B. Whitham, *Linear and Nonlinear Waves*, John Wiley & Sons, New York, 1974.
[26] H.M. Zhang, *A theory of nonequilibrium traffic flow*, Transportation Res. B, 32 (1998), pp. 485–498.

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1163–1187

THE EFFECT OF CONTACT LINES ON THE RAYLEIGH INSTABILITY WITH ANISOTROPIC SURFACE ENERGY*

K. F. GURSKI†, G. B. MCFADDEN‡, AND M. J. MIKSIS§

Abstract. We determine the linear stability of a rod or wire on a substrate subject to capillary forces arising from an anisotropic surface energy for a range of contact angles between $-\pi/2$ and $\pi/2$. The unperturbed rod is assumed to have infinite length with a uniform cross-section given by a portion of the two-dimensional equilibrium shape. We examine the effect of surface perturbations on the total energy. The stability of the equilibrium interface is reduced to determining the eigenvalues of a coupled system of ordinary differential equations. This system is solved both asymptotically and numerically for several types of anisotropic surface energies. We find that, in general, the presence of the substrate tends to stabilize the rod.

Key words. Rayleigh instability, contact lines, anisotropic surface energy, quantum wires, nanowires, Plateau

AMS subject classifications. 76E17, 34B24, 34D05, 34D10, 35R35, 65L15, 65L10

DOI. 10.1137/050626946

1. Introduction.

While stability studies of cylindrical rods have a long history, they are still a subject of current interest. For example, nanowires (alternatively called nanorods or quantum wires) are nanoscale crystal structures that are formed by deposition on a substrate, typically with a high lattice-mismatch that tends to produce aligned crystals on the substrate. The processing parameters that govern the growth and stability of the wires are of intense interest. Here we focus on two important issues surrounding nanowire stability: anisotropic surface tension and the contact angle between the rod and substrate.

The analysis of capillary driven instabilities spurring cylindrical rods to break up into droplets was initiated in 1873 by Plateau [27], who showed that breakup will occur when the rod, with isotropic surface energy, is subject to axisymmetric perturbations whose wavelength exceeds the circumference of the cylinder. Lord Rayleigh [28] determined that the length scale of the instability is controlled by the perturbations having the fastest temporal growth rate. The tendency for preferential beading has subsequently become known as the Rayleigh instability. A nice review by Michael may be found in [22]. Molares et al. [23] recently performed an experimental demonstration using the Rayleigh instability of a nanowire to produce long chains of nanospheres.

The surface energy of a liquid-solid or vapor-solid interface is generally anisotropic due to the underlying crystal lattice and depends on the orientation of the local normal

*Received by the editors March 16, 2005; accepted for publication (in revised form) December 15, 2005; published electronically March 24, 2006. The first author was supported by a National Research Council Postdoctoral Fellowship, and the second author was supported by the Microgravity Research Division of NASA. Part of this research was performed at Northwestern University with the support of the NSF Nanoscale Interdisciplinary Research Teams Program under grant DMR-0102794.

http://www.siam.org/journals/siap/66-4/62694.html

†National Institute of Standards and Technology, Gaithersburg, MD 20899-8910. Current address: Department of Mathematics, George Washington University, Washington, DC 20052 (kgurski@gwu.edu).

‡National Institute of Standards and Technology, Gaithersburg, MD 20899-8910 (mcfadden@nist.gov).

§Northwestern University, Department of Engineering Sciences and Applied Mathematics, Evanston, IL 60208 (miksis@northwestern.edu).

vector at each point of the interface [12, 25, 30]. As a first step in applying the analysis of Plateau to a nanowire, the effects of the substrate may be ignored, and the stability of the isolated rod determined strictly by the consideration of anisotropic surface energy. The experimental studies of Kondo and Takayanagi [18] show an apparent stability of elongated nanowires that are grown in a bridge configuration, in contrast to the expected nanoislands or quantum dots predicted by the Rayleigh instability for the isotropic case.

Cahn [2] studied the effect of anisotropic surface energy on the Rayleigh instability for isolated rods with circular cross-sections that are subject to axisymmetric perturbations; the underlying surface free energy was assumed to have transverse isotropy, resulting in closed-form solutions to the stability problem. Glaeser and Stölken [9, 32] extended Cahn's analysis and evaluated the effect of the axisymmetric surface energy anisotropy on evolution kinetics. Gurski and McFadden [10] considered general anisotropic surface energies by computing the second variation of the surface free energy of a freestanding rod whose cross-section is smooth and given by a two-dimensional equilibrium shape. The analysis was applied to examples with uniaxial or cubic anisotropy, which illustrated that anisotropic surface energy plays a significant role in establishing the stability of the rod. It was found that both the magnitude and sign of the anisotropy determine whether the contribution stabilizes or destabilizes the system relative to the case of isotropic surface energy.

Previous theoretical studies of the relationship between the morphological instability of a cylinder that is in contact with a substrate have concentrated on cylinders with isotropic surface energy. McCallum et al. [19] investigated the linear instability of a line of film on a substrate. The unperturbed film state was a cylinder of infinite length with a cross-sectional shape of a segment of a circle. Mass was allowed to flow by diffusion along the film surface. The results of the study found that the substrate presence was a stabilizing influence. Roy and Schwartz [31] studied the stability of liquid ridges on a substrate in the absence of gravity. In this problem the liquid region had a boundary composed of a free surface with a circular arc for a cross-section and a solid cylindrical substrate of arbitrary shape. Their results show that when a particular relationship between the curvatures of the liquid and solid interfaces and the contact angle is satisfied, the infinite liquid ridge is stable with respect to sinuous transverse modes, unlike an infinite cylindrical jet.

For an isotropic surface energy, the base state of the system is relatively simple to describe: the cross-section of the ridge is a circular arc that is determined by the equilibrium contact angles with the substrate. With surface tension anisotropy, the situation is more complicated in several respects. The axis of the nanowire relative to its underlying crystalline axes is a variable to be considered. Once the preferred orientation of the wire axis is established and the wire is assumed to be in contact with the substrate, there remains a geometrical degree of freedom represented by a rotation of the wire about its axis. This rotation exposes different sets of orientations on the crystal-vapor interface, which in turn affects the total energy of the system. Therefore, before the stability to axial perturbations of a nanowire on a substrate is addressed, the selection of the orientation of the wire relative to the substrate must be considered.

During the deposition process, various wire orientations are observed experimentally, depending on the processing conditions and the composition of the deposited crystal and substrate [6, 7, 8, 24, 26, 29]. In particular, for a given set of material parameters, it is argued that the observed orientations depend on kinetics of the (nonequilibrium) deposition process as well as on the effects of surface energy and

surface stress of the crystal-substrate interface [4, 5, 6, 11, 15]. In our simplified model, we consider an inert (nondeforming) substrate with an isotropic crystal-substrate surface energy and ignore kinetic effects by focusing on equilibrium states.

Even with these simplifying assumptions, the identification of the lowest energy orientation of a nanowire on a substrate is more complicated than the simpler problem of determining the lowest energy crystalline orientation of a planar epitaxial layer deposited on a substrate [5, 15, 29], since a nonfacetted nanowire contains a range of crystal-vapor surface orientations instead of the single orientation of a planar film. Within our model, we are able to investigate the low energy orientations of the nanowire on a substrate and perform a stability analysis to axial perturbations (Rayleigh instability).

Our results depend in detail on the anisotropy of the crystal-vapor surface energy. Experimental measurements of this anisotropy for metallic systems are uncommon, although there has been considerable progress recently in atomic-scale simulations of the surface energy anisotropy, specifically for crystal-melt interfaces [1, 14]. Here we adopt a simple one-parameter model for the surface energy anisotropy of the cubic material with a form that is often used to fit the simulation results [1, 14].

In this paper we examine how both the anisotropy of the surface energy of the wire and the interaction of the rod with a substrate affect the stability of the rod. As in the work of Roy and Schwartz [31] on the stability of liquid ridges, we use a variational approach using an energy functional and constant volume condition. Using general anisotropic surface energies, we derive an associated eigenproblem. The eigenproblem is described by a pair of coupled second-order ordinary differential equations with periodic boundary conditions along the axis of the rod and boundary conditions arising from the contact angles between the rod and substrate. We consider the effects of the overall orientation of the crystal relative to the substrate and examine a range of contact angles. The substrate is assumed to be rigid with an isotropic surface energy. We apply the analysis to a number of examples, including the case of a cubic material, and compute the stability of the rod to perturbations when the axis of the rod is aligned parallel to the high symmetry orientations [001], [011], and [111]. When the anisotropy is sufficiently small, the stability of the rod can be computed approximately with asymptotics. For larger levels of anisotropy, the solution is computed numerically.

2. The model. We consider the stability of an infinite rod deposited on a planar substrate below a vapor phase. The rod extends uniformly in the z direction of a Cartesian coordinate system (x, y, z) with the y direction normal to the plane of the substrate. The cross-section of the rod is uniform in z and defined by a two-dimensional equilibrium shape which is determined by surface energy considerations and the angle of contact between crystal and substrate. The vapor-substrate surface energy is denoted by γ_V and the crystal-substrate surface energy is denoted by γ_S. In this model we assume that both γ_V and γ_S are isotropic (constants that are independent of orientation).

Since we will be considering the stability of the rod under an arbitrary shape perturbation, we need to consider the three-dimensional crystal-vapor surface energy for general orientations. The crystal-vapor surface energy will be expressed in terms of the local normal vector to the crystal-vapor interface written in terms of spherical coordinates (ρ, θ, ϕ) in which z is the polar axis, $\rho = \sqrt{x^2 + y^2 + z^2}$ is the radius, θ the polar angle, and ϕ the azimuthal angle as shown in Figure 2.1. The crystal-vapor surface energy is assumed to be anisotropic (orientation-dependent) and is denoted by $\gamma = \gamma(\phi, \theta)$. The unit normal to the unperturbed rod lies in the plane

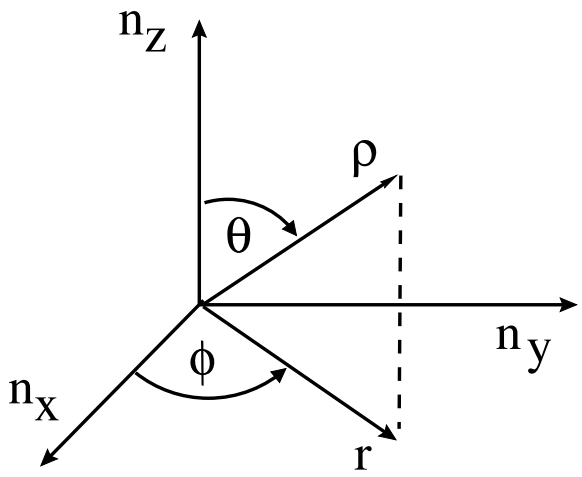


FIG. 2.1. *Schematic diagram of the spherical coordinate system* (ρ, θ, ϕ) *used for the definition of the surface energy* $\gamma(\phi, \theta)$.

$\theta = \pi/2$ and is given by $\hat{\rho}(\phi, \theta = \pi/2) = \hat{r}(\phi) = (\cos\phi, \sin\phi, 0)$, with $\hat{\theta} = -\hat{z}$ and $\hat{\phi}(\phi, \theta = \pi/2) = (-\sin\phi, \cos\phi, 0)$. The axis of the unperturbed rod is parallel to the plane of constant y which represents the substrate. Variables with hats are unit vectors in their corresponding directions.

We restrict our consideration to differentiable surface energies with anisotropies that are mild enough that the surface of the rod is smooth and does not exhibit any missing orientations.

2.1. Three-dimensional surface energy: General formulation. In order to examine the stability of the rod using a variational approach, we will need the general energy functional. This formula and the constant volume condition will be perturbed about the equilibrium rod. The higher order terms in this perturbation expansion will produce a condition for stability. Simply put, for constant volume, if the perturbation increases the energy, the equilibrium state is stable; otherwise it is unstable. This approach parallels the method used by Gurski and McFadden [10], who study the stability of a free rod, but here it is necessary to account for the presence of the substrate.

We will consider the stability of the rod to small amplitude disturbances in the z direction of wavelength $\lambda = 2\pi/k$, where k is the axial wave number. Hence, only the energy and the volume of a portion of the rod of length λ need to be determined. The effect of the substrate on the stability of the rod is local, so only a finite section of the substrate large enough to contain the perturbed rod needs to be examined. In particular, we will consider a rectangular section of the substrate of length λ in the z direction and width $2L_R$ in the x direction.

The total energy of our rod-substrate system, E, can be written as

$$E = E_{CV} + E_{VS} + E_{CS}, \tag{2.1}$$

where E_{CV} is the energy of the crystal-vapor interface, E_{VS} is the energy of the vapor-substrate system, and E_{CS} is the energy of the crystal-substrate system. Letting A_{CS} be the surface area of the crystal-substrate interface, we have that $E_{CS} = \gamma_S A_{CS}$ and $E_{VS} = \gamma_V(2\lambda L_R - A_{CS})$.

The energy of the crystal-vapor interface, E_{CV}, is equal to the surface integral of γ along this interface. If γ is constant, the energy equals γ times the surface area of the crystal-vapor interface. We wish to consider the anisotropic case, so the integral will depend on the orientation of the unit normal to the interface. To compute the associated surface integral, it is helpful to introduce some notation. Let $\vec{X} = \vec{X}(u, v)$ be the position vector of a point along the surface of the rod, where u and v denote surface coordinates. The normal vector field to the rod interface is given by $\vec{P} = \vec{X}_u \times \vec{X}_v$.

Following Gurski and McFadden [10], it is convenient to introduce a generalized surface energy function defined by

$$\Gamma(\vec{P}) = |\vec{P}|\gamma(\Phi, \Theta), \tag{2.2}$$

where

$$\Theta = \tan^{-1}\left(\sqrt{P_x^2 + P_y^2}/P_z\right), \qquad \Phi = \tan^{-1}\left(P_y/P_x\right) \tag{2.3}$$

are the corresponding spherical angles based on the normal vector $\vec{P}$.

A formula for the surface energy can now be obtained by noting that $\gamma(\Phi, \Theta)dA = \Gamma(\vec{P})du\,dv$. The surface energy E_{CV} can therefore be written as

$$E_{CV} = \int\int \Gamma(\vec{P})du\,dv. \tag{2.4}$$

The total energy E is then obtained by substituting (2.4) into (2.1).

It should be noted that the generalized surface energy Γ is closely related to the three-dimensional Cahn–Hoffmann vector [3, 13], $\xi = \nabla\left[\rho\gamma(\phi, \theta)\right]$; in fact,

$$\xi_j(\vec{P}) = \frac{\partial\Gamma(\vec{P})}{\partial P_j}. \tag{2.5}$$

The dimensionless three-dimensional equilibrium shape of a solid particle in a vapor is given by $\vec{\xi}(\phi, \theta)$ for $0 \le \phi \le 2\pi$ and $0 \le \theta \le \pi$, and its normal is $\hat{\rho}(\phi, \theta)$. In the plane $\theta = \pi/2$, this relation reduces to

$$\vec{\xi}(\phi, \pi/2) = \gamma(\phi, \pi/2)\,\hat{r}(\phi) + \gamma_\phi(\phi, \pi/2)\,\hat{\phi}(\phi) - \gamma_\theta(\phi, \pi/2)\,\hat{z}, \tag{2.6}$$

where we have used $\hat{r}(\phi) = (\cos\phi, \sin\phi, 0)$, $\hat{\phi}(\phi) = (-\sin\phi, \cos\phi, 0)$, and $\hat{\theta} = -\hat{z}$ in the plane $\theta = \pi/2$. Note that here partial derivatives are denoted by subscripts, e.g., $\gamma_\phi = \partial\gamma/\partial\phi$. If $\gamma_\theta(\phi, \pi/2) = 0$, then the two-dimensional equilibrium shape defined by (2.6) is characterized by a constant weighted mean curvature $[\gamma + \gamma_{\phi\phi}]\,\mathcal{K}$ [33]. Missing orientations can occur if $\gamma + \gamma_{\phi\phi} < 0$ [35]; here we will assume $\gamma + \gamma_{\phi\phi} > 0$. If $\gamma_\theta(\phi, \pi/2) \neq 0$, then the curve $\vec{\xi}(\phi, \pi/2)$ is out of the plane $z = 0$, but its projection onto the plane represents the two-dimensional equilibrium shape corresponding to $\gamma = \gamma(\phi)$. These two-dimensional shapes define the cross-sections of our rod. We will choose the surface coordinates $(u, v) = (\phi, z)$, which is a natural choice for studying the stability of the equilibrium rod along a substrate to small perturbations.

2.2. The equilibrium rod. The cross-section of the unperturbed rod is a portion of a two-dimensional equilibrium shape parameterized by the vector $\vec{X}^{(0)}(\phi) = (X^{(0)}(\phi), Y^{(0)}(\phi))$, where

$$X^{(0)}(\phi) = \frac{\ell}{\gamma_0}\left[\gamma\left(\phi, \frac{\pi}{2}\right)\cos\phi - \gamma_\phi\left(\phi, \frac{\pi}{2}\right)\sin\phi\right] \tag{2.7}$$

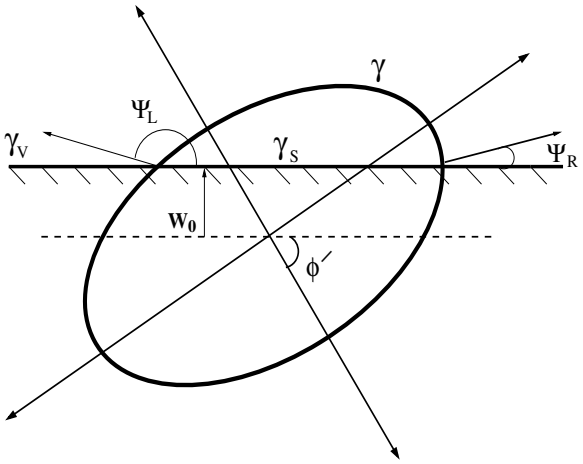


FIG. 2.2. *Ellipse rotated an angle of ϕ' about the original axes. Contact angles between the ellipse and substrate are ψ_R and ψ_L.*

and

$$Y^{(0)}(\phi) = \frac{\ell}{\gamma_0}\left[\gamma\left(\phi, \frac{\pi}{2}\right)\sin\phi + \gamma_\phi\left(\phi, \frac{\pi}{2}\right)\cos\phi\right] \tag{2.8}$$

for $\psi_R \le \phi \le \psi_L$, where $\gamma(\phi,\theta)$ is the surface energy, ℓ is a characteristic length scale, and γ_0 is a characteristic surface energy (see, e.g., [35]). Anticipating the subsequent perturbation expansion, we associate variables having a superscript (0) with the equilibrium rod. The rod is in contact with the substrate over the range $X^{(0)}(\psi_L) \le x \le X^{(0)}(\psi_R)$, and the surface of the substrate in this coordinate system is $y = W_0 = Y^{(0)}(\psi_R) = Y^{(0)}(\psi_L)$. The contact angle that the rod makes with the substrate thus is described by ψ_R and ψ_L, as illustrated in Figure 2.2. Winterbottom [36] (see also [3]) shows that the conditions for equilibrium at the contact line are satisfied for the choice $W_0 = (\ell/\gamma_0)[\gamma_V - \gamma_S]$; we also derive this result below in the course of the energy minimization. Unless otherwise noted, we will henceforth assume that all variables are dimensionless, based on the units of length ℓ and energy γ_0.

3. Stability under perturbations. We determine the stability of the rod by computing the total energy of a volume-preserving perturbation to the rod. The unperturbed interface can be written in the form $\vec{X}^{(0)}(\phi) + z\,\hat{z}$. As in our previous development (see [10]), we then consider a perturbed interface of the form

$$\vec{X}(\phi, z) = \vec{X}^{(0)}(\phi) + z\,\hat{z} + \epsilon h(\phi, z)\hat{r}(\phi) + \frac{\epsilon^2}{2}h_2\hat{r}(\phi) + \cdots, \tag{3.1}$$

where ϵ is a small parameter, $h(\phi, z)$ is the height of the perturbation along the normal $\hat{r}$ to the unperturbed shape, and the constant h_2 is a second-order shape correction introduced to satisfy the volume constraint at $O(\epsilon^2)$. The domain of ϕ is given by $\psi_R(z) \le \phi \le \psi_L(z)$. Note that the contact angles depend on both z and ϵ since the contact angles are determined by the boundary conditions at the contact point. In particular, we will assume that ψ_i has a regular expansion in ϵ of the form

$$\psi_i(z) = \psi_i^{(0)} + \epsilon\psi_i^{(1)}(z) + \frac{\epsilon^2}{2}\psi_i^{(2)}(z) + \cdots \tag{3.2}$$

for $i = L, R$. At the substrate we have

$$Y(\psi_i, z) = W_0, \tag{3.3}$$

so that at leading order we have

$$W_0 = Y^{(0)}(\psi_i^{(0)}) = \gamma(\psi_i^{(0)}, \pi/2)\sin\psi_i^{(0)} + \gamma_\phi(\psi_i^{(0)}, \pi/2)\cos\psi_i^{(0)} \tag{3.4}$$

for $i = L, R$. The case $W_0 = 0$, $\psi_R^{(0)} = 0$, $\psi_L^{(0)} = \pi$ represents a contact line at the orientation of an equatorial plane of symmetry of the equilibrium shape and also corresponds to the upper half of a freely suspended rod [10].

At the contact line the variables X, Y, h, and ψ are related through (3.1)–(3.3). Expanding (3.3) to first order yields

$$Y_\phi^{(0)}(\psi_i^{(0)})\psi_i^{(1)} + Y^{(1)}(\psi_i^{(0)}, z) = 0, \tag{3.5}$$

where from (3.1) the first-order shape change is given by

$$X^{(1)}(\phi, z) = h(\phi, z)\cos\phi, \quad Y^{(1)}(\phi, z) = h(\phi, z)\sin\phi. \tag{3.6}$$

Since $X_\phi^{(0)}(\psi_i^{(0)}) = -(\gamma+\gamma_{\phi\phi})\sin\psi_i^{(0)}$ and $Y_\phi^{(0)}(\psi_i^{(0)}) = (\gamma+\gamma_{\phi\phi})\cos\psi_i^{(0)}$, (3.5) yields

$$\psi_i^{(1)}(z) = \frac{-h(\psi_i^{(0)}, z)\sin\psi_i^{(0)}}{(\gamma + \gamma_{\phi\phi})\cos\psi_i^{(0)}}, \tag{3.7}$$

which relates h and $\psi^{(1)}$ at the contact line.

The geometry of the perturbed rod is determined by the two tangent vectors $\vec{X}_\phi$ and $\vec{X}_z$, and their cross product, $\vec{P} = \vec{X}_\phi \times \vec{X}_z$, which is normal to the interface. The area element on the interface is given by $dA = |\vec{P}|\, d\phi\, dz$. Evaluating the tangent vectors by using (3.1) and taking their cross product, we find that the interface normal has the expansion

$$\vec{P}(\phi, z) = \vec{P}^{(0)}(\phi) + \epsilon\vec{P}^{(1)}(\phi, z) + \frac{\epsilon^2}{2}\vec{P}^{(2)}(\phi, z) + O(\epsilon^3), \tag{3.8}$$

where

$$\vec{P}^{(0)} = (\gamma + \gamma_{\phi\phi})\,\hat{r}, \tag{3.9}$$

$$\vec{P}^{(1)} = h\,\hat{r} - h_\phi\,\hat{\phi} - (\gamma + \gamma_{\phi\phi})\,h_z\,\hat{z}, \tag{3.10}$$

$$\vec{P}^{(2)} = h_2\,\hat{r} - 2hh_z\,\hat{z}. \tag{3.11}$$

3.1. Volume. The shape perturbation (3.1) is required to preserve the volume of the rod over a given length with a period of the perturbation equal to $\lambda = 2\pi/k$. As in Gurski and McFadden [10], we can write the volume as a surface integral by using the divergence theorem. Then using the expansion (3.1), we find that to $O(\epsilon^2)$ the volume is given by

$$\begin{aligned} V &= \frac{1}{2}\int\int\int \nabla\cdot(x, y, 0)\, dV \\ &= \frac{1}{2}\int_0^{2\pi/k}\int_{\psi_R(z)}^{\psi_L(z)} \vec{P}(\phi, z)\cdot[\vec{X}(\phi, z) - z\,\hat{z}]\, d\phi\, dz \end{aligned} \tag{3.12}$$

$$
\begin{aligned}
&-\frac{1}{2}\int_0^{2\pi/k} W_0\left[X(\psi_R(z),z)-X(\psi_L(z),z)\right]\,dz \\
&=\frac{1}{2}\int_0^{2\pi/k}\int_{\psi_R(z)}^{\psi_L(z)}\left\{\left[\gamma+\epsilon h+\frac{\epsilon^2}{2}h_2\right]\left[(\gamma+\gamma_{\phi\phi})+\epsilon h+\frac{\epsilon^2}{2}h_2\right]-\epsilon\gamma_\phi h_\phi\right\}d\phi\,dz \\
&\quad-\frac{1}{2}\int_0^{2\pi/k} W_0\left[X(\psi_R(z),z)-X(\psi_L(z),z)\right]\,dz.
\end{aligned}
$$

Expanding in ϵ then gives

$$V = V^{(0)} + \epsilon V^{(1)} + \frac{\epsilon^2}{2}V^{(2)} + O(\epsilon^3) + \cdots, \tag{3.13}$$

where

$$V^{(0)} = \frac{\lambda}{2}\int_{\psi_R^{(0)}}^{\psi_L^{(0)}} \gamma\left(\gamma+\gamma_{\phi\phi}\right)\,d\phi - \frac{\lambda}{2}W_0[X^{(0)}(\psi_R^{(0)}) - X^{(0)}(\psi_L^{(0)})], \tag{3.14}$$

$$V^{(1)} = \int_0^{2\pi/k}\int_{\psi_R^{(0)}}^{\psi_L^{(0)}} h\left(\gamma+\gamma_{\phi\phi}\right)\,d\phi\,dz, \tag{3.15}$$

$$V^{(2)} = \int_0^{2\pi/k}\int_{\psi_R^{(0)}}^{\psi_L^{(0)}} \left\{h^2 + h_2\left(\gamma+\gamma_{\phi\phi}\right)\right\}d\phi\,dz - \int_0^{2\pi/k} h^2\tan\phi\Big|_{\psi_R^{(0)}}^{\psi_L^{(0)}}\,dz. \tag{3.16}$$

A perturbation $h(\phi,z)$ that is periodic in z with mean zero makes $V^{(1)}=0$, and the condition $V^{(2)}=0$ then determines the appropriate value of the constant h_2. A perturbation $h(\phi)$ that is independent of z does not automatically make $V^{(1)}=0$. This represents a special case for the stability calculation that is treated in the appendix.

3.2. Energy. The stability of the rod is determined by expanding the total energy through $O(\epsilon^2)$ for $|\epsilon| \ll 1$, and, for a given volume, examining whether the shape perturbation, constrained to maintain constant volume of the rod, raises or lowers the energy of the rod. Since the rod is assumed to be infinite in the z direction, an analysis in terms of Fourier components allows us to consider shape perturbations that are periodic in z. The contact angle is now a function of z as well, $\psi_i = \psi_i(z)$ for both $i = L, R$, and the energy E in the region $-L < x < L$ and $0 < z < 2\pi/k$ from (2.1) is given by

$$
\begin{aligned}
E = &\int_0^{2\pi/k} dz \int_{\psi_R(z)}^{\psi_L(z)} \Gamma(\vec{P}(\phi,z))\,d\phi + \frac{4\pi\gamma_V L_R}{k} \\
&-(\gamma_V-\gamma_S)\int_0^{2\pi/k}\left[X(\psi_R(z),z)-X(\psi_L(z),z)\right]dz.
\end{aligned} \tag{3.17}
$$

Expanding in powers of ϵ, we find

$$E = E^{(0)} + \epsilon E^{(1)} + \frac{\epsilon^2}{2}E^{(2)} + O(\epsilon^3), \tag{3.18}$$

where

$$E^{(0)} = \int_0^{2\pi/k} dz \int_{\psi_R^{(0)}}^{\psi_L^{(0)}} \gamma\left(\gamma + \gamma_{\phi\phi}\right)\, d\phi + \frac{4\pi\gamma_V L}{k} + \left(\gamma_V - \gamma_S\right) \int_0^{2\pi/k} \left(\gamma\cos\phi - \gamma_\phi \sin\phi\right) \Big|_{\psi_R^{(0)}}^{\psi_L^{(0)}} dz. \tag{3.19}$$

The first variation of the energy is

$$\frac{E^{(1)}}{2} = \int_0^{2\pi/k} dz \int_{\psi_R^{(0)}}^{\psi_L^{(0)}} \left\{(\gamma + \gamma_{\phi\phi})h + \gamma_\theta\left(\gamma + \gamma_{\phi\phi}\right) h_z\right\} d\phi - \int_0^{2\pi/k} dz \left(\frac{h(\phi, z)}{\cos\phi}\left\{[\gamma_S - \gamma_V] + [\gamma\sin\phi + \gamma_\phi\cos\phi]\right\}\right) \Big|_{\psi_R^{(0)}}^{\psi_L^{(0)}}, \tag{3.20}$$

where we used (3.7) to simplify the second integral.

The second variational term is

$$\begin{aligned}\frac{E^{(2)}}{2} = &\int_0^{2\pi/k} dz \int_{\psi_R^{(0)}}^{\psi_L^{(0)}} \left[h_\phi^2 + (\gamma + \gamma_{\phi\phi})(\gamma + \gamma_{\theta\theta})h_z^2 - 2\gamma_{\phi\theta}h_\phi h_z + \gamma h_2 + 2\gamma_\theta h h_z\right] d\phi \\ &+ \int_0^{2\pi/k} \left\{\left[h_2\gamma_\phi(\psi_L^{(0)}) - 2\gamma_\theta(\psi_L^{(0)})h(\psi_L^{(0)}, z)h_z(\psi_L^{(0)}, z)\tan\psi_L^{(0)}\right]\right. \\ &\left. - \left[h_2\gamma_\phi(\psi_R^{(0)}) - 2\gamma_\theta(\psi_R^{(0)})h(\psi_R^{(0)}, z)h_z(\psi_R^{(0)}, z)\tan\psi_R^{(0)}\right]\right\} dz.\end{aligned} \tag{3.21}$$

3.3. Three-dimensional eigenvalue problem. Our choice of $h(\phi, z)$ makes the integral of the sum of the second and fourth terms in (3.21) identically zero. Eliminating h_2 from (3.21) by using the volume condition, $V^{(2)} = 0$ in (3.16), and integrating by parts, we find that

$$\begin{aligned}\frac{E^{(2)}}{2} = &-\int_0^{2\pi/k} dz \int_{\psi_R^{(0)}}^{\psi_L^{(0)}} h\mathcal{L}h\, d\phi \\ &+ \int_0^{2\pi/k} \left\{\left[h(\psi_L^{(0)}, z)\tan\psi_L^{(0)} + h_\phi(\psi_L^{(0)}, z) - \gamma_{\theta\phi}(\psi_L^{(0)})h_z(\psi_L^{(0)}, z)\right]h(\psi_L^{(0)}, z)\right. \\ &\left. - \left[h(\psi_R^{(0)}, z)\tan\psi_R^{(0)} + h_\phi(\psi_R^{(0)}, z) - \gamma_{\theta\phi}(\psi_R^{(0)})h_z(\psi_R^{(0)}, z)\right]h(\psi_R^{(0)}, z)\right\} dz,\end{aligned} \tag{3.22}$$

where

$$\mathcal{L}h = h_{\phi\phi} + (\gamma + \gamma_{\phi\phi})(\gamma + \gamma_{\theta\theta})h_{zz} - \gamma_{\phi\theta}h_{\phi z} - \left(\gamma_{\theta\phi}h_z\right)_\phi + h. \tag{3.23}$$

Recall that the aim of this calculation is to determine perturbations h such that $E^{(2)} > 0$. This condition can be satisfied if the eigenvalue problem

$$\mathcal{L}h = \mu h \tag{3.24}$$

with boundary conditions

$$h(\psi_i^{(0)}, z)\sin\psi_i^{(0)} + h_\phi(\psi_i^{(0)}, z)\cos\psi_i^{(0)} - \gamma_{\theta\phi}(\psi_i^{(0)})h_z(\psi_i^{(0)}, z)\cos\psi_i^{(0)} = 0 \tag{3.25}$$

for $i = L, R$ has only negative eigenvalues μ. The differential operator $\mathcal{L}$ is identical to that for the isolated rod as given in [10], but instead of periodicity in ϕ we now have boundary conditions that apply at $\psi_R^{(0)}$ and $\psi_L^{(0)}$.

Assuming a discrete set of eigenvalues μ_n and eigenfunctions h_n, for $n = 0, 1, 2, \ldots$, we can rewrite the eigenvalue problem (3.24)–(3.25) as

$$\partial_{\phi\phi}h_n + (\gamma + \gamma_{\phi\phi})(\gamma + \gamma_{\theta\theta})\partial_{zz}h_n - \gamma_{\phi\theta}\partial_{\phi z}h_n - (\gamma_{\theta\phi}\partial_z h_n)_\phi + h_n = \mu_n h_n, \tag{3.26}$$

where the eigenfunctions h_n satisfy the boundary conditions given by (3.25) for $i = L, R$. Note that (3.26) must also satisfy periodicity in the z direction, which can be satisfied by assuming the solution has the form $h_n(\phi, z) = H_n(\phi)\sin kz + G_n(\phi)\cos kz$. Substituting this into (3.26) implies

$$\frac{d^2H_n}{d\phi^2} + (1 - k^2(\gamma + \gamma_{\phi\phi})(\gamma + \gamma_{\theta\theta}))H_n + k\gamma_{\phi\theta}\frac{dG_n}{d\phi} + k\left(\gamma_{\theta\phi}G_n\right)_\phi = \mu_n H_n, \tag{3.27}$$

$$\frac{d^2G_n}{d\phi^2} + (1 - k^2(\gamma + \gamma_{\phi\phi})(\gamma + \gamma_{\theta\theta}))G_n - k\gamma_{\phi\theta}\frac{dH_n}{d\phi} - k\left(\gamma_{\theta\phi}H_n\right)_\phi = \mu_n G_n \tag{3.28}$$

with the boundary conditions

$$H_n \sin\psi_i^{(0)} + \frac{dH_n}{d\phi}\cos\psi_i^{(0)} + k\gamma_{\theta\phi}(\psi_i^{(0)})G_n\cos\psi_i^{(0)} = 0, \tag{3.29}$$

$$G_n \sin\psi_i^{(0)} + \frac{dG_n}{d\phi}\cos\psi_i^{(0)} - k\gamma_{\theta\phi}(\psi_i^{(0)})H_n\cos\psi_i^{(0)} = 0 \tag{3.30}$$

for $i = L, R$. This coupled system of equations must be solved to determine the eigenfunctions and eigenvalues. Note that if γ is independent of θ, the equations decouple.

4. Rotation and contact angles. In the next two sections we consider two related aspects of the stability of a two-dimensional rod on a substrate. We first consider the preferred, low energy orientations of the two-dimensional rod neglecting axial perturbations. For this evaluation, we fix the rod axis that lies parallel to the substrate and compute the energy of the system as the rod is rotated about this axis. Given the specification of the axis of the rod, we assume that preferred orientations correspond to minima of the energy as a function of the rotation angle. Once the low energy orientations are determined, we go on to consider the further effect of axial perturbations on the stability of rods aligned in the preferred orientation.

4.1. Ellipse. Consider the special case of a rod whose cross-section is given by a two-dimensional ellipse,

$$\frac{x^2}{a_x^2} + \frac{y^2}{a_y^2} = 1. \tag{4.1}$$

The major and minor axes of the ellipse are then rotated with respect to the x-axis by an angle ϕ', as shown in Figure 2.2. The corresponding surface free energy γ is given by

$$\gamma(\phi) = \sqrt{a_x^2\cos^2(\phi + \phi') + a_y^2\sin^2(\phi + \phi')}. \tag{4.2}$$

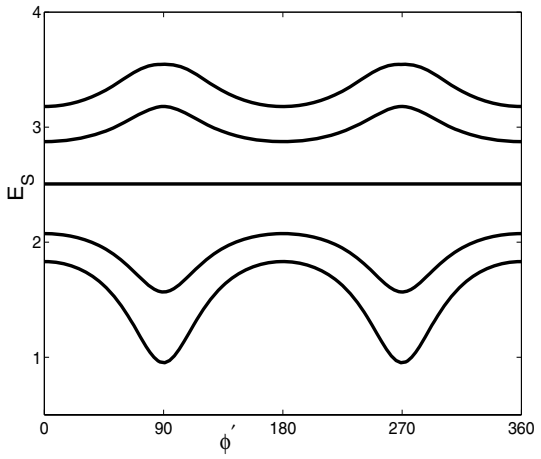


FIG. 4.1. *Scaled energy versus rotation angle for the ellipse. From the top at* $\phi' = 0$ *the curves correspond to* $\tilde{W}_0 = -1.0, -0.5, 0, 0.5, 0.75$.

The surface of the substrate is the plane $y = W_0 = (\gamma_V - \gamma_S)$, and the ellipse makes contact with the substrate with two angles ψ_R and ψ_L, which are roots of the equation

$$0 = W_0 - [\gamma(\phi)\sin\phi + \gamma_\phi(\phi)\cos\phi]. \tag{4.3}$$

If ϕ' is zero or $\pi/2$, then $\psi_L = \pi - \psi_R$.

First, let us consider how the energy depends on the angle of rotation ϕ' and the parameter W_0. Since the volume of a rod with an elliptical cross-section will vary as W_0 varies, we must find a normalization for the energy

$$E = \lambda \int_{\psi_R}^{\psi_L} \gamma\,(\gamma + \gamma_{\phi\phi})\; d\phi - \lambda W_0\,(X(\psi_R) - X(\psi_L)) + 2\lambda L\gamma_V. \tag{4.4}$$

The corresponding formula for the volume of the rod is

$$V = \frac{\lambda}{2} \int_{-\pi/2}^{3\pi/2} \gamma\,(\gamma + \gamma_{\phi\phi})\; d\phi - \lambda \int_{\psi_L}^{\psi_R+2\pi} (W_0 - Y)\; \frac{\partial X(\phi)}{\partial \phi}\, d\phi. \tag{4.5}$$

Using these definitions, we can define a normalized energy E_S as

$$E_S = \frac{E - 2L\gamma_V}{\sqrt{2V\lambda}} = \frac{\int_{\psi_R}^{\psi_L} \gamma\,(\gamma + \gamma_{\phi\phi})\; d\phi - W_0\,(X(\psi_R) - X(\psi_L))}{\sqrt{\int_{-\pi/2}^{3\pi/2} \gamma\,(\gamma + \gamma_{\phi\phi})\; d\phi - 2\int_{\psi_L}^{\psi_R+2\pi} (W_0 - Y)\; \frac{\partial X(\phi)}{\partial \phi}\, d\phi}}. \tag{4.6}$$

For our numerical calculations we set $a_x = 1$, $a_y = 2$. Figure 4.1 shows that when $W_0 > 0$, the lowest scaled energy E_S is attained when $\phi' = 90$ (i.e., the major axis is horizontal) and the highest scaled energy is reached when $\phi' = 0$ degrees (i.e., the major axis is vertical). Since $W_0 = \gamma_V - \gamma_S > 0$, the lowest energy state of the ellipse will be where contact between the crystal and the substrate is maximized and where contact between the substrate and vapor is minimized, i.e., when the semimajor axis is horizontal. The scaled energy is independent of ϕ' when W_0 is zero, showing that when $\gamma_V = \gamma_S$, there is no preferred rod orientation. When W_0 is negative, the scaled energy is higher at $\phi' = 90$ degrees and lower at $\phi' = 0$ degrees as the crystal orients itself to minimize the crystal-substrate interface.

4.2. Cubic materials. A simple model of the surface energy anisotropy for a cubic material is given by the dimensionless expression [20]

$$\gamma(n'_x, n'_y, n'_z) = \left\{1 + 4\epsilon_4([n'_x]^4 + [n'_y]^4 + [n'_z]^4)\right\}, \tag{4.7}$$

where we employ a primed coordinate system that is attached to the crystal axes. We will consider rod directions z that coincide with the high symmetry orientations [001], a fourfold axis; [011], a twofold axis; and [111], a threefold axis. We will use appropriate preliminary rotations of the crystal axes in each case to bring these axes into alignment with the z-axis of the rod, which will be fixed in the unprimed coordinate system.

The shapes are smooth for $-1/18 < \epsilon_4 < 1/12$ (see [20]). For $\epsilon_4 < 0$ the shapes resemble rounded cubes, with [110] edges first forming at $\epsilon_4 = -1/18 \approx -0.0556$. As ϵ_4 decreases below $-1/18$, the edges extend toward the [111] directions, merging to form a corner for $\epsilon_4 = -5/68 \approx -0.07735$. For $\epsilon_4 > 0$ the shapes are octahedral, with [100] corners first forming at $\epsilon_4 = 1/12 \approx 0.0833$.

4.2.1. Rod axis parallel to [001] orientation. If the axis of the rod is aligned with the [001] orientation of the crystal, the dimensionless surface energy resulting from (4.7) is given by [16, 17, 21]

$$\gamma(\phi, \theta) = 1 + \epsilon_4 \left[4\cos^4\theta + \sin^4\theta\,(3 + \cos 4\phi)\right]. \tag{4.8}$$

In the plane $\theta = \pi/2$,

$$\gamma = (1 + 3\epsilon_4) + \epsilon_4 \cos 4\phi. \tag{4.9}$$

If we allow the crystal to rotate on the substrate, we must include the effect of the rotation angle ϕ',

$$\gamma = (1 + 3\epsilon_4) + \epsilon_4 \cos 4\,(\phi + \phi'). \tag{4.10}$$

We can determine the scaled energies using (4.6). Results are shown in Figure 4.2 for several values of ϵ_4 in the range of -0.0556 to 0.0833. The two extreme heights of $W_0 = \pm W_M$, where

$$W_M = \min\left(0.95\sqrt{X(0)^2 + Y(0)^2}, 0.95\sqrt{X(\pi/2)^2 + Y(\pi/2)^2}\right), \tag{4.11}$$

are shown. When W_0 is near $-W_M$, i.e., the origin of the coordinate system is nearly at a maximum height above the substrate surface, the effects of rotation are very slight, with maxima at $\phi' = 0$, 90, 180, 270, and 360 degrees for negative ϵ_4. When ϵ_4 is positive, these maxima switch to minima. The plot of the scaled energy versus angle of rotation for $\epsilon_4 = -0.0556$ has the largest oscillations with an amplitude of 0.001. The fourfold symmetry of the [001] oriented crystal is responsible for the 90 degree spacing. When W_0 is near W_M, the location of the maxima and minima reverse with respect to their locations at $W_0 = -W_M$, and the effect of the rotation becomes more pronounced.

4.2.2. Rod axis parallel to [011] orientation. If the axis of the rod is aligned with the [011] orientation of the crystal, then an appropriate rotation of the crystal axes relative to the rod axis is given by [20]

$$\begin{pmatrix} n'_x \\ n'_y \\ n'_z \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix}. \tag{4.12}$$

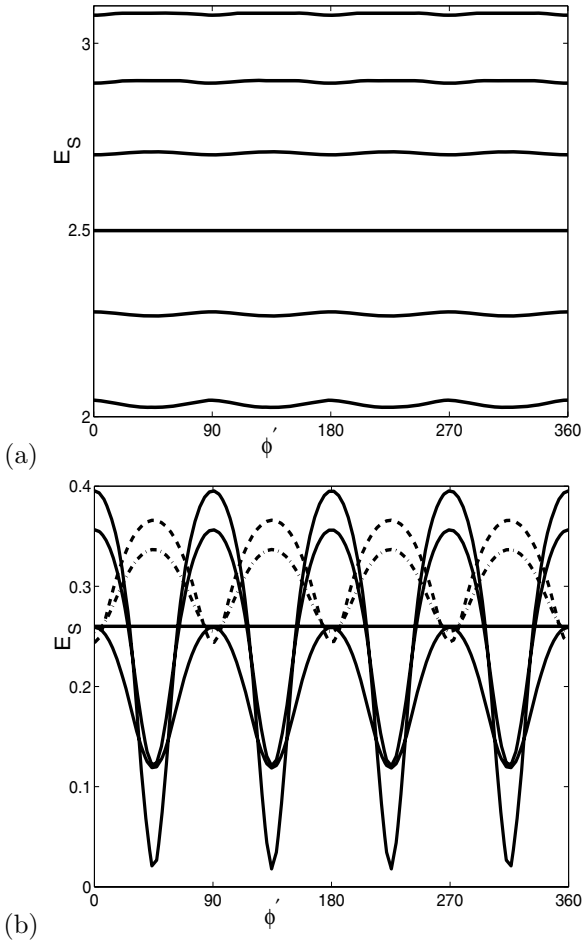


FIG. 4.2. *Scaled energy versus ϕ' for the* [001] *cubic orientation.* (a) *From the bottom, the curves are* $\epsilon_4 = -0.0556,\ -0.02778,\ 0.0,\ 0.02778,\ 0.0556,\ 0.0833$. $W_0 = -W_M$. (b) *From the top, the two dashed curves are* $\epsilon_4 = -0.0556$ *and* -0.02778, *and the solid curves are* $\epsilon_4 = 0.0,\ 0.02778,\ 0.0556,\ 0.0833$. $W_0 = W_M$.

This rotation gives

$$\gamma = 1 + 4\epsilon_4 \left(n_x^4 + \frac{n_y^4}{2} + \frac{n_z^4}{2} + 3n_y^2 n_z^2 \right), \tag{4.13}$$

which reduces to

$$\gamma = 1 + 2\epsilon_4 \left(\cos^4\theta + 6\cos^2\theta \sin^2\theta \sin^2\phi + 2\sin^4\theta \cos^4\phi + \sin^4\theta \sin^4\phi \right). \tag{4.14}$$

When $\theta = \pi/2$ and rotation of the crystal about the substrate is included, then

$$\gamma = 1 + \epsilon_4 \left[\frac{9}{4} + \cos 2(\phi + \phi') + \frac{3}{4} \cos 4(\phi + \phi') \right]. \tag{4.15}$$

The rod is smooth for $-5/68 \le \epsilon_4 \le 1/12$, which is a larger range than the [001] case.

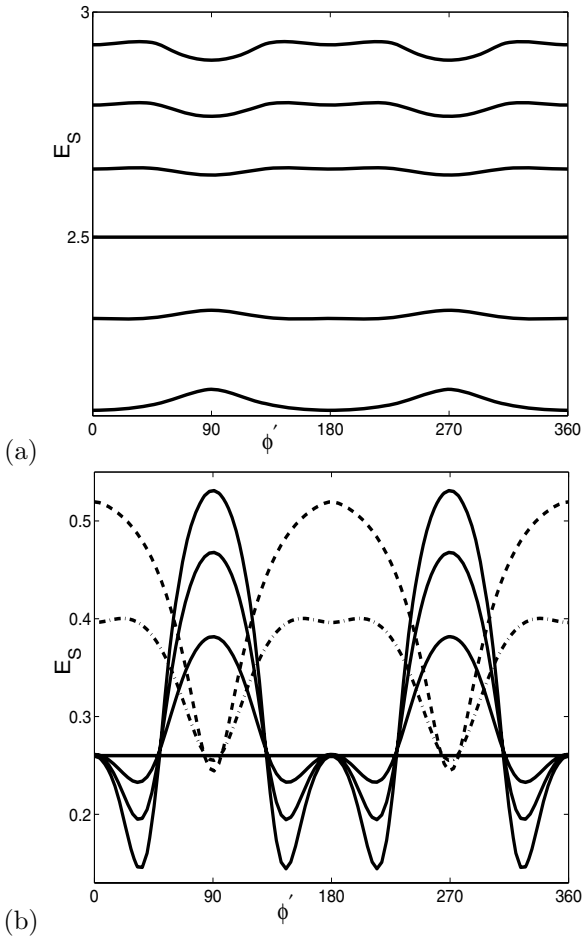


FIG. 4.3. *Scaled energy versus θ' for the* [011] *cubic orientation.* (a) *From the bottom, the curves are* $\epsilon_4 = -0.0556, -0.02778, 0.0, 0.02778, 0.0556, 0.0833$. $\tilde{W}_0 = -W_{M'}$. (b) *At* $\theta' = 0$ *from the bottom, the two dashed curves correspond to* $\epsilon_4 = -0.0556$ *and* -0.02778*, and the solid curves correspond to* $\epsilon_4 = 0.0, 0.02778, 0.0556, 0.0833$. $\tilde{W}_0 = W_M$.

We can determine the scaled energies using (4.6). Results are shown in Figure 4.3 for several values of ϵ_4 in the range of -0.0556 to 0.0833 for the two extreme heights of $W_0 = \pm W_M$. The twofold symmetry of the [011] crystal is apparent in the spacing of maxima and minima shown in Figure 4.3. When ϵ_4 is negative, the minima (maxima) are located at 0, 180, and 360 degrees for $W_0 = -W_M(+W_M)$. When ϵ_4 is positive, the minima are located at 90 and 270 degrees for $W_0 = -W_M$. At $W_0 = W_M$, maxima are maintained at 90 and 270 degrees for positive ϵ_4. We see that secondary local maxima form at $\phi' = 0$, 180, 360 for positive ϵ_4 at $W_0 = W_M$.

4.2.3. Rod axis parallel to [111] orientation. If the axis of the rod is aligned with the [111] orientation of the crystal, then an appropriate rotation of the crystal axes relative to the rod axis is given by [20]

$$\begin{pmatrix} n'_x \\ n'_y \\ n'_z \end{pmatrix} = \begin{pmatrix} \sqrt{2}/\sqrt{3} & 0 & 1/\sqrt{3} \\ -1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \\ -1/\sqrt{6} & -1/\sqrt{2} & 1/\sqrt{3} \end{pmatrix} \begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix}. \tag{4.16}$$

This leads to the form

$$\gamma = 1 + 4\epsilon_4\left(\frac{n_x^4}{2} + \frac{n_y^4}{2} + \frac{n_z^4}{3} + n_x^2 n_y^2 + 2n_x^2 n_z^2 + 2n_y^2 n_z^2 + \frac{2\sqrt{2}}{3} n_x^3 n_z - 2\sqrt{2} n_x n_y^2 n_z\right), \tag{4.17}$$

which reduces to

$$\gamma = 1 + 4\epsilon_4\left(\frac{1}{3}\cos^4\theta + 2\cos^2\theta\sin^2\theta + \frac{1}{2}\sin^4\theta + \frac{2\sqrt{2}}{3}\cos\theta\sin^3\theta\cos 3\phi\right). \tag{4.18}$$

Although the [111] orientation is not isotropic, in the $\theta = \pi/2$ plane,

$$\gamma = 1 + 2\epsilon_4. \tag{4.19}$$

Therefore, the effect of the ϕ dependence is lost and the surface energy for the [111] orientation is unchanged by a rotation of ϕ' about the substrate.

5. Linear stability calculations. Next we investigate the linear stability of the system by examining the eigenvalue problem associated with diagonalizing the second variation of the energy for a fixed orientation of the rod on the substrate. We consider a number of examples, including an ellipsoidal surface energy anisotropy and several variants of cubic anisotropy. The three-dimensional study includes both numerical and asymptotic results. A discussion on the effect of rotation on stability is covered in the appendix.

5.1. Ellipsoidal anisotropy. We first discuss an anisotropic surface energy that leads to an ellipsoidal equilibrium shape described by

$$\frac{x^2}{a_x^2} + \frac{y^2}{a_y^2} + \frac{z^2}{a_z^2} = 1. \tag{5.1}$$

We consider an axisymmetric shape with $a_x = a_y = 1$. The corresponding surface free energy is given by $\gamma(\phi,\theta) = \sqrt{\sin^2\theta + a_z^2\cos^2\theta}$, and in the plane $\theta = \pi/2$ we have $\gamma = 1$, $\gamma_{\phi\phi} = 0$, $\gamma_{\theta\phi} = 0$, and $\gamma_{\theta\theta} = a_z^2 - 1$. The eigenvalue problem (3.27)–(3.28) decouples, leading to the single equation

$$\frac{\partial^2 H_n}{\partial\phi^2} + (1 - K^2)H_n = \mu_n H_n, \tag{5.2}$$

where $K = a_z k$. The boundary conditions for $i = L, R$ are

$$H_n(\psi_i^{(0)})\sin\psi_i^{(0)} + H_{n\phi}(\psi_i^{(0)})\cos\psi_i^{(0)} = 0. \tag{5.3}$$

We solve this problem numerically using a pseudospectral Chebyshev method for a range of contact angles determined by solutions to (4.3), where $-1 \le W_0 \le 1$. To eliminate the change in length scale with the contact angle, we define $\kappa = KR_e$ as the dimensionless axial wave number based on the effective radius of the cross-section.

In Figure 5.1(a) we show the most unstable mode μ_0 for a range of contact angles ψ_R as a function of κ. The results indicate that the ellipsoid is stabilized with respect to long wavelengths for large ψ_R. In Figure 5.1(b) we plot the square of the critical wavenumber κ_C, which corresponds to the transition between stable and unstable behavior, as a function of the contact angle ψ_R.

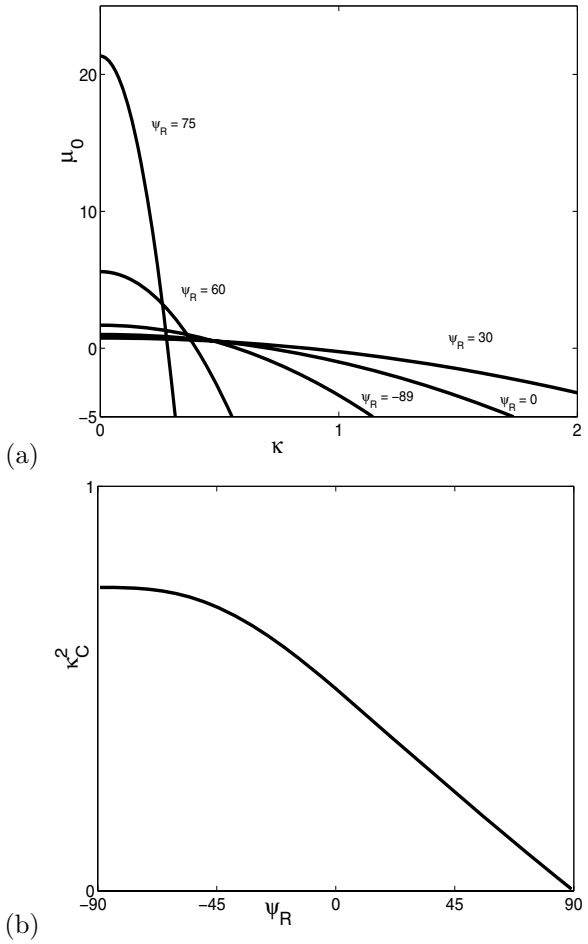


FIG. 5.1. (a) *Eigenvalues for $n = 0$ versus $\kappa = R_e K$ for the ellipsoid case. From the lower curve at $\kappa = 0$: $\psi_R^{(0)} = -89, 0, 30, 60, 75$ degrees.* (b) *The square of the rescaled wave number, κ_C^2, versus ψ_R for the ellipsoid.*

5.2. Cubic materials. For numerical determination of the eigenvalues we used a pseudospectral Chebyshev discretization of the (3.27), (3.28) with boundary conditions from (3.25) for contact angles, where $-\pi \leq \psi_R \leq \pi$. For this problem we may choose a value ψ_R, thereby fixing the value of W_0. If we restrict ψ_R between $-\pi/2$ and $\pi/2$, then ψ_L is the solution to

$$W_0 - (\gamma(\phi)\sin\phi + \gamma_\phi(\phi)\cos\phi) = 0 \tag{5.4}$$

for a value of ϕ between $\pi/2$ and $3\pi/2$. Therefore, fixing the value of ψ_R determines the value of ψ_L.

Since γ varies with respect to ϵ_4 for each of the three high symmetry cubic orientations [001], [011], and [111], we must determine the effective radius of the cross-section R_e. To eliminate the change in length scale with ϵ_4, we set $\kappa = kR_e$, which is the dimensionless axial wave number based on the effective radius of the cross-section.

In addition, we performed an asymptotic expansion of the problem with a plane of symmetry at $\phi = \pi/2$, i.e., when $\psi_L = \pi - \psi_R$. The results for the numerical cal-

culation will be compared to the results for the asymptotic expansion in the following sections.

5.2.1. Rod axis parallel to [001] orientation. The dimensionless surface energy is given by (4.8) for the [001] orientation. In the plane $\theta = \pi/2$, we then have $\gamma_{\theta\phi} = 0$, and

$$\gamma = (1 + 3\epsilon_4) + \epsilon_4 \cos 4\phi, \tag{5.5}$$

$$(\gamma + \gamma_{\phi\phi})(\gamma + \gamma_{\theta\theta}) = \left(1 - 6\epsilon_4 - \frac{9}{2}\epsilon_4^2\right) - (18\epsilon_4 - 126\epsilon_4^2)\cos 4\phi + \frac{45}{2}\epsilon_4^2 \cos 8\phi. \tag{5.6}$$

The rod is smooth for $-1/18 \le \epsilon_4 \le 1/12$.

For this orientation, $\gamma_{\theta\phi}(\phi, \pi/2)$ vanishes, leading to a decoupling of (3.27), (3.28) that leaves a single equation,

$$\frac{\partial H_n}{\partial_{\phi\phi}} + \left(1 - k^2\left[1 + \epsilon_4 A_1(\phi) + \epsilon_4^2 A_2(\phi)\right]\right) H_n = \mu_n H_n, \tag{5.7}$$

where

$$A_1 = -6[1 + 3\cos 4\phi], \quad A_2 = -\frac{9}{2}[1 - 28\cos 4\phi - 5\cos 8\phi]. \tag{5.8}$$

For the asymptotics, we assume a symmetry condition about $\phi = \pi/2$. If $-\pi/2 \le \psi_R \le \pi/2$, then $\psi_L = \pi - \psi_R$. Then the boundary conditions become

$$H_n(\psi_R)\sin\psi_R + H_{n\phi}(\psi_R)\cos\psi_R = 0, \tag{5.9}$$

$$H_{n\phi}(\pi/2) = 0. \tag{5.10}$$

We take the simple expansions of H_n and μ_n in terms of small ϵ_4:

$$H_n(\phi) = H_n^{(0)}(\phi) + \epsilon_4 H_n^{(1)}(\phi) + O\left(\epsilon_4^2\right), \tag{5.11}$$

$$\mu_n = \mu_n^{(0)} + \epsilon_4 \mu_n^{(1)} + O(\epsilon_4^2). \tag{5.12}$$

The formal asymptotic expansion gives

$$H_n(\phi) = C_1 \cos\beta_n(\pi/2 - \phi) + O(\epsilon_4), \tag{5.13}$$

$$\mu_n = 1 - k^2 - \beta_n^2 + \epsilon_4\left\{6k^2 - \frac{9k^2\beta_n}{(\eta_n + \sin\eta_n)}\left[\sin 4\psi_R^{(0)} + \frac{\sin(4\psi_R^{(0)} - \eta_n)}{(\beta_n + 2)} - \frac{\sin(4\psi_R^{(0)} + \eta_n)}{(\beta_n - 2)}\right]\right\} + O(\epsilon_4^2), \tag{5.14}$$

where $\eta_n = 2\beta_n(\pi/2 - \psi_R^{(0)})$ and the value of β_n must satisfy

$$\cos\beta_n(\pi/2 - \psi_R)\sin\psi_R + \beta_n \sin\beta_n(\pi/2 - \psi_R)\cos(\psi_R) = 0.$$

In Figure 5.2, the first two terms of the asymptotic expansion for the most unstable mode, μ_0, are compared against the numerical value for several choices of contact angle, $\psi_R^{(0)}$, for $\epsilon_4 = 0.20$. The asymptotic results, shown by the dashed curves, are

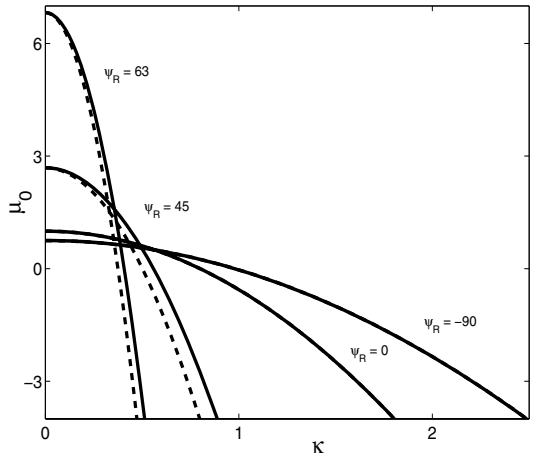


FIG. 5.2. *Eigenvalues for $n = 0$ versus the rescaled wave number $\kappa = R_e k$ for the* [001] *orientation. Reading from the solid lower curve to the solid upper curve at $\kappa = 0$, the corresponding values of ψ_R are $-90, 0, 45, 63$ degrees as indicated. The solid curve represents the numerical solution, and the dashed curve represents the asymptotic solution for $\epsilon_4 = 0.020$.*

close to the numerical results represented by the solid curves. These results show that the larger contact angles are more stable with respect to large dimensionless axial wave numbers. In addition, the [001] orientation is more stable with respect to the isotropic case for negative ϵ_4 and destabilized for positive ϵ_4. The $\psi_R = 0$ case corresponds to the free rod discussed in [10].

The corresponding critical wave number κ_C, defined where μ_0 is zero, as a function of ϵ_4 has the form $\kappa_C = R_e k$. If we define $\sigma_n^{(1)} = k^2\tau$, then the critical dimensionless wave number is as follows:

$$\kappa_C^2 = R_e^2(1 - \beta_n^2)(1 + \epsilon_4\tau) + O(\epsilon_4^2). \tag{5.15}$$

Figure 5.3 shows the results for the square of the rescaled critical wave number versus the contact angle ψ_R. The isotropic case, where ϵ_4 is zero, matches the results found in McCallum et al. [19] for the isotropic rod in contact with a substrate. It is clear from Figure 5.3 that the negative values of ϵ_4 stabilize the rod for all contact angles with respect to the isotropic rod, while the positive values of ϵ_4 destabilize the rod. This behavior also was observed for the freestanding rod [10].

5.2.2. Rod axis parallel to [011] orientation. The dimensionless surface energy, γ, for the [011] orientation is given by (4.15). In the plane $\theta = \pi/2$, we then have $\gamma_{\theta\phi} = 0$, and

$$\gamma = \left(1 + \frac{9}{4}\epsilon_4\right) + \epsilon_4 \cos 2\phi + \frac{3}{4}\epsilon_4 \cos 4\phi, \tag{5.16}$$

(5.17)

$$\begin{aligned}(\gamma + \gamma_{\phi\phi})\,(\gamma + \gamma_{\theta\theta}) = \Big(&1 + \frac{3}{2}\epsilon_4[5 - 12\cos 2\phi - 9\cos 4\phi] \\ &+ \frac{9}{32}\epsilon_4^2[167 + 136\cos 2\phi - 148\cos 4\phi + 312\cos 6\phi + 45\cos 8\phi]\Big).\end{aligned}$$

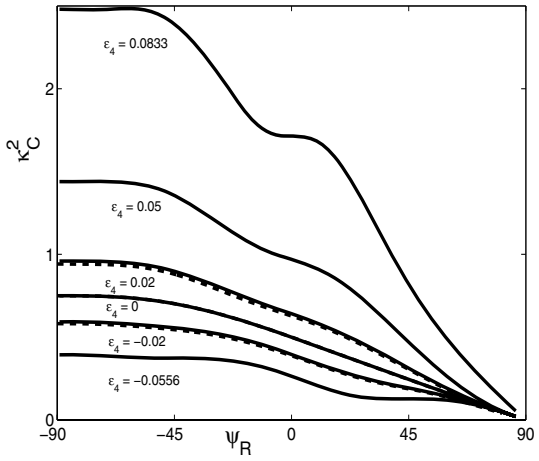


FIG. 5.3. *The square of the critical rescaled wavenumber, κ_C^2, versus ψ_R for the* [001] *orientation for* $\epsilon_4 = -0.0556, -0.02, 0, 0.02, 0.05, 0.0833$. *The solid curves represent the numerical solutions and the dashed curves the asymptotic solutions. The asymptotic solutions are given only for* $\epsilon_4 = -0.02, 0.0.02$.

The asymptotic expansion for the [011] orientation is similar to the [001] orientation. Therefore, we merely state the results for the first two terms of the eigenvalue expansion:

$$\mu_n^{(0)} = 1 - k^2 - \beta_n^2, \tag{5.18}$$

$$\begin{aligned}\mu_n^{(1)} = -\frac{15}{2}k^2 - \frac{9k^2\beta_n}{(\eta_n + \sin\eta_n)}\Bigg[&2\sin 2\psi_R^{(0)} + \frac{3}{4}\sin 4\psi_R^{(0)} \\ &-\left(\frac{2\sin 2\psi_R^{(0)}}{\beta_n^2 - 1} + \frac{3\sin 4\psi_R^{(0)}}{\beta_n^2 - 4}\right)\cos\eta_n \\ &-\left(\frac{2\beta_n\cos 2\psi_R^{(0)}}{\beta_n^2 - 1} + \frac{3\beta_n\cos 4\psi_R^{(0)}}{2(\beta_n^2 - 4)}\right)\sin\eta_n\Bigg],\end{aligned} \tag{5.19}$$

where η_n and β_n are as defined for the [001] orientation. Figure 5.4(a) shows numerical results for the range $-0.05 \le \epsilon_4 \le 0.08$ and the asymptotic results for $\epsilon_4 = -0.02, 0, 0.02$. The extreme ends of the smooth ϵ_4 range present some numerical difficulties and are therefore not shown. Likewise the extreme case, where ψ_R is approaching 90 degrees and the cross-sectional area of the rod is approaching zero, prevents us from calculating over the entire range of ψ_R. The curve $\epsilon_4 = 0$ corresponds to the isotropic case. The substrate acts as a stabilizing influence on the rod; even the negative ϵ_4 case, while less stable than the positive ϵ_4 case, is less unstable than it is for a freestanding rod. The large wave number instability associated with negative values of $(\gamma + \gamma_{\theta\theta})$ seen in the freestanding rod [10] when ϵ_4 is in the range $-5/68 \le \epsilon_4 \le -1/18$ is possibly related to the growing κ_C^2 values observed for $\epsilon_4 = -0.05$ over a positive range of ψ_R.

5.2.3. Rod axis parallel to [111] orientation. If the axis of the rod is aligned with the [111] orientation of the crystal axes relative to the rod axis, then γ is given by the following in the plane $\theta = \pi/2$:

$$\gamma = 1 + 2\epsilon_4, \tag{5.20}$$

$$(\gamma + \gamma_{\phi\phi})(\gamma + \gamma_{\theta\theta}) = 1 + 12\epsilon_4 + 20\epsilon_4^2. \tag{5.21}$$

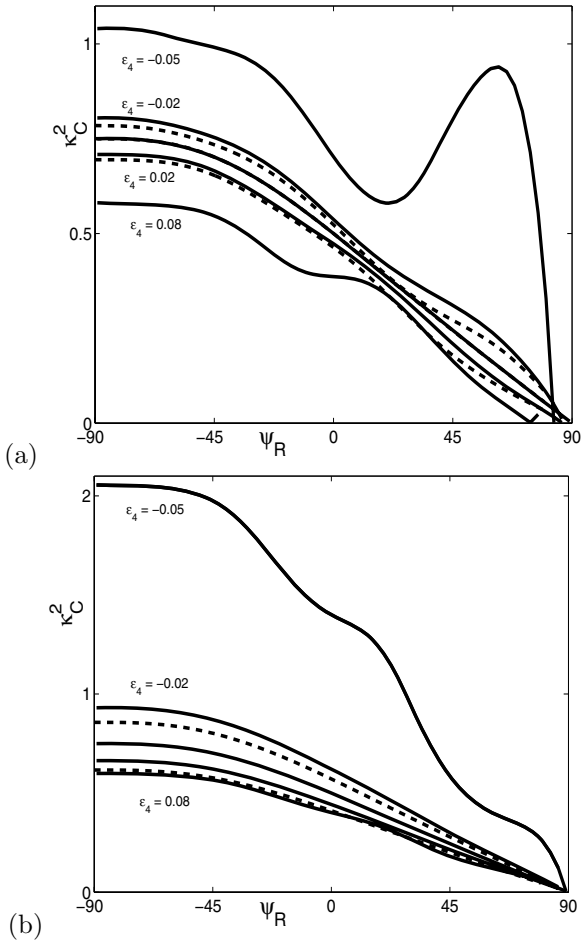


FIG. 5.4. *The square of the critical rescaled wave number, κ_C^2, versus ψ_R for $\epsilon_4 = -0.05, -0.02, 0, 0.02, 0.08$ for* (a) *the* [011] *orientation and* (b) *the* [111] *orientation. The solid curves represent the numerical solutions, and the dashed curves represent the asymptotic solutions. The asymptotic solutions are given only for $\epsilon_4 = -0.02, 0.0.02$.*

The asymptotic expansion for this orientation differs from the expansions for the [001] and [011] orientations since in the [111] orientation (3.27) and (3.28) are coupled through the nonzero term

$$\gamma_{\theta\phi} = 8\sqrt{2}\epsilon_4 \sin 3\phi. \tag{5.22}$$

The zeroth-order conditions and results are the same as those for the [001] and [011] orientations, as expected, so only the details for the first-order terms are shown here. We begin with the equations that must be solved to determine $\mu_n^{(1)}$:

$$\begin{aligned} &- H_{n\phi}^{(1)}(\psi_0)u(\psi_0) + H_n^{(1)}(\psi_0)u_\phi(\psi_0) \\ &- \int_{\psi_R^{(0)}}^{\pi/2} [(\mu_n^{(1)} + 12k^2)H_n^{(0)} - k\gamma_{\theta\phi}(\phi)G_{n\phi}^{(0)} - k(\gamma_{\theta\phi}(\phi)G_n^{(0)})_\phi]u\, d\phi = 0 \end{aligned} \tag{5.23}$$

and

$$-G^{(1)}_{n\phi}(\psi^{(0)}_R)u(\psi^{(0)}_R)+G^{(1)}_{n}(\psi^{(0)}_R)u_\phi(\psi^{(0)}_R) - \int_{\psi^{(0)}_R}^{\pi/2}[(\mu^{(1)}_n+12k^2)G^{(0)}_n+k\gamma_{\theta\phi}(\phi)H^{(0)}_{n\phi}+k(\gamma_{\theta\phi}(\phi)H^{(0)}_n)_\phi]u\,d\phi=0 \tag{5.24}$$

with the boundary conditions

$$H^{(1)}_{n\phi}(\pi/2)=0, \tag{5.25}$$

$$G^{(1)}_{n\phi}(\pi/2)=0, \tag{5.26}$$

$$H^{(1)}_n(\psi^{(0)}_R)\sin\psi^{(0)}_R+H^{(1)}_{n\phi}(\psi^{(0)}_R)\cos\psi^{(0)}_R+k\gamma_{\theta\phi}(\psi^{(0)}_R)G^{(0)}_n\cos\psi^{(0)}_R=0, \tag{5.27}$$

$$G^{(1)}_n(\psi^{(0)}_R)\sin\psi^{(0)}_R+G^{(1)}_{n\phi}(\psi^{(0)}_R)\cos\psi^{(0)}_R-k\gamma_{\theta\phi}(\psi^{(0)}_R)H^{(0)}_n\cos\psi^{(0)}_R=0. \tag{5.28}$$

Concentrating on (5.23), one sees that the equation can be rewritten as

$$-\big(H^{(1)}_{n\phi}(\psi^{(0)}_R)+k\gamma_{\theta\phi}(\psi^{(0)}_R)G^{(0)}_n(\psi^{(0)}_R)\big)u(\psi^{(0)}_R)+H^{(1)}_n(\psi^{(0)}_R)u_\phi(\psi^{(0)}_R) -(\mu^{(1)}_n+12k^2)\int_{\psi^{(0)}_R}^{\pi/2}H^{(0)}_n u\,d\phi+k\int_{\psi^{(0)}_R}^{\pi/2}[\gamma_{\theta\phi}(\phi)G^{(0)}_{n\phi}u(\phi)-\gamma_{\theta\phi}(\phi)G^{(0)}_n u_\phi(\phi)]\,d\phi=0. \tag{5.29}$$

The boundary terms vanish since the operator is self-adjoint. In addition, since $G^{(0)}_n(\phi)$ and $u(\phi)$ differ only by a constant, the second integral vanishes as well. Thus one finds that

$$\mu^{(1)}_n=-12k^2. \tag{5.30}$$

Applying similar logic to (5.24), one finds the same result.

The numerical results for the [111] orientation for $-0.05\le\epsilon_4\le 0.08$ and the asymptotic results for $\epsilon_4=-0.02,0,0.02$ are shown in Figure 5.4(b). The extreme ends of the smooth ϵ_4 range present some numerical difficulties and are therefore not shown. Figure 5.4(b) shows that for nonnegative ϵ_4, the substrate is a stabilizing presence, but even this added stability is unable to overcome the instability associated with positive values of ϵ_4.

6. Conclusion. We have examined rotation effects and the linear stability of a rod on a substrate, in which the rod has a uniform cross-section given by a two-dimensional equilibrium shape. This work extends our previous treatment of a free-standing rod, where the stability analysis produces an associated eigenvalue problem with periodic boundary conditions. The effect of the rod making contact with the substrate involves instead mixed boundary conditions for the eigenvalue problem. The eigenvalues are determined numerically with asymptotic solutions given for the limiting case of small anisotropy. The eigenproblem is a coupled pair of second-order ordinary differential equations with coefficients that are periodic along the axis of the rod and depend on the second derivatives with respect to the orientation variables. We assumed a weak anisotropic surface energy to eliminate missing orientations on the rod.

As was found in our previous exploration of the freestanding anisotropic rod, the magnitude and the sign of the anisotropy determine the relative stability in comparison to the isotropic case. The overall effect of the substrate is stabilizing to the anisotropic rod. In general, as the contact angle ψ_R tends to 90 degrees, the rod on the substrate becomes more stable, which is analogous to the stability of a

three-dimensional planar film, where the anisotropy is not strong enough to make the problem ill-posed. In particular, the rod on a substrate with any of the high symmetry cubic orientations maintains the same relationship as the freestanding rods as to whether or not a positive or negative anisotropy enhances or diminishes stability. When the contact angle ψ_R between the rod and the substrate approaches -90 degrees, the stability does not revert to that of the freestanding rod. In this limiting case, the rods are pinned to the surface at one point, and this single point of contact increases the stability of the rod with respect to the unpinned case. This effect was also seen by McCallum et al. [19] for the isotropic rod.

We have considered the effect of rotation on the stability of a two-dimensional rod whose cross-section is either elliptical or a shape determined by one of the three high symmetry cubic orientations. In order to describe these cross-sections, the coordinate system is fixed to the center of the rod, thereby defining a substrate height above this center. When the major axis of the ellipse is horizontal, the ellipse is most stable with a positive substrate height (see Figure 2.2) and is least stable with a negative substrate height. For this elliptical case, the observation is consistent with the remark above that contact angles near 90 degrees are more stable than contact angles near -90 degrees. But the observation concerning substrate height applies to more general cases. In particular, the stability of rods whose cross-sections are determined by high symmetry cubic orientations mimic this reversal of stability when the substrate height is at a maximum over the coordinate axes in comparison to a minimum. Similar to the linear stability results, the stability under rotation of these rods with the [001] and [011] cubic orientations depends greatly on the sign and magnitude of the anisotropy. Negative anisotropy corresponds to a rod with a cross-section of a rounded cube; positive anisotropy corresponds to a rod with a smoothed octahedral cross-section. In general we observe that for the [001] cubic orientation, the more negative the anisotropy, the more unstable the rod. The situation is reversed for the [011] and [111] cubic orientations.

The two- or fourfold symmetry of the orientations is reflected in the effect of rotation of the rods on the stability. The stability of the rod with the [111] cubic orientation is found to be unchanged by rotation.

These results may be potentially useful in the manufacture of stable long rods or wires with an axis oriented along the high symmetry orientations [001], [011], and [111].

Appendix. The effect of rotations on perturbation stability. In this section we discuss the effect of rotating a crystal on the second variational energy of the contact line problem. We start with a study of the two-dimensional shapes from sections 4.1 and 4.2: an ellipse and three two-dimensional cubic crystals in high symmetry orientations. For the ellipse, the major and minor axes are rotated with respect to the x-axis at an angle of ϕ'. Additionally, the substrate may be moved upwards from the original x-axis at a distance of W_0. In this particular case the effect of W_0 will be seen indirectly, as it does not appear explicitly in the relevant equations. However, W_0 determines the contact angles ψ_R and ψ_L that the crystal makes with the substrate.

In two dimensions the stability problem (i.e., the second variation of the energy) reduces to the following:

$$\frac{E^{(2)}}{2} = -\int_{\psi_R^{(0)}}^{\psi_L^{(0)}} [h_{\phi\phi}(\phi) + h(\phi)]h(\phi)\, d\phi \tag{A.1}$$

with the boundary conditions

$$h(\psi_i^{(0)})\sin\psi_i^{(0)} + h_\phi(\psi_i^{(0)})\cos\psi_i^{(0)} = 0 \tag{A.2}$$

for $i = L, R$, subject to the constraints where

$$\int_{\psi_R^{(0)}}^{\psi_L^{(0)}} h(\phi)\left[\gamma(\phi+\phi') + \gamma_{\phi\phi}(\phi+\phi')\right] d\phi = 0, \tag{A.3}$$

and where

$$\int_{\psi_R^{(0)}}^{\psi_L^{(0)}} h^2(\phi)\, d\phi \tag{A.4}$$

is minimized.

Therefore we can formulate the problem as

$$h_{\phi\phi}(\phi) + h(\phi) = \mu\, h(\phi) + \tau\, \left[\gamma(\phi+\phi') + \gamma_{\phi\phi}(\phi+\phi')\right] \tag{A.5}$$

with the boundary conditions given as above in (A.2). We solve this problem using two different numerical techniques. The first uses a pseudospectral Chebyshev calculation [34] to determine a basis of eigenvectors for the problem

$$h_{\phi\phi}(\phi) + h(\phi) = \mu\, h(\phi), \tag{A.6}$$

subject to the boundary conditions, that are orthogonal to the vector

$$\gamma(\phi+\phi') + \gamma_{\phi\phi}(\phi+\phi'). \tag{A.7}$$

In addition, we can solve (A.2), (A.5) with a double shooting method.

The unstable eigenmodes are those where $\mu_n > 0$. We find that the largest eigenvalue, μ_0, is zero to numerical accuracy for all cases, indicating that none of the modes are unstable. Analytically we can obtain some insight by noting that (A.6) admits solutions of the form

$$-n^2 + 1 = \mu_n, \tag{A.8}$$

$$h_n(\phi) = A_n \sin(n\phi) + B_n \cos(n\phi). \tag{A.9}$$

Note that μ_n is negative for $n \geq 2$, resulting in stable shape perturbations. The boundary conditions show that if $n = 0$, then $h_0 = 0$. The next allowed n is $n = 1$, which gives $\mu_1 = 0$ and $h_1 = B_1 \cos\phi$, an allowable solution. This is an eigenmode as long as it satisfies the orthogonality constraint given in (A.3).

Acknowledgments. The authors are grateful for helpful discussions with D. L. Cotrell.

REFERENCES

[1] M. Asta, J. J. Hoyt, and A. Karma, *Calculation of alloy solid-liquid interfacial free energies from atomic-scale simulations*, Phys. Rev. B, 66 (2002), 100101.

[2] J. W. Cahn, *Stability of rods with anisotropic surface free energy*, Scripta Metall., 13 (1979), pp. 1069–1071.

[3] J. W. Cahn and D. W. Hoffmann, *A vector thermodynamics for anisotropic interfaces*, II. *Curved and facetted surfaces*, Acta Metall., 22 (1974), pp. 1205–1214.
[4] S.-C. Chang, J.-M. Shieh. B-T. Dai, M.-S. Feng, and Y.-H. Li, *The effect of plating current densities on self-annealing behaviors of electroplated copper films*, J. Electrochem. Soc., 149 (2002), G535.
[5] J. S. Chen, B. C. Lim, and J. P. Wang, *Controlling the crystallographic orientation and the axis of magnetic anisotropy in LI(o) FePT films*, Appl. Phys. Lett., 81 (2002), pp. 1848–1850.
[6] J. Diao, K. Gall, and M. L. Dunn, *Surface stress driven reorientation of gold nanowires*, Phys. Rev. B, 70 (2004), 075413.
[7] T. Gao, G. W. Meng, Y. Wang, S. Sun, and J. Zhang, *Electrochemical synthesis of copper nanowires*, J. Phys. Condens. Matter, 14 (2002), pp. 355–363.
[8] T. Gao, G. W. Meng, J. Zhang, Y. W. Wang, C. H. Liang, J. C. Fan, and L. D. Zhang, *Template synthesis of single-crystal Cu nanowire arrays by electrodeposition*, Appl. Phys. A, 73 (2001), pp. 251–254.
[9] A. M. Glaeser, *A new approach to investigating surface transport in ceramics,* in Mass and Charge Transport in Ceramics, K. Koumoto, H. Matsubara, and L. M. Sheppard, eds., Ceramic Trans. 71, American Ceramic Society, Westerville, OH 1996, pp. 117–136.
[10] K. F. Gurski and G. B. McFadden, *The effect of anisotropic surface energy on the Raleigh instability*, R. Soc. Cond. Proc. Ser. A Math. Phys. Eng. Sci., 459 (2003), pp. 2575–2598.
[11] J. M. E. Harper, C. Cabral, Jr., P. C. Andricacos, L. Gignac, I. C. Noyan, K. P. Rodbell, and C. K. Hu, *Mechanisms for microstructure evolution in electroplated copper thin films near room temperature*, J. Appl. Phys., 86 (1999), pp. 2516–2525.
[12] C. Herring, *The use of classical macroscopic concepts in surface-energy problems*, in Structure and Properties of Solid Surfaces, R. Gomer and C. S. Smith, eds., University of Chicago Press, Chicago, 1953, pp. 5–72.
[13] D. W. Hoffmann and J. W. Cahn, *A vector thermodynamics for anisotropic interfaces*. I. *Fundamentals and applications to plane surface junctions*, Surf. Sci., 31 (1972), pp. 368–388.
[14] J. J. Hoyt, M. Asta, and A. Karma, *Method for computing the anisotropy of the solid-liquid interfacial free energy*, Phys. Rev. Lett., 86 (2001), pp. 5530–5533.
[15] H. H. Huang, M. H. Hon, and M. C., Wang, *Effect of NH3 on the growth characterization of TiN films at low temperature*, J. Crystal Growth, 240 (2002), pp. 513–520.
[16] D. A. Kessler and H. Levine, *Growth velocity of three-dimensional dendritic crystals*, Phys. Rev. A, 36 (1987), pp. 4123–4126.
[17] D. A. Kessler and H. Levine, *Pattern selection in three dimensional dendritic growth*, Acta Metall., 36 (1988), pp. 2693–2706.
[18] Y. Kondo and K. Takayanagi, *Gold nanobridge stabilized by surface structure*, Phys. Rev. Lett., 79 (1997), pp. 3455–3458.
[19] M. S. McCallum, P. W. Voorhees, M. J. Miksis, S. H. Davis, and H. Wong, *Capillary instabilities in solid thin films: Lines*, J. Appl. Phys., 79 (1996), pp. 7604–7611.
[20] G. B. McFadden, S. R. Coriell, and R. F. Sekerka, *Effect of surface tension anisotropy on cellular morphologies*, J. Crystal Growth, 91 (1988), pp. 180–198.
[21] G. B. McFadden, S. R. Coriell, and R. F. Sekerka, *Effect of surface free energy anisotropy on dendrite tip shape*, Acta Mater., 48 (2000), pp. 3177–3181.
[22] D. H. Michael, *Meniscus stability*, Ann. Rev. Fluid Mech., 13 (1981), pp. 189–215.
[23] M. E. Toimil Molares, A. G. Balogh, T. W. Cornelius, R. Neumann, and C. Trautmann, *Fragmentation of nanowires driven by Rayleigh instability*, Appl. Phys. Lett., 85 (2004), pp. 5337–5339.
[24] M. E. Toimil Molares, J. Brötz, V. Buschmann, D. Dobrev, R. Neumann, R. Scholz, I. U. Schuchert, C. Trautmann, and J. Vetter, *Etched heavy ion tracks in polycarbonate as template for copper nanowires*, Nucl. Instrum. Methods Phys. Res. B, 185 (2001), pp. 192–197.
[25] W. W. Mullins, *Solid surface morphologies governed by capillarity*, in Metal Surfaces: Structure, Energetics, and Kinetics, ASM, Metals Park, OH, 1963, pp. 17–66.
[26] Y. T. Pang, G. W. Meng, Y. Zhang, Q. Fang, and L. D. Zhang, *Copper nanowire arrays for infrared polarizer,* Appl. Phys. A, 76 (2003), pp. 533–536.
[27] J. A. F. Plateau, *Experimental and theoretical researches on the figures of equilibrium of a liquid mass withdrawn from the action of gravity*, Annual Reports of the Smithsonian Institution (1863–1866), 1873, pp. 270–285.
[28] Lord Rayleigh, *On the instabilities of jets*, Proc. Lond. Math. Soc., 10 (1878), pp. 4–13.

[29] G. Riveros, H. Gómez, A. Cortes, R. E. Marotti, and E. A. Dalchiele, *Crystallographically-oriented single-crystalline copper nanowire arrays electrochemically grown into nanoporous anodic templates*, Appl. Phys. A, 81 (2005), pp. 17–24.

[30] C. Rottman and M. Wortis, *Statistical mechanics of equilibrium crystal shapes: Interfacial phase diagrams and phase transitions*, Phys. Rep., 103 (1984), pp. 59–79.

[31] R. V. Roy and L. W. Schwartz, *On the stability of liquid ridges*, J. Fluid Mech., 391 (1999), pp. 293–318.

[32] J. S. Stölken and A. M. Glaeser, *The morphological evolution of cylindrical rods with anisotropic surface free energy via surface diffusion*, Scripta Metall. Mater., 27 (1992), pp. 449–453.

[33] J. E. Taylor, *Mean curvature and weighted mean curvature*, Acta Metall. Mater., 40 (1992), pp. 1475–1485.

[34] R. G. Voigt, D. Gottlieb, and M. Y. Hussaini, *Spectral Methods for Partial Differential Equations*, SIAM, Philadelphia, 1984.

[35] P. W. Voorhees, S. R. Coriell, G. B. McFadden, and R. F. Sekerka, *The effect of anisotropic crystal-melt surface tension on grain boundary groove morphology*, J. Crystal Growth, 67 (1984), pp. 425–440.

[36] W. L. Winterbottom, *Equilibrium shape of a small particle in contact with a foreign substrate*, Acta Metall., 15 (1967), pp. 303–310.

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1188–1208

THE EXIT PROBLEM IN A NONLINEAR SYSTEM DRIVEN BY $1/f$ NOISE: THE DELAY LOCKED LOOP*

S. LANDIS†, B. Z. BOBROVSKY†, AND Z. SCHUSS‡

Abstract. The frequency generated by high frequency oscillators contains a small but significant noise component known as phase noise, also known as oscillator noise or phase jitter. The phase noise belongs to the family of stochastic processes with spectra $1/f^\alpha$, which exhibits scale invariance (or self-similarity) and a long-term correlation structure that decays polynomially in time. Both the phase and thermal noises cause errors in receivers that contain the oscillators. In particular, they cause losses of lock in phase tracking systems such as the phase locked loop in coherent systems, which include cellular phones, global positioning systems (GPS), and radar (e.g., synthetic aperture radar (SAR)), and in the delay locked loop (DLL), which is an important component of code division multiple access receivers and interface to modern memory modules, such as double data rate synchronous dynamic random access memory. The mean time to lose lock (MTLL) is well known to be an important design objective for various tracking loops. The evaluation of the MTLL is known in the mathematical literature as the exit problem for a dynamical system driven by noise, which is the problem of calculating the mean time for the noisy trajectories to reach the boundary of the domain of attraction of a stable point of the noiseless dynamics. In this paper we develop an analytic approach to the evaluation of the leading order term for MTLL of a second order DLL, due to both the non-Markovian $1/f^\alpha$ noise and to thermal white noise. The method is applicable to more general systems driven by a wide class of phase noises. The keys to the solution of this exit problem are the construction of a series of higher order Markovian processes that converge to the non-Markovian $1/f^\alpha$ noise and the asymptotic solution to a multidimensional elliptic boundary value problem that the mean first passage time (MFPT) satisfies.

Key words. exit problem, phase noise, fractional Brownian motion, loss of lock, delay locked loop, mean time to lose lock, phase locked loop, $1/f$ noise

AMS subject classifications. 34E05, 34E20, 93C10, 93C35

DOI. 10.1137/050627666

1. Introduction.

In communication theory random signals are often described as the output of a dynamical system driven by noise [1]. Various types of noise are used to model different signals, including white noise, colored noise, shot noise, and so on. When the signal is carried by the phase of the wave, such as in phase modulation (PM), frequency modulation (FM), and various keying modulations [1], [2], as well as in global positioning systems (GPS) and other signal tracking systems, the main component that generates the signal is an oscillator [3]. The frequency generated by high frequency oscillators is not very stable; it drifts randomly and contains a small but significant noise component known as *phase noise*. Both the random drift and thermal noise cause errors in the receivers of these signals. In particular, they cause losses of lock in phase tracking systems such as the phase locked loop (PLL) in coherent systems [1], [2], [4], which include cellular phones [3], GPS [5], and radar (e.g., synthetic aperture radar (SAR)) [6], and in the delay locked loop (DLL) [7], [8], which is an important component of code division multiple access (CDMA) receivers [3], GPS, and interface to modern memory modules such as double data rate synchronous

*Received by the editors March 25, 2005; accepted for publication (in revised form) December 28, 2005; published electronically March 24, 2006.
http://www.siam.org/journals/siap/66-4/62766.html
†Department of Electrical Engineering – Systems, School of Electrical Engineering, Tel-Aviv University, Ramat-Aviv, 69978 Tel-Aviv, Israel (slandis@eng.tau.ac.il, bobi@eng.tau.ac.il).
‡Department of Mathematics, Tel-Aviv University, 69978 Tel-Aviv, Israel (schuss@post.tau.ac.il).

dynamic random access memory (DDR SDRAM) [9]. The mean time to lose lock (MTLL) is well known to be an important design objective for various tracking loops [2], [4].

Phase noise, also known as oscillator noise or phase jitter, is a well-known problem that does not yet have a full physical model (a recent example of a physical model is found in [10]) or extensive tools for mathematical manipulation. The phase drift and noise may be due to impurities, imperfections, thermal fluctuations, and other factors in the oscillator's crystal. The phase noise is usually described as having four parts [3]: the first is "frequency flicker" with power spectrum $1/f^3$; the second is "flat frequency" with power spectrum $1/f^2$; the third is "phase flicker" with power spectrum of the form $1/f$; and finally, the fourth is a "flat spectrum" phase. Often a white noise term is added to represent thermal noise. This thermal noise should not be confused with the "flat spectrum" part of the phase noise. The family of stochastic processes with spectra $1/f^\alpha$ is of growing interest in many fields of research due to the wide variety of data for which they are inherently suited [11], [12]. This family of processes exhibits scale invariance (or self-similarity) and a long-term correlation structure that decays polynomially in time rather than exponentially, as is the case for the well-studied family of autoregressive moving average (ARMA) processes [13]. The long-term correlation structure of $1/f$ noise is due to the absence of a low frequency cutoff in the spectrum, which results in nonstationarity. This means that an approximation that has a low frequency cutoff (flattening of the spectrum below a certain frequency) has a finite correlation structure. Furthermore, "ideal" $1/f$ Gaussian noise cannot exist since the single-time variance of the process is not finite. A model that has a low frequency cutoff, as in our case, can be Gaussian. A considerable body of work has been devoted to $1/f^\alpha$ processes (see [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], and references therein). An approximation to $1/f$ noise by an output of a linear system of first order stochastic differential equations, driven by a vector of white noises, is given in [23]. The structure of this model is similar to that of the present one but is derived from different considerations. The model in [23] is based on a physical description of diffusion and transport that leads to $1/f$ spectrum. In contrast, our model is based on a purely mathematical construction and can be generated for any $1/f^\alpha$ spectrum and, in general, for a wide class of spectra. Further, our model is constructed as the output of a system driven by a single noise source, whereas Milotti's model is driven by many noise sources. Our construction may be more appropriate for linear estimation problems driven by $1/f$ noise because in the absence of observation noise, completely accurate reconstruction of the state of the system is possible only if the rank of the observed variable is not lower than the rank of the noise driving the system [24]. Both models can be used for the calculation of the mean first passage time (MFPT) because they are based on a system of first order stochastic differential equations driven by white noise.

The MTLL of coherent and noncoherent pseudonoise code tracking loops, due to white (e.g., thermal) noise, has been studied extensively in the literature [7], [8]. A more analytic approach to the second order DLL was given in [25], where an asymptotic expansion of the solution for the MTLL was derived using singular perturbation theory. The MTLL in systems driven by $1/f$ noise, or more specifically, the effects of phase noise on a DLL, has not been solved analytically in the literature to date, although its existence and significance are well known [3]. Our aim in the specific application considered in this paper is to develop an analytic approach to the evaluation of the leading order term for the MTLL of a second order DLL due to both phase and thermal white noise. It should be noted that in this article we calculate

the asymptotic rate at which the MTLL grows exponentially when the dimensionless noise strength tends to zero. The result is a single leading order term in an asymptotic expansion.

The evaluation of the MTLL is known in the mathematical literature as the *exit problem* for a dynamical system driven by noise [26], which is the problem of calculating the mean time for the noisy trajectories to reach the boundary of the domain of attraction of a stable point of the noiseless dynamics. The problem has been extensively studied for the case of Itô dynamics driven by small white noise (see the comprehensive review [27]). The key to the calculation of the MFPT to the boundary is the construction of an asymptotic solution to a boundary value problem for the elliptic partial differential equation that the MFPT satisfies.

When the noise is non-Markovian or cannot be imbedded in a higher order Markovian process, the calculation of the MFPT is complicated by the absence of differential or integral equations that the MFPT satisfies. To circumvent this difficulty, methods have been developed to approximate the noise by a Markov process, e.g., by an output of a finite-dimensional Itô system. The MFPT is then calculated for the approximating Itô system, and convergence is shown as the order of approximation increases [28]. This procedure requires the development of an asymptotic method for the calculation of the MFPT in Itô systems of high dimension.

Our approach is based on constructing a convergent sequence of rational approximations to $1/f$ (see [28] and references therein). The approximate $1/f$ noise is a component of a higher dimensional (Markovian) diffusion process. We construct a sequence of approximations to the noise by multidimensional diffusion processes, whose power spectral densities are obtained by truncating the continued fraction expansion of the Laplace transform of the function $1/\sqrt{s}$ about $s = 1$. The resulting approximation to the $1/f$ noise is represented by a set of first order linear Itô equations. The resulting rational spectra approximate $1/f$ in finite intervals without ripples. The width of these intervals can be increased arbitrarily by increasing the degree of the truncation. Our model is amenable to asymptotic analysis of the exit problem for high-dimensional diffusion processes. We develop a multidimensional singular perturbation theory for the solution of the exit problem for multidimensional diffusion processes and apply it to the calculation of the leading order term of MTLL in a DLL with an approximate $1/f^3$ phase noise. The dynamics of the delay estimator error is driven by the derivative of the phase noise, which for $1/f^3$ is $1/f$ noise. It is shown that truncation of the model for the phase noise results in a convergent sequence of approximations for the first approximation of the MTLL (the exponential growth rate).

Our main result is the calculation of the leading order term for the MTLL in a specific second order DLL loop under the combined influence of $1/f^3$ phase noise and additive thermal white noise. Specifically, we find explicit asymptotic expressions for the MTLL for Markovian approximations to the phase noise. The calculations indicate that as the degree of approximation increases, the MTLL converges to a finite limit. Actually, no significant change in the MTLL is observed after truncation of the phase noise model at order 10. From a practical point of view it is most important to optimize the system at a low carrier-to-noise ratio (CNR), because this is where the system will stop operating. It turns out that as CNR decreases, the needed order of truncation decreases. We also show that the filter can be optimized for maximal MTLL when both composite phase and thermal white noise drive the system.

2. The leading order term for the MTLL problem for a second order DLL with approximated $1/f^3$ phase noise. The baseband equivalent model for

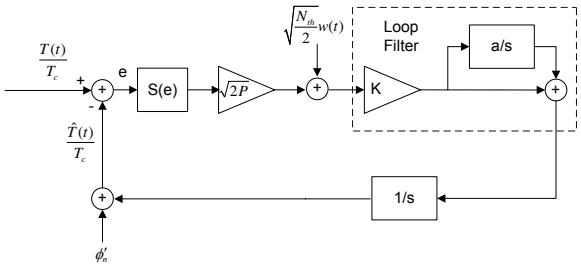


FIG. 1. *Baseband equivalent model for nonlinear DLL.*

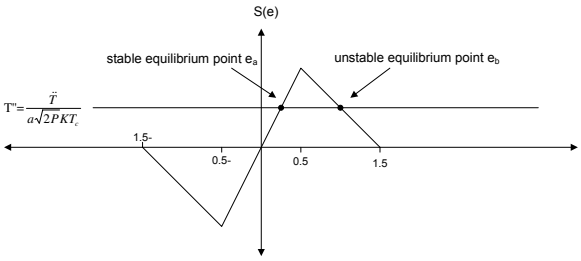


FIG. 2. *S-curve for early-late DLL problem and the corresponding equilibrium points.*

the second order nonlinear early-late DLL shown in Figure 1 [29] has normalized channel propagation delay $T(t)/T_c$, total power of received signal P, gains K and a that define the loop's filter, and phase noise $\tilde{\phi}_n$ with power spectrum

$$S_{\tilde{\phi}_n\tilde{\phi}_n}(f) = \frac{N_{ph}}{2\,|2\pi f|^3}. \tag{2.1}$$

The loop filter used is of proportional integration type [30] with a zero at $s = -a$. The parameter $K[\sqrt{\text{Hz}}]$ determines the loop's gain, and the dimensionless parameter $a \neq 0$ stabilizes the loop. This type of loop can handle relative velocity $\dot{T}$ without a steady state error, and relative acceleration $\ddot{T}$ with a steady state error. The piecewise linear S-curve for the early-late DLL used in spread spectrum synchronization of long PN (pseudonoise) sequences is given by [29] (see Figure 2)

$$S(e) = \begin{cases} 2e, & |e| \le \frac{1}{2}, \\ 1.5 - e, & \frac{1}{2} < |e| < \frac{3}{2}, \\ -1.5 - e, & -\frac{3}{2} < |e| < -\frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

The resulting equations describing the system are

$$e = \frac{T}{T_c} + \tilde{\phi}_n - \frac{\hat{T}}{T_c}, \tag{2.2}$$

$$\dot{z} = aK\sqrt{2P}S(e) + K\sqrt{\frac{N_{th}}{2}}\,\dot{w}(t),$$

$$\frac{\dot{\hat{T}}}{T_c} = z + K\sqrt{2P}S(e) + K\sqrt{\frac{N_{th}}{2}}\,\dot{w}(t),$$

where e is the delay estimation error, $\dot{z}$ is the output of the loop filter, and the last equation is the output of the integrator. Here $w(t)$ is the standard Wiener process (Brownian motion), whose "derivative" $\dot{w}(t)$ is standard δ-correlated Gaussian white noise [26], independent of $\nu(t)$.

Differentiating e and setting $\tilde{z} = z - \frac{\dot{T}}{T_c}$, equations (2.2) become

$$\begin{aligned}
\dot{e} &= -\tilde{z} - K\sqrt{2P}S(e) + \dot{\tilde{\phi}}_n - K\sqrt{\frac{N_{th}}{2}}\,\dot{w}(t),\\
\dot{\tilde{z}} &= aK\sqrt{2P}S(e) - \frac{\ddot{T}}{T_C} + K\sqrt{\frac{N_{th}}{2}}\,\dot{w}(t).
\end{aligned} \tag{2.3}$$

The spectral power density (2.1) indicates that the new noise process $\phi_n = \dot{\tilde{\phi}}_n$ is well defined and that its power spectral density function is given by $S_{\phi_n\phi_n}(f) = \frac{N_{ph}}{2|2\pi f|}$.

Next, we construct an approximate $1/f$ Gaussian noise by passing white Gaussian noise through a filter, whose response in the Laplace domain is [20]

$$H(s) = \frac{1}{\sqrt{s}}. \tag{2.4}$$

This filter cannot be realized as a Markovian process in a straightforward fashion, and therefore the standard tools of Markov processes are not available for the study of the effects of the $1/f$ noise in dynamical systems. We construct a sequence of rational approximations to (2.4) by truncating its continued fraction representation in a fashion similar to that used in [28] and references therein. Although (2.4) is not an analytic function near the origin, it is analytic at any nonzero s such as $s = \omega_0 > 0$, so it has the continued fraction representation

$$\frac{1}{\sqrt{\tilde{s}+\omega_0}} = \cfrac{1}{1+\cfrac{1}{\cfrac{2\omega_0}{\tilde{s}}+\cfrac{1}{2+\cfrac{1}{\cfrac{2\omega_0}{\tilde{s}}+\cfrac{1}{2+\ddots}}}}}, \tag{2.5}$$

where $\tilde{s} = \omega_0(s-1)$. Thus (2.4) is

$$H(s) = \cfrac{1}{1+\cfrac{1}{\cfrac{2}{s-1}+\cfrac{1}{2+\cfrac{1}{\cfrac{2}{s-1}+\cfrac{1}{2+\ddots}}}}}, \tag{2.6}$$

which converges for $|s-\omega_0| < \omega_0$. Next, we define an approximate $1/f$ noise through the Laplace transform relation

$$\Phi_n(s) = H(s)V(s), \tag{2.7}$$

where the power spectral density of a white Gaussian process is given by

$$S_{vv}(f) = N_{ph}. \tag{2.8}$$

Truncating the continued fraction and using (2.6), (2.7), we obtain the system of $2N$ equations

$$\begin{aligned}
V(s) &= \Phi_n(s) + Y_1(s), \quad \Phi_n(s) = \frac{2}{s-1}Y_1(s) + Y_2(s),\\
Y_1(s) &= 2Y_2(s) + Y_3(s), \quad Y_2(s) = \frac{2}{s-1}Y_3(s) + Y_4(s),\\
Y_3(s) &= 2Y_4(s) + Y_5(s), \quad Y_4(s) = \frac{2}{s-1}Y_5(s) + Y_6(s),\\
&\vdots \qquad\qquad\qquad\quad \vdots\\
Y_{2N-3}(s) &= 2Y_{2N-2}(s) + Y_{2N-1}(s), \quad Y_{2N-2}(s) = \frac{2}{s-1}Y_{2N-1}(s) + Y_{2N}(s),\\
2Y_{2N-1}(s) &= 4Y_{2N}(s) + (s-1)Y_{2N},
\end{aligned} \tag{2.9}$$

where N denotes the order of approximation to the $1/f$ noise.

We note that all the state variables $Y_{2j+1}(s)$ in (2.9) can be eliminated by a linear transformation. To transform the system (2.9) into the time domain, we denote by $\nu(t)$ a standard Gaussian white noise and denote the state variables in the time domain by lowercase letters. Then (2.9) is transformed into the Itô system

$$\begin{aligned}
\dot{y}_{2N}(t) &= y_{2N}(t) - y_{2N}(0) + 2\left[\nu(t) - \phi_n(t) - 2\sum_{k=1}^{N} y_{2k}(t)\right],\\
\dot{y}_{2N-2}(t) &= y_{2N-2}(t) - y_{2N-2}(0) - 4y_{2N}(t) + 4\left[\nu(t) - \phi_n(t) - 2\sum_{k=1}^{N-1} y_{2k}(t)\right],\\
\dot{y}_{2N-4}(t) &= y_{2N-4}(t) - y_{2N-4}(0) - 4y_{2N}(t) - 8y_{2N-2}(t)\\
&\quad + 6\left[\nu(t) - \phi_n(t) - 2\sum_{k=1}^{N-2} y_{2k}(t)\right],\\
&\vdots\\
\dot{y}_2(t) &= y_2(t) - y_2(0) - 4\sum_{m=1}^{N-1} m y_{2(N-m+1)}(t) + 2N\left[\nu(t) - \phi_n(t) - 2y_2(t)\right],\\
\dot{\phi}_n(t) &= -(2N+1)\phi_n(t) - \phi_n(0) - 4\sum_{m=1}^{N} m y_{2(N-m+1)}(t) + 2(N+1)\nu(t).
\end{aligned} \tag{2.10}$$

Thus the Nth approximation to the $1/f$ noise process is an output of a Markovian system of $N+1$ linear stochastic differential equations of Itô type [26]. Finally, since

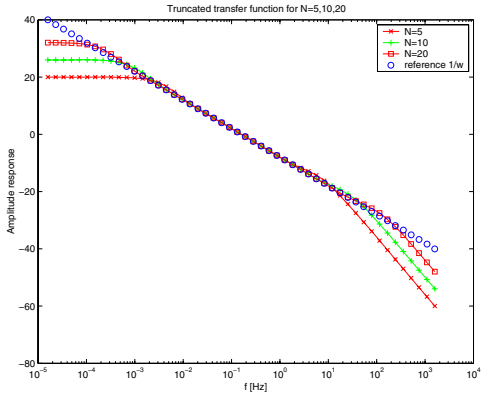


FIG. 3. *Frequency response of continued fraction approximations of* $1/\omega$ *with* ω_0=1, *truncated at* $N = 5$ *(red* $\times$*'s),* $N = 10$ *(green), and* $N = 20$ *(red squares). The reference* $1/\omega$ *is blue circles.*

the expected value of the $1/f$ noise process is zero, we find the initial conditions by taking the expectation of (2.10)

$$\phi_n(0) = y_2(0) = y_4(0) = \cdots = y_{2N}(0) = 0. \tag{2.11}$$

An interesting feature of our model is that for $N = 0$, (2.10) becomes an Ornstein–Uhlenbeck process. This type of process is commonly used for colored Gaussian noise models. In Figure 3 the frequency response [31] of the truncated transfer function $H(s)$ (2.6) is given for the approximation of $1/\omega$, where $\omega = 2\pi f$, with $N = 5$, $N = 10$, and $N = 20$, to show how the range of validity of the approximation expands with increasing N, and that the approximation is without ripples.

Thus, using (2.3), (2.10), we obtain the system

$$\begin{aligned}
\dot{e}(t) &= -\tilde{z}(t) - K\sqrt{2P}S(e(t)) + \phi_n(t) - K\sqrt{\frac{N_{th}}{2}}\,\dot{w}(t), \qquad (2.12)\\
\dot{\tilde{z}}(t) &= aK\sqrt{2P}S(e(t)) - \frac{\ddot{T}}{T_C} + K\sqrt{\frac{N_{th}}{2}}\,\dot{w}(t),\\
\dot{y}_{2N}(t) &= y_{2N}(t) + 2\left[\sqrt{\frac{N_{ph}}{2}}\,\nu(t) - \phi_n(t) - 2\sum_{k=1}^{N} y_{2k}(t)\right],\\
\dot{y}_{2N-2}(t) &= y_{2N-2}(t) - 4y_{2N}(t) + 4\left[\sqrt{\frac{N_{ph}}{2}}\,\nu(t) - \phi_n(t) - 2\sum_{k=1}^{N-1} y_{2k}(t)\right],\\
\dot{y}_{2N-4}(t) &= y_{2N-4}(t) - 4y_{2N}(t) - 8y_{2N-2} + 6\left[\sqrt{\frac{N_{ph}}{2}}\,\nu(t) - \phi_n(t) - 2\sum_{k=1}^{N-2} y_{2k}(t)\right],\\
&\vdots\\
\dot{y}_2(t) &= y_2(t) - 4\sum_{m=1}^{N-1} m y_{2(N-m+1)} + 2N\left[\sqrt{\frac{N_{ph}}{2}}\,\nu(t) - \phi_n(t) - 2y_2(t)\right],
\end{aligned}$$

$$\dot{\phi}_n(t) = -(2N+1)\phi_n(t) - 4\sum_{m=1}^{N} m y_{2(N-m+1)}(t) + 2(N+1)\sqrt{\frac{N_{ph}}{2}}\,\nu(t).$$

Next, we normalize the equations so that the noise term converges to zero as the CNR term P/N_{ph} increases to infinity. The CNR, measured in Hz, is a well-accepted engineering quantity [5]. We introduce dimensionless time and define the auxiliary variables

$$\tilde{t} = \sqrt{2P}Kt, \quad \beta = \frac{\tilde{z}}{a}, \quad T'' = \frac{\ddot{T}}{a\sqrt{2P}KT_c} \tag{2.13}$$

to convert to the nondimensional system

$$\begin{aligned}
\dot{e}\left(\tilde{t}\right) &= -\frac{a}{\sqrt{2P}K}\beta\left(\tilde{t}\right) - S\left(e\left(\tilde{t}\right)\right) + \frac{\phi_n\left(\tilde{t}\right)}{\sqrt{2P}K} - \sqrt{\frac{KN_{th}}{2\sqrt{2P}}}\,\dot{w}(\tilde{t}),\\
\dot{\beta}\left(\tilde{t}\right) &= S\left(e\left(\tilde{t}\right)\right) - T'' + \sqrt{\frac{KN_{th}}{2\sqrt{2P}}}\,\dot{w}(\tilde{t}),\\
\dot{y}_{2N}\left(\tilde{t}\right) &= \frac{1}{K\sqrt{2P}}\left\{y_{2N}\left(\tilde{t}\right) + 2\left[-\phi_n\left(\tilde{t}\right) - 2\sum_{k=1}^{N} y_{2k}\left(\tilde{t}\right)\right]\right\} + 2\sqrt{\rho}\,\nu\left(\tilde{t}\right),\\
\dot{y}_{2N-2}\left(\tilde{t}\right) &= \frac{1}{K\sqrt{2P}}\left\{y_{2N-2}\left(\tilde{t}\right) - 4y_{2N}\left(\tilde{t}\right) + 4\left[-\phi_n\left(\tilde{t}\right) - 2\sum_{k=1}^{N-1} y_{2k}\left(\tilde{t}\right)\right]\right\}\\
&\quad + 4\sqrt{\rho}\,\nu\left(\tilde{t}\right),\\
\dot{y}_{2N-4}\left(\tilde{t}\right) &= \frac{1}{K\sqrt{2P}}\left\{y_{2N-4}\left(\tilde{t}\right) - 4y_{2N}\left(\tilde{t}\right) - 8y_{2N-2}\left(\tilde{t}\right) + 6\left[-\phi_n\left(\tilde{t}\right) - 2\sum_{k=1}^{N-2} y_{2k}\left(\tilde{t}\right)\right]\right\}\\
&\quad + 6\sqrt{\rho}\,\nu\left(\tilde{t}\right),\\
&\vdots\\
\dot{y}_2\left(\tilde{t}\right) &= \frac{1}{K\sqrt{2P}}\left\{y_2\left(\tilde{t}\right) - 4\sum_{m=1}^{N-1} m y_{2(N-m+1)}\left(\tilde{t}\right) + 2N\left[-\phi_n\left(\tilde{t}\right) - 2y_2\left(\tilde{t}\right)\right]\right\}\\
&\quad + 2N\sqrt{\rho}\,\nu\left(\tilde{t}\right),\\
\dot{\phi}_n\left(\tilde{t}\right) &= \frac{1}{K\sqrt{2P}}\left\{-(2N+1)\phi_n\left(\tilde{t}\right) - 4\sum_{m=1}^{N} m y_{2(N-m+1)}\left(\tilde{t}\right)\right\} + 2(N+1)\sqrt{\rho}\,\nu\left(\tilde{t}\right),
\end{aligned} \tag{2.14}$$

where the dimensionless noise level is given by $\rho = \frac{N_{ph}}{2K\sqrt{2P}}$. For small values of ρ the system (2.14) can be viewed as a small stochastic perturbation of a nonlinear dynamical system that has a stable equilibrium at the point

$$e_a = \frac{T''}{2}, \quad \beta_a = -\frac{\ddot{T}}{T_c a^2}, \quad \phi_{n,a} = 0, \quad y_{2i,a} = 0, \quad 1 \le i \le N, \tag{2.15}$$

and an unstable equilibrium point at

$$e_b = \frac{3}{2} - |T''|, \quad \beta_b = -\frac{\ddot{T}}{T_c a^2}, \quad \phi_{n,b} = 0, \quad y_{2i,b} = 0, \quad 1 \le i \le N, \tag{2.16}$$

which we refer to as the *saddle point*.

The stable equilibrium point (2.15) of the system (2.14) has a domain of attraction D. This means that any noiseless trajectory of (2.14) starting in D converges to the stable equilibrium point (2.15). The boundary of the region D is denoted ∂D. As long as a trajectory of the stochastic system (2.14) remains in D, the DLL is said to be in a locked state. Upon exiting the region D through the boundary ∂D, the DLL is said to have lost lock. The exact description of the boundary ∂D is complex and is omitted here; however, in the limit of weak noise, the exit from D occurs in the immediate neighborhood of the saddle point. Thus the calculation of the MTLL is the classical exit problem of a dynamical system from the domain of attraction of a stable point under the influence of small noise [26].

We denote a trajectory of (2.14) by

$$\boldsymbol{x}^T(t) = [e(t),\ \beta(t),\ y_{2N}(t),\ y_{2N-2}(t),\ \ldots,\ y_2(t),\ \phi_n(t)]\,.$$

For each trajectory of (2.14) that starts in D, we denote by τ_D the first time it reaches the boundary ∂D (the first passage time to the boundary),

$$\tau_D = \inf\{t \geq 0 \,|\, \boldsymbol{x}(t) \in \partial D,\, \boldsymbol{x}(0) \in D\}\,, \tag{2.17}$$

and its conditional expectation

$$\bar{\tau}_D(\boldsymbol{x}) = E[\tau_D \,|\, \boldsymbol{x}(0) = \boldsymbol{x}].$$

The MTLL is defined as

$$\bar{t}_L(\boldsymbol{x}) = 2\bar{\tau}_D(\boldsymbol{x}) \tag{2.18}$$

because once on ∂D, a trajectory is equally likely to return to D immediately or to leave D for a long time [32]. In the case of small ρ an analytic approximation to the MTLL can be obtained, as described below.

3. Review of the multidimensional exit problem for diffusions. The approximation scheme used in the previous section reduces the loss of lock problem for a DLL with $1/f^3$ phase noise to the classical problem of escape of a multidimensional diffusion process from the domain of attraction of an attractor [26], [32], [33], [34], [35]. In the following, we outline a singular perturbation method for constructing an asymptotic approximation to the solution of the Fokker–Planck equation (FPE) and to the escape problem in high dimensions.

Consider the autonomous multidimensional system

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{a}(\boldsymbol{x}) + \sqrt{2\varepsilon}\,\boldsymbol{b}(\boldsymbol{x})\boldsymbol{\nu}(t), \tag{3.1}$$

$$\boldsymbol{x}(0) = \boldsymbol{x} \tag{3.2}$$

in a domain D in $\boldsymbol{R}^n$ for piecewise smooth flows (vector fields) $\boldsymbol{a}(\boldsymbol{x})$ and an $n \times m$ noise matrix $\boldsymbol{b}(\boldsymbol{x})$. Here $\boldsymbol{\nu}(t)$ is a vector of m independent standard Gaussian white noises [26]. We assume that the noiseless dynamics

$$\dot{\boldsymbol{x}} = \boldsymbol{a}(\boldsymbol{x}) \tag{3.3}$$

has a unique critical point $\boldsymbol{x}_0$ in D and that it is a global attractor in D. This means that

$$\boldsymbol{a}(\boldsymbol{x}_0) = \boldsymbol{0},$$

and we assume that the eigenvalues of the matrix

$$\boldsymbol{A} = \left\{ \frac{\partial a^i(\boldsymbol{x}_0)}{\partial x^j} \right\}_{i,j=1}^{n} \tag{3.4}$$

of the linearized system

$$\dot{\boldsymbol{z}} = \boldsymbol{A}\boldsymbol{z} \tag{3.5}$$

have negative real parts. Thus the trajectories of the system (3.3) that start in D cannot reach the boundary ∂D.

The FPE for the stationary probability density function $p_\varepsilon(\boldsymbol{y} \,|\, \boldsymbol{x})$ of the solution $\boldsymbol{x}(t, \varepsilon)$ of (3.1), (3.2), with a source at $\boldsymbol{x}$ and absorption in ∂D, is [34]

$$-\sum_{i=1}^{n} \frac{\partial}{\partial y^i} \left[a^i(\boldsymbol{y}) p_\varepsilon(\boldsymbol{y} \,|\, \boldsymbol{x}) \right] + \sum_{i,j=1}^{n} \varepsilon \frac{\partial^2}{\partial y^i \partial y^j} \left[\sigma^{i,j}(\boldsymbol{y}) p_\varepsilon(\boldsymbol{y} \,|\, \boldsymbol{x}) \right] = -\delta(\boldsymbol{y} - \boldsymbol{x}) \tag{3.6}$$

with the boundary condition

$$p_\varepsilon(\boldsymbol{y} \,|\, \boldsymbol{x}) \,|\, _{\boldsymbol{x} \in D, \boldsymbol{y} \in \partial D} = 0,$$

where

$$\boldsymbol{\sigma}(\boldsymbol{y}) = \boldsymbol{b}(\boldsymbol{y}) \boldsymbol{b}^T(\boldsymbol{y}).$$

The function $p_\varepsilon(\boldsymbol{y} \,|\, \boldsymbol{x})$ develops singularities in the domain and on its boundary as $\varepsilon \to 0$. We resolve these singularities by constructing an approximate solution that contains all the singularities of $p_\varepsilon(\boldsymbol{y} \,|\, \boldsymbol{x})$ at $\varepsilon = 0$. First, we transform the FPE (3.6) by seeking a solution in the WKB form

$$p_\varepsilon(\boldsymbol{y} \,|\, \boldsymbol{x}) = K_\varepsilon(\boldsymbol{y} \,|\, \boldsymbol{x}) \exp\left\{ -\frac{\Psi(\boldsymbol{y})}{\varepsilon} \right\}, \tag{3.7}$$

where $K_\varepsilon(\boldsymbol{y} \,|\, \boldsymbol{x})$ has an asymptotic series expansion in powers of ε for $\boldsymbol{x}$ in the domain and $\Psi(\boldsymbol{y})$ is a regular function. The asymptotic approximation to the MFPT for small ε is given by

$$\bar{\tau}(\boldsymbol{x}) = C(\varepsilon) \exp\left\{ \frac{\hat{\Psi}}{\varepsilon} \right\} (1 + o(1)), \tag{3.8}$$

where $C(\varepsilon)$ has an asymptotic series expansion in powers of ε and $\hat{\Psi}$ is the minimum of the eikonal function $\Psi(\boldsymbol{x})$ on the boundary of the domain of attraction of the stable equilibrium point $\boldsymbol{x}_0$. Equation (3.8) indicates that to leading order $\bar{\tau}(\boldsymbol{x})$ is independent of $\boldsymbol{x}$.

The essential singularity of $p_\varepsilon(\boldsymbol{y} \,|\, \boldsymbol{x})$ inside D is captured by the exponential term in (3.7) and that on ∂D by the preexponential factor $K_\varepsilon(\boldsymbol{y} \,|\, \boldsymbol{x})$. Substituting (3.7)

in (3.6) and collecting like powers of ε, we obtain at the leading order the first order *eikonal equation*

$$\sum_{i,j=1}^{n} \sigma^{i,j}(\boldsymbol{y}) \frac{\partial \Psi(\boldsymbol{y})}{\partial y^i} \frac{\partial \Psi(\boldsymbol{y})}{\partial y^j} + \sum_{i=1}^{n} a^i(\boldsymbol{y}) \frac{\partial \Psi(\boldsymbol{y})}{\partial y^i} = 0. \tag{3.9}$$

The eikonal equation has the form of a Hamilton–Jacobi equation [37] and is solved by the method of characteristics [38], [39]. Setting $\boldsymbol{p} = \nabla \Psi(\boldsymbol{x})$, (3.9) becomes

$$\sum_{i,j=1}^{n} \sigma^{i,j}(\boldsymbol{x}) p^i p^j + \sum_{i=1}^{n} a^i(\boldsymbol{x}) p^i = 0,$$

and the characteristic equations (or rays [33]) are the solutions of

$$\frac{d\boldsymbol{x}}{ds} = 2\boldsymbol{\sigma}(\boldsymbol{x})\boldsymbol{p} + \boldsymbol{a}(\boldsymbol{x}), \tag{3.10}$$

$$\frac{d\boldsymbol{p}}{ds} = -\nabla_{\boldsymbol{x}} \boldsymbol{p}^T \boldsymbol{\sigma}(\boldsymbol{x})\boldsymbol{p} - \nabla_{\boldsymbol{x}} \boldsymbol{a}^T(\boldsymbol{x})\boldsymbol{p}, \tag{3.11}$$

$$\frac{d\Psi}{ds} = \boldsymbol{p}^T \boldsymbol{\sigma}(\boldsymbol{x})\boldsymbol{p}. \tag{3.12}$$

Since the rays that begin near the attractor $(\boldsymbol{x}_0, \boldsymbol{0})$ diverge, we can cover the domain of attraction of $(\boldsymbol{x}_0, \boldsymbol{0})$ with rays emanating from a small neighborhood of $(\boldsymbol{x}_0, \boldsymbol{0})$. To integrate the characteristic equations (3.10), (3.11), initial conditions can be imposed near the attractor $(\boldsymbol{x}_0, \boldsymbol{0})$ by constructing $\Psi(\boldsymbol{x})$ in the form of a power series. The truncation of the power series near the attractor provides an approximation to $\Psi(\boldsymbol{x})$ and to $\boldsymbol{p} = \nabla \Psi(\boldsymbol{x})$, whose error can be made arbitrarily small. Expanding the functions $\Psi(\boldsymbol{x})$, $\boldsymbol{a}(\boldsymbol{x})$, and $\boldsymbol{\sigma}(\boldsymbol{x})$ in powers of $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{x}_0$, we find from the eikonal equation (3.9) that $\nabla \Psi(\boldsymbol{x}_0) = \boldsymbol{0}$, so that the power series expansion of $\Psi(\boldsymbol{x})$ begins as a quadratic form

$$\Psi(\boldsymbol{x}) = \frac{1}{2} \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} + o\left(|\boldsymbol{x}|^2\right) \tag{3.13}$$

and $\boldsymbol{Q}$ is the solution of the Riccati equation [36]

$$2\boldsymbol{Q}\boldsymbol{\sigma}(\boldsymbol{x}_0)\boldsymbol{Q} + \boldsymbol{Q}\boldsymbol{A} + \boldsymbol{A}^T\boldsymbol{Q} = \boldsymbol{0}. \tag{3.14}$$

Obviously, the first term in the power series expansion of $\boldsymbol{p}$ is given by

$$\boldsymbol{p} = \boldsymbol{Q}\boldsymbol{x} + O\left(|\boldsymbol{x}|^2\right). \tag{3.15}$$

Taking the contour

$$\frac{1}{2} \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} = \delta, \tag{3.16}$$

for some small positive δ, as the initial surface for the system (3.10)–(3.12) and using the approximate initial values $\Psi(\boldsymbol{x}) = \delta$ and (3.15) at each point of the surface, we can integrate the system (3.10)–(3.12) analytically or numerically. Once the domain D is covered with characteristics, the approximate value of $\Psi(\boldsymbol{x})$ can be determined

at each point $\boldsymbol{x} \in D$ as the value of the solution $\Psi(s)$ of (3.12) at s, such that the solution of (3.10) satisfies

$$\boldsymbol{x}(s) = \boldsymbol{x}. \tag{3.17}$$

The initial condition on the surface (3.16) determines the unique trajectory of the system (3.10)–(3.12) that satisfies (3.17) for some s. It can be found numerically by the method of shooting.

4. The exit problem for the DLL. The eikonal equation (3.9) corresponding to the stochastic system (2.14) is given by

(4.1)

$$\begin{aligned} H = &\left(-\frac{a}{\sqrt{2P}K}\beta - S(e) + \frac{\phi_n}{\sqrt{2P}K}\right)\frac{\partial\Psi}{\partial e} + \left(S(e) - \frac{T''}{2}\right)\frac{\partial\Psi}{\partial\beta} \\ &+ \frac{1}{\sqrt{2P}K}\sum_{i=1}^{N}\left(y_{2i} - 4\sum_{l=1}^{N-i} l y_{2(N-l+1)} + 2(N-i+1)\left(-\phi_n - 2\sum_{l=1}^{i} y_{2l}\right)\right)\frac{\partial\Psi}{\partial y_{2i}} \\ &+ \frac{1}{\sqrt{2P}K}\left(-(2N+1)\phi_n - 4\sum_{l=1}^{N} l y_{2(N-l+1)}\right)\frac{\partial\Psi}{\partial\phi_n} \\ &+ \left(\sum_{l=1}^{N} 2(N-l+1)\frac{\partial\Psi}{\partial y_{2l}} + 2(N+1)\frac{\partial\Psi}{\partial\phi_n}\right)^2 + \left(-K\sqrt{\frac{N_{th}}{N_{ph}}}\frac{\partial\Psi}{\partial e} + K\sqrt{\frac{N_{th}}{N_{ph}}}\frac{\partial\Psi}{\partial\beta}\right)^2 \\ &= 0. \end{aligned}$$

The solution in the slab $-1/2 \le e \le 1/2$, corresponding to the linear part of the S-curve (see Figure 2), is the quadratic form (3.13), determined by the solution to Lyapunov's (Riccati's) equation (3.14). The system in the linear region can be written as

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{\nu}(t), \tag{4.2}$$

where $\boldsymbol{\nu}(t)$ is a standard Gaussian white noise vector composed of the uncorrelated white noises $w(t)$ and $\nu(t)$ in (2.14) and the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are given by

$$\boldsymbol{A} = \frac{1}{\sqrt{2P}K} \times \begin{bmatrix} -2\sqrt{2P}K & -a & 0 & 0 & 0 & 0 & \cdots & 0 & -1 \\ 2\sqrt{2P}K & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -3 & -4 & -4 & -4 & \cdots & -4 & -2 \\ 0 & 0 & -4 & -7 & -8 & -8 & \cdots & -8 & -4 \\ 0 & 0 & -4 & -8 & -11 & -12 & \cdots & -12 & -6 \\ \vdots & & & & & & & & \\ 0 & 0 & -4 & -8 & -12 & \cdots & -4(N-1) & -4N-1 & -2N \\ 0 & 0 & -4 & -8 & -12 & -16 & \cdots & -4N & -2N-1 \end{bmatrix}$$

and

$$\boldsymbol{B}^T = \sqrt{\frac{N_{ph}}{2K\sqrt{2P}}} \begin{bmatrix} -K\sqrt{\frac{N_{th}}{N_{ph}}} & K\sqrt{\frac{N_{th}}{N_{ph}}} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 2 & 4 & \cdots & 2N & 2(N+1) \end{bmatrix},$$

respectively. The explicit solution of (3.14) can be obtained by using standard symbolic mathematical packages such as Maple or Mathematica.

To solve (4.1) outside the strip $|e| < 1/2$, we use the method of characteristics, as described in section 3 above. We define the components of the vector $\boldsymbol{p}$ by the equations

$$p = \frac{\partial \Psi}{\partial e}, \quad q = \frac{\partial \Psi}{\partial \beta}, \quad \alpha_1 = \frac{\partial \Psi}{\partial y_2}, \quad \alpha_2 = \frac{\partial \Psi}{\partial y_4}, \quad \ldots, \quad \alpha_N = \frac{\partial \Psi}{\partial y_{2N}}, \quad r = \frac{\partial \Psi}{\partial \phi_n}.$$

Now, taking the total derivative of H with respect to time, we get

$$\begin{aligned} \frac{dH}{dt} = 0 = {} & \frac{\partial H}{\partial e}\frac{de}{dt} + \frac{\partial H}{\partial \beta}\frac{d\beta}{dt} + \frac{\partial H}{\partial \phi_n}\frac{d\phi_n}{dt} \\ & + \sum_{i=1}^{N} \frac{\partial H}{\partial y_{2i}}\frac{dy_{2i}}{dt} + \frac{\partial H}{\partial p}\frac{dp}{dt} + \frac{\partial H}{\partial q}\frac{dq}{dt} + \frac{\partial H}{\partial r}\frac{dr}{dt} + \sum_{i=1}^{N} \frac{\partial H}{\partial \alpha_i}\frac{d\alpha_i}{dt}. \end{aligned} \tag{4.3}$$

The characteristic equations (3.10)–(3.12) are given by

$$\begin{aligned} & \frac{de}{dt} = \frac{\partial H}{\partial p}, \quad \frac{dp}{dt} = -\frac{\partial H}{\partial e}, \quad \frac{d\beta}{dt} = \frac{\partial H}{\partial q}, \quad \frac{dq}{dt} = -\frac{\partial H}{\partial \beta}, \\ & \frac{dy_2}{dt} = \frac{\partial H}{\partial \alpha_1}, \quad \frac{d\alpha_1}{dt} = -\frac{\partial H}{\partial y_2}, \quad \frac{dy_4}{dt} = \frac{\partial H}{\partial \alpha_2}, \quad \frac{d\alpha_2}{dt} = -\frac{\partial H}{\partial y_4}, \\ & \qquad \vdots \\ & \frac{dy_{2N}}{dt} = \frac{\partial H}{\partial \alpha_N}, \quad \frac{d\alpha_N}{dt} = -\frac{\partial H}{\partial y_{2N}}, \quad \frac{d\phi_n}{dt} = \frac{\partial H}{\partial r}, \quad \frac{dr}{dt} = -\frac{\partial H}{\partial \phi_n}. \end{aligned} \tag{4.4}$$

Inserting (4.1) and (4.3) into (4.4), we get

$$\begin{aligned} \frac{de}{dt} = {} & -\frac{a}{K\sqrt{2P}}\beta - S(e) + \frac{\phi_n}{K\sqrt{2P}} - 2K^2\frac{N_{th}}{N_{ph}}(-p+q), \\ \frac{d\beta}{dt} = {} & S(e) - T'' + 2K^2\frac{N_{th}}{N_{ph}}(-p+q), \\ \frac{dy_{2i}}{dt} = {} & \frac{1}{K\sqrt{2P}}\left\{ y_{2i} - 4\sum_{l=1}^{N-i} l y_{2(N-l+1)} - 2(N-i+1)\left(\phi_n + 2\sum_{l=1}^{i} y_{2l}\right)\right\} \\ & + 4(N-i+1)\left(\sum_{l=1}^{N} 2(N-l+1)\alpha_l + 2(N+1)r\right) \quad \text{for all } 1 \le i \le N, \end{aligned} \tag{4.5}$$

$$\frac{d\phi_n}{dt} = \frac{1}{K\sqrt{2P}}\left\{-(2N+1)\,\phi_n - \sum_{l=1}^{N} 4l y_{2(N-l+1)}\right\}$$
$$+ 4(N+1)\left(\sum_{l=1}^{N} 2(N-l+1)\alpha_l + 2(N+1)\,r\right),$$
$$\frac{dp}{dt} = (p-q)\,S'(e),$$
$$\frac{dq}{dt} = \frac{a}{K\sqrt{2P}}p$$

and

$$\frac{d\alpha_i}{dt} = -\frac{1}{K\sqrt{2P}}\left\{\alpha_i - 4\sum_{l=i+1}^{N}(N-l+1)\,\alpha_l - 4(N-i+1)\sum_{l=1}^{i}\alpha_i\right\}$$
$$+ \frac{r}{K\sqrt{2P}}4(N-i+1) \quad \text{for all } 1 \le i \le N,$$
$$\frac{dr}{dt} = -\frac{p}{K\sqrt{2P}} + \frac{1}{K\sqrt{2P}}\sum_{i=1}^{N} 2(N-i+1)\,\alpha_i + \frac{r}{K\sqrt{2P}}(2N+1).$$

To complete the solution of (4.1), we must show that $H = 0$ for at least one point. Taking the total derivative $\frac{d\Psi}{dt}$ along a characteristic, and using (4.1) and (4.5), we get

$$\frac{d\Psi}{dt} = \frac{\partial\Psi}{\partial e}\frac{de}{dt} + \frac{\partial\Psi}{\partial\beta}\frac{d\beta}{dt} + \frac{\partial\Psi}{\partial\phi_n}\frac{d\phi_n}{dt} + \sum_{i=1}^{N}\frac{\partial\Psi}{\partial y_{2i}}\frac{dy_{2i}}{dt} \tag{4.6}$$
$$= H + \sum_{i=1}^{N} 2(N-i+1)\,\alpha_i\left(\sum_{l=1}^{N} 2(N-l+1)\,\alpha_l + 2(N+1)\,r\right)$$
$$+ 2(N+1)\left(\sum_{l=1}^{N} 2(N-l+1)\,\alpha_l + 2(N+1)\,r\right) r$$
$$= H + \left(\sum_{l=1}^{N} 2(N-l+1)\,\alpha_l + 2(N+1)\,r\right)^2.$$

Thus, $H = 0$ if

$$\frac{d\Psi}{dt} = \left(\sum_{l=1}^{N} 2(N-l+1)\,\alpha_l + 2(N+1)\,r\right)^2 + K^2\frac{N_{th}}{N_{ph}}(-p+q)^2. \tag{4.7}$$

Equations (4.5) and (4.7) represent the solution of (4.1) on each characteristic curve. The boundary ∂D is spanned by characteristic curves that converge to the saddle point, and Ψ decreases on each characteristic to its value $\hat{\Psi}$ at the saddle point [26].

5. The MTLL in a second order DLL with $1/f^3$ phase noise. The S-curve $S(e)$ for a DLL is given in Figure 2. The stable equilibrium point of the system, in the

absence of relative motion between the transmitter and the receiver, is the point where the S-curve vanishes with positive slope, and the two unstable equilibrium points (the saddle points) are the points where it vanishes with negative slopes. In case there is relative motion with constant acceleration $\ddot{T}$ (see (2.13)), the equilibrium points of the dynamics (2.14) are the points where the S-curve intersects the line $S = T''$ (see Figure 2). For $\ddot{T} \neq 0$ there are one stable equilibrium point and one unstable equilibrium point, given by (2.15) and (2.16), respectively.

The minimum value $\hat{\Psi}$ determines the leading order term (or small ρ exponential growth rate) of the MTLL (see (3.8)), so we have to determine it by finding the characteristic that hits the saddle point and the limiting value of Ψ there. To this end, we start the characteristic on the hyperplane $e = 1/2$, where Ψ is given explicitly by (3.13). Since the characteristic equations are linear in the half-space to the right of the hyperplane $e = 1/2$, all characteristics diverge exponentially fast, except one that corresponds to the only negative eigenvalue of the system matrix (see below). Thus, the starting point of the desired characteristic is the column corresponding to this eigenvalue in the matrix that reduces the system matrix into its Jordan canonical form.

Specifically, we observe in (4.5) and (4.7) that the quasi-potential Ψ is dependent only on the variables $\{p, q, \boldsymbol{\alpha}, r\}$. We define the state vector

$$\boldsymbol{v} = \{p, q, \boldsymbol{\alpha}, r\}^T. \tag{5.1}$$

In the strip $|e| < 1/2$ the S-curve is linear so that the system (4.5) is linear, and with the notation (5.1) it can be written as

$$\dot{\boldsymbol{v}} = \boldsymbol{M}\boldsymbol{v}. \tag{5.2}$$

Since we are looking for the minimum $\hat{\Psi} = \Psi(\boldsymbol{x}_b)$, and $\boldsymbol{x}_b$ is the saddle point of the system (2.14), we need only find the starting point of the characteristic that hits the saddle point (2.16). In the linear strip $|e| < 1/2$ we can use (3.13) and start shooting characteristic trajectories (4.4) from the hyperplane $e = 1/2$. For clarity, we explain the method for finding $\hat{\Psi}$ by considering the simplest case of a noise approximation of order $N = 1$, loop parameters $P = 1/2$, $K = a = 1$, and without thermal noise ($N_{th} = 0$). For these parameters the value of $\Psi(\boldsymbol{x})$ in the linear domain $|e| < 1/2$ is given by

$$\begin{aligned}\Psi(\boldsymbol{x}) = \frac{1}{64}\Bigg\{ & -424\phi_n e + 1408 y_2\beta + 832 y_2 e - 704 y_2\phi_n + 900\beta\frac{\ddot{T}}{T_c} \\ & + 388 e\frac{\ddot{T}}{T_c} + 784 e\beta - 436\phi_n\frac{\ddot{T}}{T_c} - 648\phi_n\beta + 992 y_2\frac{\ddot{T}}{T_c} + 646\beta^2 + 792 y_2^2 \\ & + 170\phi_n^2 + 396 e^2 + 353\left(\frac{\ddot{T}}{T_c}\right)^2\Bigg\}.\end{aligned} \tag{5.3}$$

The coordinate of the point on the hyperplane $e = 1/2$, where the shooting begins, is denoted by $\boldsymbol{v}_0$. It is chosen as the coordinate $(\beta_0, \boldsymbol{y}_0, \phi_{n,0})$ of a point on the unique stable characteristic trajectory of (5.2). We write (5.2) in the linear domain $e \geq 1/2$ as the linear system

$$\dot{\boldsymbol{u}} = \boldsymbol{\Lambda}\boldsymbol{u}, \tag{5.4}$$

where $\boldsymbol{\Lambda}$ is diagonal with the eigenvalues of $\boldsymbol{M}$ on its diagonal. We can write

$$\boldsymbol{v} = \boldsymbol{P}\boldsymbol{u}, \quad \boldsymbol{\Lambda} = \boldsymbol{P}^{-1}\boldsymbol{M}\boldsymbol{P}. \tag{5.5}$$

The columns of $\boldsymbol{P}$ are the eigenvectors of the matrix $\boldsymbol{M}$ with respect to the eigenvalues on the diagonal $\boldsymbol{\Lambda}$. The matrix $\boldsymbol{M}$ has only one negative eigenvalue,

$$\lambda_1 = -\frac{1}{2} - \frac{\sqrt{5}}{2}, \tag{5.6}$$

and thus only the eigenvector corresponding to that eigenvalue leads to a stable solution of (5.4). Assuming that the negative eigenvalue is the first element in $\boldsymbol{\Lambda}$, we need only take the first element in $\boldsymbol{u}$, replacing the others by zeros.

The initial values are defined by

$$p_0 = \frac{\partial \Psi(1/2,\, \beta_0, \boldsymbol{y}_0, \phi_{n,0})}{\partial e}, \quad q_0 = \frac{\partial \Psi(1/2,\, \beta_0, \boldsymbol{y}_0, \phi_{n,0})}{\partial z}, \tag{5.7}$$

$$r_0 = \frac{\partial \Psi(1/2,\, \beta_0, \boldsymbol{y}_0, \phi_{n,0})}{\partial r}, \quad \boldsymbol{\alpha}_{i,0} = \frac{\partial \Psi(1/2,\, \beta_0, \boldsymbol{y}_0, \phi_{n,0})}{\partial y_{2i}} \quad \text{for all } 1 \le i \le N.$$

Using (5.3), (5.6), and (5.7), we get

$$p_0 = -\left(\frac{11}{8} + \frac{7}{8}\sqrt{5}\right) u_{1,0}, \quad q_0 = \left(\frac{3}{2} + \frac{\sqrt{5}}{4}\right) u_{1,0},$$

$$r_0 = -\left(\frac{7}{8} + \frac{\sqrt{5}}{8}\right) u_{1,0}, \quad \alpha_{1,0} = u_{1,0}. \tag{5.8}$$

It follows from (5.4) that $u_1 = u_{1,0} \exp\{\lambda_1 \tilde{t}\}$. Now, we can solve for the initial conditions $(\beta_0, y_0, \phi_{n,0})$ from equations (5.3), (5.7), and (5.8). Having found the initial conditions, we can proceed to integrate (4.7) to find the minimal value $\hat{\Psi}$ on ∂D,

$$\hat{\Psi} = \Psi(1/2,\, \beta_0, \boldsymbol{y}_0, \phi_{n,0}) + 4 \int_0^\infty \left\{ \sum_{l=1}^{N} (N - l + 1)\alpha_l(t) + (N+1) r(t) \right\}^2 dt, \tag{5.9}$$

where $\alpha_l(t)$ and $r(t)$ are calculated on the characteristic that starts at $(1/2,\, \beta_0,\, \boldsymbol{y}_0,\, \phi_{n,0})$. For the case of zero acceleration between transmitter and receiver, i.e., $\ddot{T}/T_c = 0$, we get $\hat{\Psi} = 0.9120$. The analogous computation for Nth order approximations can be done by solving the Lyapunov equation numerically or symbolically (e.g., with Maple or Mathematica), finding the negative eigenvalue, and determining the matrix $\boldsymbol{P}$. This was done with the results in Table 1. As can be seen in Table 1, the minimum value $\hat{\Psi}$ changes very slightly for $N \ge 5$. Thus, approximation of order $N = 5$ for the noise is sufficient for the calculation of the MTLL of 100–1000 seconds. This range of values of the MTLL is chosen because for MTLL less than 100 seconds the leading order approximation is insufficient. The problem of optimizing the MTLL is most critical at low CNR, where the majority of losses of lock occur and the MTLL is still below 1000 seconds. The MTLL increases exponentially with the CNR, so higher accuracy of $\hat{\Psi}$ is needed. Long MTLLs are of less interest in the optimization process. The range of validity of the leading order approximation is for values of the CNR

TABLE 1
$\hat{\Psi}$ *and MTLL for different* N.

N	$\hat{\Psi}$	$MTLL$	$MTLL$	$MTLL$
		$cnr = 1, \rho = 0.25$	$cnr = 1.5, \rho = 0.1667$	$cnr = 2.5, \rho = 0.1$
1	0.91204	77	476	18280
2	0.90916	76	468	17761
3	0.91027	76	471	17959
4	0.91077	76	472	18049
5	0.91094	76	473	18080
6	0.91102	76	473	18094
8	0.91108	77	473	18105
10	0.91110	77	473	18109
20	0.91111	77	473	18111
30	0.91111	77	473	18111

that result in $\rho \ll \hat{\Psi}$. We have disregarded the preexponential factor in (3.8) because the main contribution to the MTLL comes from the exponential term. Furthermore, since we assume a constant prefactor, the results of the simulations might be slightly displaced from the theoretical line. The prefactor can be resolved by simulations for small MTLL and then applied to large MTLL, where simulations are impractical. Our derivation results in a single quantity in an asymptotic expansion. This result is the exponential growth rate of MTLL as the dimensionless noise strength ρ tends to zero.

This can be understood as follows. The loop's noise equivalent bandwidth is 1/2Hz, which is much smaller than the region of validity of the truncated continued fraction approximation to $1/f$. Furthermore, for MTLL of the order of 100–1000 seconds, the corresponding frequency range is $10^{-3}\text{Hz} \leq f \leq 10^{-2}\text{Hz}$. Since the region of validity of the approximation for $N = 5$ is in the range $10^{-3}\text{Hz} \leq f \leq 10\text{Hz}$ (Figure 3), it is understandable that using the approximation for the $1/f$ noise with $N = 5$ results in an accurate value of $\hat{\Psi}$, which in turn accurately approximates the exponential growth rate of the MTLL.

In general one would expect that as more phase noise enters the DLL, the MTLL will become smaller. In Figure 3 we see that as the approximation order becomes larger, more energy enters at very low and very high frequencies. However, we see in Table 1 that $\hat{\Psi}$ actually increases monotonically for $N \geq 2$. The answer to this paradox can be solved by taking a closer look at the transfer function of our $1/f$ approximation. Before flattening out below a certain low frequency, the transfer function displays a "knee" that is above the $1/f$ curve. Further, for a specific MTLL only frequencies larger than 1/MTLL need be considered and only frequencies smaller than the loop's bandwidth should be accounted for. In this frequency band the noise entering the loop actually decreases as the phase noise approximation order is increased, since the "knee" moves to lower frequencies that are irrelevant to the problem at hand. For example, let us consider an MTLL of 100 seconds. The frequency band in question is $0.01 \leq f \leq 0.5$. In Table 2 the energy in the $0.01 \leq f \leq 0.5$ frequency band is given for different approximation orders N. Since for $N < 6$ the "knee" is above $f = 0.01$Hz, the decrease in energy for increasing N explains the increasing $\hat{\Psi}$. From Table 1 we learn that for $N \geq 6$ the changes in $\hat{\Psi}$ result in an insignificant rise in MTLL. In fact, the difference between MTLL for CNR that gives MTLL of 100 seconds for approximation order $N = 6$ and MTLL given for approximation order $N = 30$ with the same CNR is less than 0.1%.

Next we present Monte-Carlo results of the MTLL for approximation of order $N = 5$ compared to the analytic MTLL calculated above for that order of approximation

TABLE 2
Energy in $0.01\text{Hz} \leq f \leq 0.5\text{Hz}$ *frequency band for approximation orders* N.

N	Energy
1	4.7995
2	4.3968
3	4.1909
4	4.0139
5	3.9280
6	3.9050
7	3.9048
8	3.9086
9	3.9112
10	3.9121

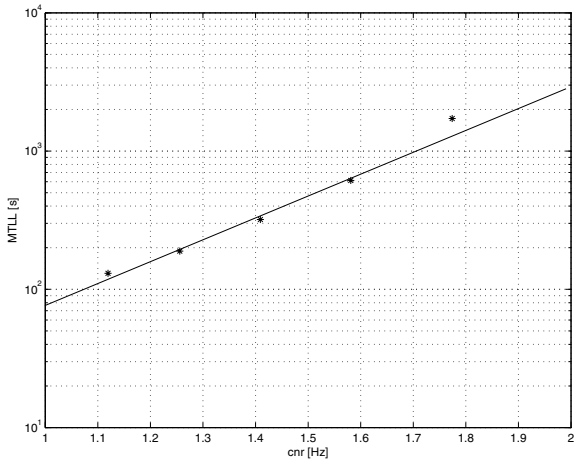


FIG. 4. *The MTLL of a second order DLL under the influence of* $1/f^3$ *noise approximation of order* $N = 5$. *The loop parameters were taken as* $P = 1/2, a = K = 1$. *The solid line is the derived analytic leading order of the MTLL, and the asterisks denote the Monte-Carlo results (each asterisk represents the mean result of* 50 *trials).*

(see Figure 4). Furthermore, Monte-Carlo results of the MTLL under the influence of exact $1/f$ discrete noise generated according to [20] are presented along with the results of the MTLL under the influence of $1/f$ noise approximation with $N = 20$ in Figure 5.

Figures 4 and 5 show that the analytic calculation of the leading order term of the MTLL results in a model that fits well the Monte-Carlo results. Furthermore, the similarity of the analytic calculation of the leading order term of the MTLL to those calculated via Monte-Carlo trials under the influence of exact discrete $1/f$ proves that the truncation of our model with the appropriate N provides very accurate results for the calculation of the MTLL for the second order DLL. The dependence of the loop's parameters on the power P can be eliminated by a proper AGC (automatic gain control) loop.

6. Discussion and conclusions. First, we apply the results to the optimization of the loop's parameters by finding the values of a and K that yield the maximum $\hat{\Psi}$. In our case it turns out that the best result is obtained in the limit $K \to \infty$. This can be easily derived analytically and is quite predictable from (2.14). In a real system

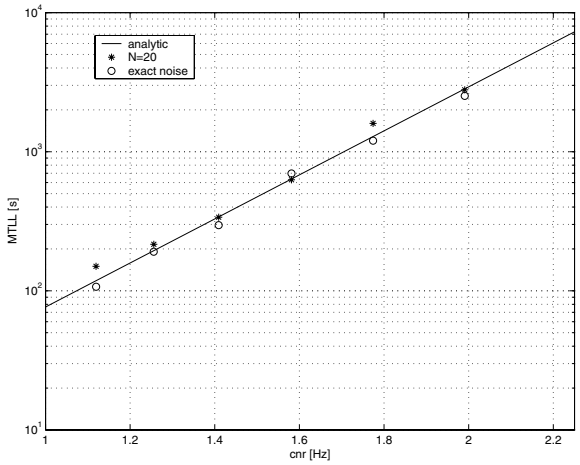


FIG. 5. *The MTLL of a second order DLL under the influence of* $1/f^3$ *noise approximation of order* $N = 20$ *(denoted by asterisks—mean of* 50 *trials each) along with the MTLL under the influence of exact discrete noise (denoted by pluses—mean of* 50 *trials each). The loop parameters were taken as* $P = 1/2, a = K = 1$. *The solid line is the derived analytic leading order of the MTLL.*

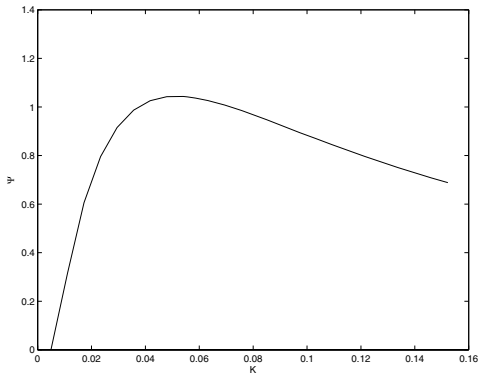


FIG. 6. *Optimizing for loop filter parameter* K *for* $a = 100$.

with additive channel thermal noise, increasing K will increase the thermal noise entering the loop, thus limiting the profitability of increasing K. We also analyzed the case with additive thermal noise by choosing $N_{th} = \frac{N_{ph}}{10}$, $P = \frac{1}{2}$, and $\ddot{T} = 0.5$. It is clear from (2.14) that for the system to remain stable the condition $aK \geq \frac{\ddot{T}}{\sqrt{2P}T_c}$ has to be satisfied, which simplifies in our case to $aK \geq \frac{1}{2}$. It turns out that for every a there is a $K_{opt}(a)$ that maximizes $\hat{\Psi}$. We found that $\hat{\Psi}$ increases monotonically as a increases, but only to a very slight extent, e.g., beyond $a = 100$. In a real system the loop filter coefficients cannot be chosen arbitrarily large, and thus a has to be chosen as large as possible for any given realizable K. Finally, a graph showing $\hat{\Psi}$ for $a = 100$ is given in Figure 6.

In summary, we give here the first analytic approach to the determination of the effects of $1/f^3$ noise on the leading order term for the MTLL of a second order DLL. The analytic derivation of the leading term in a small noise expansion of the MTLL makes it possible to find the loop parameters that maximize the leading order term

of the MTLL. The concept outlined in this paper can be further applied to other tracking loops, such as the commonly used PLL. It sets the foundation for a more complete analysis, to be explored in a future investigation of a comprehensive model that includes additional elements of phase noise ($1/f^2$, $1/f$, and flat segment).

REFERENCES

[1] H. Van Trees, *Detection, Estimation, and Modulation Theory*, Parts I, II, Wiley, New York, 1970.

[2] W. C. Lindsey, *Synchronous Systems in Communications and Control*, Prentice–Hall, Englewood Cliffs, NJ, 1972.

[3] R. Ziemer and R. Peterson, *Introduction to Digital Communication*, Prentice–Hall, Englewood Cliffs, NJ, 2001.

[4] A. J. Viterbi, *Principles of Coherent Communications*, McGraw–Hill, New York, 1966.

[5] E. D. Kaplan, ed., *Understanding GPS: Principles and Applications*, Artech House, Norwood, MA, 1996.

[6] G. W. Stimson, *Introduction to Airborne Radar*, SciTech, Mendham, NJ, 1998.

[7] J. K. Holmes, *Coherent Spread Spectrum Systems*, Wiley, New York, 1982.

[8] M. K. Simon, J. Omura, R. A. Scholtz, and B. Levitt, *Spread Spectrum Communications*, Vols. I–III, Computer Science Press, Rockville, MD, 1985.

[9] 256Mbit GDDR3 SDRAM, Technical specification, Samsung Electronics, Seoul, Korea, 2004.

[10] T. E. Lee and A. Hajimiri, *Oscillator phase noise: A tutorial*, IEEE J. Solid-State Circuits, 35 (2000), pp. 326–336.

[11] B. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, San Francisco, 1982.

[12] M. S. Keshner, *$1/f$ noise*, Proc. IEEE, 70 (1982), pp. 212–218.

[13] B. B. Mandelbrot and J. W. Van Ness, *Fractional Brownian motions, fractional noises and applications*, SIAM Rev., 10 (1968), pp. 422–437.

[14] A. Demir, A. Mehrotra, and J. Roychowdhury, *Phase noise in oscillators: A unifying theory and numerical methods for characterization*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 47 (2000), pp. 655–674.

[15] J. A. Barnes and D. W. Allan, *A statistical model of flicker noise*, Proc. IEEE, 54 (1966), pp. 176–178.

[16] A. van der Ziel, *On the noise spectra of semiconductor noise and of flicker effect*, Physica, 16 (1950), pp. 359–372.

[17] G. W. Wornell, *A Karhunen-Loéve-like expansion for $1/f$ processes via wavelets*, IEEE Trans. Inform. Theory, 36 (1990), pp. 859–861.

[18] G. W. Wornell, *Synthesis, Analysis, and Processing of Fractal Signals*, RLE Technical report 566, Massachusetts Institute of Technology, Cambridge, MA, 1991.

[19] G. W. Wornell and A. V. Oppenheim, *Estimation of fractal signals from noisy measurements using wavelets*, IEEE Trans. Signal Process., 40 (1992), pp. 611–623.

[20] N. J. Kasdin, *Discrete simulation of colored noise and stochastic processes and a $1/f$ power law noise generation*, Proc. IEEE, 83 (1995), pp. 802–827.

[21] T. E. Duncan and B. Pasik-Duncan, *Fractional Brownian motion and stochastic equations in Hilbert spaces*, Stoch. Dyn., 2 (2002), pp. 225–250.

[22] D. Feyel and A. D. L. Pradelle, *On fractional Brownian processes*, Potential Anal., 10 (1999), pp. 273–288.

[23] E. Milotti, *Linear processes that produce $1/f$ or flicker noise*, Phys. Rev. E (3), 51 (1995), pp. 3087–3103.

[24] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*, Wiley, New York, 1972.

[25] A. L. Welti and B. Z. Bobrovsky, *Mean time to lose lock for a coherent second-order PN-code tracking loop—The singular perturbation approach*, IEEE J. Select. Areas Commun., 8 (1990), pp. 809–818.

[26] Z. Schuss, *Theory and Applications of Stochastic Differential Equations*, Wiley, New York, 1980.

[27] P. Hänngi, P. Talkner, and M. Borkovec, *Reaction-rate theory: Fifty years after Kramers*, Rev. Modern Phys., 62 (1990), pp. 251–341.

[28] M. M. Dygas, B. J. Matkowsky, and Z. Schuss, *A singular perturbation approach to non-Markovian escape rate problems with state dependent friction*, J. Chem. Phys., 84 (1986), pp. 3731–3738.

[29] R. Ziemer and R. Peterson, *Digital Communications and Spread Spectrum Systems*, Macmillan, New York, 1985.

[30] R. Dorf and R. H. Bishop, *Modern Control Systems*, 10th ed., Prentice–Hall, Englewood Cliffs, NJ, 2004.
[31] J. G. Proakis, *Probability and stochastic processes*, in Digital Communications, 3rd ed., McGraw–Hill, New York, 1995, pp. 68–72.
[32] Z. Schuss and B. J. Matkowsky, *The exit problem: A new approach to diffusion across potential barriers*, SIAM J. Appl. Math., 36 (1979), pp. 604–623.
[33] D. Ludwig, *Persistence of dynamical systems under random perturbations*, SIAM Rev., 17 (1975), pp. 605–640.
[34] T. Naeh, M. M. Kłosek, B. J. Matkowsky, and Z. Schuss, *A direct approach to the exit problem*, SIAM J. Appl. Math., 50 (1990), pp. 595–627.
[35] D. Ryter and H. Meyr, *Theory of phase tracking systems of arbitrary order: Statistics of cycle slips and probability distribution of the state vector*, IEEE Trans. Inform. Theory, 24 (1978), pp. 1–7.
[36] D. Ryter and P. Jordan, *A way to solve the stationary Fokker-Planck equation for metastable systems*, Phys. Lett. A, 104 (1984), pp. 193–195.
[37] R. S. Maier and D. L. Stein, *Limiting exit location distributions in the stochastic exit problem*, SIAM J. Appl. Math., 57 (1997), pp. 752–790.
[38] R. Courant and D. Hilbert, *Methods of Mathematical Physics. Vol.* II. *Partial Differential Equations*, Wiley-Interscience, New York, 1989.
[39] I. Sneddon, *Elements of Partial Differential Equations*, McGraw–Hill International Editions (Mathematics Series), McGraw–Hill, New York, 1985.

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1209–1226

MOMENTS OF THE INVERSE SCATTERING OPERATOR OF THE BOLTZMANN EQUATION: THEORY AND APPLICATIONS*

S. C. BRUGGER†, A. SCHENK†, AND W. FICHTNER†

Abstract. In this paper useful physical objects called moments of the inverse scattering operator (MISO) of the Boltzmann equation (BE) are studied. The existence and uniqueness of the MISO is proven and a simple, generally valid, iterative scheme to actually compute those objects is given. The applications of the MISO extend from the computation of the solution for the space-homogeneous BE for small electric and magnetic fields to the exact computation of any transport parameter. This can be done for all moments of the space-inhomogeneous BE and for arbitrary electric and magnetic field intensities. The concept of MISO offers an elegant way to avoid the relaxation time approximation (RTA) every time it comes into play, not only theoretically but also in practical computations.

Key words. Boltzmann equation, transport model, beyond the relaxation time approximation, inverse scattering operator, semiconductor

AMS subject classifications. 82D37, 82C70, 82C05, 82C80

DOI. 10.1137/050633275

1. Introduction.

In contemporary semiconductor device simulation two approaches are common: For industrial application the so-called transport models (TMs) are widely used to compute terminal currents and to perform small signal and noise analysis. On the research side, one prefers, when possible, to directly solve the space-inhomogeneous Boltzmann equation (BE) either using Monte Carlo (MC) methods or directly. The two methods often lead to different results. When trying to understand the origin of these discrepancies, one is immediately confronted with the relaxation time approximation (RTA). The RTA has been used to derive the TMs from the BE as a standard method mainly due to the lack of alternative ways to treat the problem. In the case of device simulation, it is impossible to validate the correctness of the RTA because there is no general method for directly comparing it with exact solutions.

The aim of this paper is to develop a method that avoids the RTA. It will allow, among others, the exact computation of transport parameters and noise sources for any moment of the BE using the MC method. The key feature in this formalism is the derivation and computation of moments of the inverse scattering operator (MISO). Although the formalism was primarily developed to be applied to semiconductors, it can also be used to analyze any open system described by a BE (linear or not).

The paper is organized in four sections. First, we recall the approximations in deriving a TM and point out how the knowledge of the MISO enables us to locally compare *term by term* the TM with the outcome of an MC simulation, i.e., with the solution of the BE. In section 3, we will outline the sufficient mathematical conditions under which the MISO exist, and we will present a natural way to compute them. In section 4, five useful applications are presented. First, the MISO are used to compute the solution of the space-homogeneous BE for small electric and magnetic fields. Then, transport parameters, Hall factors, and relaxation times (RTs) are computed in a very general way. At the end of the section, an exact expression is derived for

*Received by the editors June 8, 2005; accepted for publication (in revised form) January 9, 2006; published electronically March 24, 2006.
http://www.siam.org/journals/siap/66-4/63327.html
†Institut für Integrierte Systeme, ETH Zürich, CH-8092 Zürich, Switzerland (brugger@iis.ee.ethz.ch, schenk@iis.ee.ethz.ch, fw@iis.ee.ethz.ch).

the correlation functions of the Langevin noise sources for arbitrary moments of the space-homogeneous nondegenerate BE. Finally, section 5 gives conclusions and a brief outlook on future work.

2. Direct comparison of TMs with the BE.

2.1. Derivation of TMs from the BE. The first step in the derivation of any TM from the BE

$$\partial_t f + \vec{v}\cdot\nabla_{\mathbf{r}} f - \frac{q}{\hbar}\vec{E}\cdot\nabla_{\mathbf{k}} f = Sf \tag{2.1}$$

is to introduce some kind of RT τ, which ideally should contain all information about the scattering operator (SO) S:

$$\partial_t f + \vec{v}\cdot\nabla_{\mathbf{r}} f - \frac{q}{\hbar}\vec{E}\cdot\nabla_{\mathbf{k}} f = \frac{f - \frac{n}{n_{eq}} f_{eq}}{\tau(f,\vec{r},t,\vec{k},\vec{E})}. \tag{2.2}$$

The function τ may depend on moments of the distribution function f, the position in real space $\vec{r}$, the momentum $\vec{k}$, the time t, and the electric field $\vec{E}$. Multiplying both sides of (2.2) by τ results in

$$\begin{aligned}\tau(f,\vec{r},t,\vec{k},\vec{E})\partial_t f + \tau(f,\vec{r},t,\vec{k},\vec{E})\vec{v}\cdot\nabla_{\mathbf{r}} f \\ -\tau(f,\vec{r},t,\vec{k},\vec{E})\frac{q}{\hbar}\vec{E}\cdot\nabla_{\mathbf{k}} f = f - \frac{n}{n_{eq}} f_{eq}.\end{aligned} \tag{2.3}$$

In (2.1)–(2.3), f_{eq} denotes the Boltzmann distribution function normalized to one ($\int_{Bz} f_{eq}^2(\vec{k})\mathrm{d}^3k = 1$) and $n := \int_{Bz} f(\vec{k})\mathrm{d}^3k$, $n_{eq} := \int_{Bz} f_{eq}(\vec{k})\mathrm{d}^3k$, with Bz the Brillouin zone. The symbol n_{eq} is used here for convenience, although the actual equilibrium density differs from n_{eq} by a constant due to the normalization condition.

To be able to compare (2.2) and (2.3) with the exact BE (2.1), one has to invert the SO (2.4):

$$\begin{aligned}\int S^{-1}(\vec{k},\vec{k}')\partial_t f(\vec{k}')d^3k' + \int S^{-1}(\vec{k},\vec{k}')\vec{v}(\vec{k}')\cdot\nabla_{\mathbf{r}} f(\vec{k}')d^3k' \\ - \int S^{-1}(\vec{k},\vec{k}')\frac{q}{\hbar}\vec{E}\cdot\nabla_{\mathbf{k}'} f(\vec{k}')d^3k' = f(\vec{k}) - \frac{n}{n_{eq}} f_{eq}(\vec{k}).\end{aligned} \tag{2.4}$$

Note that $S^{-1}\circ S \neq \mathbb{1}$! This crucial statement will be explained in detail in section 3.

RTs are introduced because the original BE (2.1) causes a major problem when one wants to derive a transport equation: The distribution function f does not appear isolated in the equation. The only term in which f is isolated is $\partial_t f$, which is the partial derivative of f by the time, which is not an equation for f. By introducing an RT τ, f appears isolated on the rhs of the equation. This is the main reason why RTs are introduced. To do the same in an exact way, the simplest possibility is to reverse the SO. This is the only possibility for expressing in an exact mathematical way what the RTA tries to achieve on a heuristic basis. In (2.2), the term $-\frac{n}{n_{eq}} f_{eq}$ was introduced phenomenologically to express the fact that at thermodynamic equilibrium the collision term disappears. The mathematical reason is, however, that the SO S has an eigenvector with eigenvalue 0, which is the equilibrium distribution f_{eq} in the case of Boltzmann statistics. Therefore, (2.4) is the rigorous mathematical formulation of what was done approximately and heuristically by introducing an RT.

The second step is to build a moment of interest of (2.3) by multiplying (2.3) by a function $g(\vec{k})$ and integrating over the momentum space,

$$\begin{aligned}(2.5)\quad &\int_{Bz}\Big[g(\vec{k})\tau(f(\vec{k},\vec{r}))\partial_t f(\vec{k},\vec{r})\Big]\mathrm{d}^3k+\int_{Bz}\Big[g(\vec{k})\tau(f(\vec{k},\vec{r}))\vec{v}(\vec{k})\cdot\nabla_{\mathbf{r}}f(\vec{k},\vec{r})\Big]\mathrm{d}^3k\\ &-\int_{Bz}\Big[g(\vec{k})\tau(f(\vec{k},\vec{r}))\frac{q}{\hbar}\vec{E}\cdot\nabla_{\mathbf{k}}f(\vec{k},\vec{r})\Big]\mathrm{d}^3k=\int_{Bz}g(\vec{k})\left[f(\vec{k},\vec{r})-\frac{n}{n_{eq}}f_{eq}(\vec{k},\vec{r})\right]\mathrm{d}^3k,\end{aligned}$$

where $g(\vec{k})$ is a function in which we are interested. $g(\vec{k})$ could be, e.g., a power of the velocity.

Building moments in a similar fashion with (2.4) leads to

$$\begin{aligned}(2.6)\quad &\int_{Bz}\int_{Bz}\Big[g(\vec{k})S^{-1}(\vec{k},\vec{k}_1,\vec{r})\partial_t f(\vec{k}_1,\vec{r})\Big]\mathrm{d}^3k_1\mathrm{d}^3k\\ &\qquad+\int_{Bz}\int_{Bz}\Big[g(\vec{k})S^{-1}(\vec{k},\vec{k}_1,\vec{r})\vec{v}(\vec{k}_1)\cdot\nabla_{\mathbf{r}}f(\vec{k}_1,\vec{r})\Big]\mathrm{d}^3k_1\mathrm{d}^3k\\ &\qquad-\int_{Bz}\int_{Bz}\Big[g(\vec{k})S^{-1}(\vec{k},\vec{k}_1,\vec{r})\frac{q}{\hbar}\vec{E}\cdot\nabla_{\mathbf{k_1}}f(\vec{k}_1,\vec{r})\Big]\mathrm{d}^3k_1\mathrm{d}^3k\\ &:=\int_{Bz}S_g^{-1}(\vec{k}_1,\vec{r})\partial_t f(\vec{k}_1,\vec{r})\mathrm{d}^3k_1+\int_{Bz}S_g^{-1}(\vec{k}_1,\vec{r})\vec{v}(\vec{k}_1)\cdot\nabla_{\mathbf{r}}f(\vec{k}_1,\vec{r})\mathrm{d}^3k_1\\ &-\int_{Bz}S_g^{-1}(\vec{k}_1,\vec{r})\frac{q}{\hbar}\vec{E}\cdot\nabla_{\mathbf{k_1}}f(\vec{k}_1,\vec{r})\mathrm{d}^3k_1=\int_{Bz}g(\vec{k})\left[f(\vec{k},\vec{r})-\frac{n}{n_{eq}}f_{eq}(\vec{k},\vec{r})\right]\mathrm{d}^3k,\end{aligned}$$

where

$$(2.7)\qquad S_g^{-1}(\vec{k}_1,\vec{r})=\int_{Bz}g(\vec{k})S^{-1}(\vec{k},\vec{k}_1,\vec{r})\mathrm{d}^3k$$

is a moment of the inverse scattering operator (ISO) in (2.5).

The third step is to perform a partial integration of the **k**-gradient term in (2.5) and to neglect the boundary term

$$(2.8)\qquad \int_{Bz}g(\vec{k})\tau(f(\vec{k},\vec{r}))\frac{q}{\hbar}\vec{E}\cdot\nabla_{\mathbf{k}}f(\vec{k},\vec{r})\mathrm{d}^3k\simeq-\frac{q}{\hbar}\vec{E}\cdot\int_{Bz}\nabla_{\mathbf{k}}(g(\vec{k})\tau(f(\vec{k},\vec{r})))f(\vec{k},\vec{r})\mathrm{d}^3k.$$

Because the function $S_g^{-1}(\vec{k}_1,\vec{r})$ can be discontinuous in some points (for some silicon MC models, this happens on the boundary surfaces between two valleys), we cannot exactly transform the **k**-gradient term of (2.6) as in (2.8). Instead we have to decompose the domain Bz into subdomains, where the function $S_g^{-1}(\vec{k}_1,\vec{r})$ is continuous, and keep all the boundary terms, which in general will not disappear, as follows:

$$\begin{aligned}(2.9)\quad &\sum_i\int_{Bz_i}S_g^{-1}(\vec{k},\vec{r})\frac{q}{\hbar}\vec{E}\cdot\nabla_{\mathbf{k}}f(\vec{k},\vec{r})\mathrm{d}^3k\\ &=-\frac{q}{\hbar}\vec{E}\cdot\sum_i\int_{Bz_i}\nabla_{\mathbf{k}}(S_g^{-1}(\vec{k},\vec{r}))f(\vec{k},\vec{r})\mathrm{d}^3k+\frac{q}{\hbar}\vec{E}\cdot\sum_i\oint_{\partial Bz_i}S_g^{-1}(\vec{k},\vec{r})f(\vec{k},\vec{r})\vec{n}\mathrm{da},\end{aligned}$$

where $\bigcup_i Bz_i=Bz$.

The boundary term $\sum_i\oint_{\partial Bz_i}S_g^{-1}(\vec{k},\vec{r})f(\vec{k},\vec{r})\vec{n}\mathrm{d}a$ does not vanish. To our knowledge it has never been investigated numerically whether this term is negligible for all

moments and for all field strengths of practical interest, e.g., for bulk silicon. We will therefore keep this term in our numerical analysis.

To derive common TMs, only a small number of functions g is relevant. For the drift-diffusion (DD) model, only the moment corresponding to $g_1 = \vec{v}$ is needed. The hydrodynamic model is obtained by considering the moments $g_1 = \vec{v}$, $g_2 = \epsilon$ (where $\epsilon(\vec{k})$ is the energy) or $||v||^2$ and $g_3 = \tau g_2 \vec{v}$. Finally, for the so-called six moments method (see [6]), the additional moments $g_4 = \epsilon^2$ and $g_5 = \tau g_4 \vec{v}$ are considered. These moments equations, together with the Poisson equation and the current continuity equation (the contraction of the BE with the constant function 1), constitute a TM that should approximate the solution to the BE coupled with the Poisson equation. These equations are usually parametrized using the electrostatic potential, the particle density(ies), and the mean value of g_2 and g_4. To close this system, further steps are needed: The g_3 (resp., g_5) moment is inserted into the $\nabla_{\mathbf{r}}$ term of the g_2 (resp., g_4) moment and approximations are done, such as the use of the Einstein relation, the replacement of tensorial transport coefficients by scalars, and the use of a closure relation for the last moment (see, e.g., [13], [3], [1], [12]). Finally, all expressions still unknown are called transport coefficients and computed using a model (analytical and/or bulk MC simulation) and parametrized by the functions listed above.

2.2. Transport parameters. In order to illustrate potential applications of the outlined concept, we directly compare the moments of well-known macroscopic TMs, such as the DD model and the energy-balance (EB) model (see, e.g., [12, Chap. 1.1.3]), with the corresponding terms from the BE containing the exact ISO. The equation for the current in the DD model reads

$$\tau_p \partial_t(\langle \vec{v} \rangle) - \mu n \vec{E} - D \nabla_{\mathbf{r}} n = \langle \vec{v} \rangle =: -\frac{\vec{J}}{q}, \tag{2.10}$$

where τ_p is the momentum RT, μ the mobility tensor, D the diffusivity tensor, n the density, and $\langle \vec{v} \rangle := \int_{Bz} \vec{v} f \mathrm{d}^3 k$ ($\langle \vec{v} \rangle / n$ is the mean velocity).

If we compare term by term the lhs of (2.10) with the lhs of (2.6), we find the following conditions for the DD model to be exact:

$$\tau_p \partial_t(\langle v_i \rangle) \stackrel{!}{=} \int_{Bz} S_{v_i}^{-1}(\vec{k}_1, \vec{r}) \partial_t f(\vec{k}_1, \vec{r}) \mathrm{d}^3 k_1, \quad i = 1, \ldots, 3, \tag{2.11}$$

$$\left(\mu n \vec{E}\right)_i \stackrel{!}{=} -\frac{q}{\hbar} \vec{E} \cdot \int_{Bz} \nabla_{\mathbf{k}} (S_{v_i}^{-1}(\vec{k}, \vec{r})) f(\vec{k}, \vec{r}) \mathrm{d}^3 k + \frac{q}{\hbar} \vec{E} \cdot \sum_j \oint_{\partial Bz_j} S_{v_i}^{-1}(\vec{k}, \vec{r}) f(\vec{k}, \vec{r}) \vec{n} \mathrm{d}a, \quad i = 1, \ldots, 3, \tag{2.12}$$

$$(-D \nabla_{\mathbf{r}} n)_i \stackrel{!}{=} \int_{Bz} S_{v_i}^{-1}(\vec{k}_1, \vec{r}) \vec{v}(\vec{k}_1) \cdot \nabla_{\mathbf{r}} f(\vec{k}_1, \vec{r}) \mathrm{d}^3 k_1, \quad \mathrm{i} = 1, \ldots, 3, \quad \langle v_i \rangle_{eq} = 0. \tag{2.13}$$

Note that (2.13) gives a natural definition of the mobility tensor

$$(\mu)_{ij} := -\frac{q}{n\hbar} \int_{Bz} \partial_{k_j} (S_{v_i}^{-1}(\vec{k}, \vec{r})) f(\vec{k}, \vec{r}) \mathrm{d}^3 k + \frac{q}{n\hbar} \sum_l \oint_{\partial Bz_l} S_{v_i}^{-1}(\vec{k}, \vec{r}) f(\vec{k}, \vec{r}) (\vec{n} \mathrm{d}a)_j, \quad i = 1, \ldots, 3. \tag{2.14}$$

The corresponding equations for the EB model are

$$\tau_p \partial_t(\langle \vec{v} \rangle) - \mu n \vec{E} - \mu n \nabla_{\mathbf{r}} \left(\frac{k_B}{q} \frac{\langle T \rangle}{n} \right) - D' \nabla_{\mathbf{r}} n = \langle \vec{v} \rangle \,, \tag{2.15}$$

$$\frac{3}{2} k_B \left(\langle T \rangle - n \underbrace{\frac{\langle T \rangle_{eq}}{n_{eq}}}_{:=T_{eq}} \right) = -\tau_E q \langle \vec{v} \rangle \cdot \vec{E} - \tau_E \nabla_{\mathbf{r}} \left[-\kappa_n \nabla_{\mathbf{r}} \left(\frac{\langle T \rangle}{n} \right) + \frac{5 k_B}{2} \frac{\langle T \rangle \langle \vec{v} \rangle}{n} \right], \tag{2.16}$$

where τ_E is the energy RT, κ_n the heat conduction coefficient, $\langle T \rangle_{eq} / n_{eq}$ the temperature in thermodynamic equilibrium, $D' := \frac{k_B}{q} \frac{\langle T \rangle}{n} \mu$, and $\langle T \rangle := \frac{m}{3 k_B} \cdot \mathrm{Tr}(\int_{Bz} (\vec{v} \otimes \vec{v}) f \mathrm{d}^3 k)$. If we compare (2.15) and (2.16) with (2.6), we find the following conditions for the EB model to be exact:

$$-\mu \left(\nabla_{\mathbf{r}} \left(\frac{k_B \langle T \rangle}{q} \right) \right)_i \stackrel{!}{=} \int_{Bz} S^{-1}_{v_i}(\vec{k}_1, \vec{r}) \vec{v}(\vec{k}_1) \cdot \nabla_{\mathbf{r}} f(\vec{k}_1, \vec{r}) \mathrm{d}^3 k_1, \quad i = 1, \ldots, 3, \tag{2.17}$$

$$\begin{aligned} \tau_E q \langle \vec{v} \rangle \cdot \vec{E} \stackrel{!}{=} & -\frac{1}{2} m \sum_{i=1}^{3} \frac{q}{\hbar} \vec{E} \cdot \int_{Bz} \nabla_{\mathbf{k}} (S^{-1}_{v_i v_i}(\vec{k}, \vec{r})) f(\vec{k}, \vec{r}) \mathrm{d}^3 k \\ & + \frac{1}{2} m \sum_{i=1}^{3} \frac{q}{\hbar} \vec{E} \cdot \sum_j \oint_{\partial Bz_j} S^{-1}_{v_i v_i}(\vec{k}, \vec{r}) f(\vec{k}, \vec{r}) \vec{n} \mathrm{d}a, \end{aligned} \tag{2.18}$$

$$-\tau_E \nabla_{\mathbf{r}} \left[-\kappa_n \nabla_{\mathbf{r}} \left(\frac{\langle T \rangle}{n} \right) + \frac{5 k_B}{2} \frac{\langle T \rangle \langle \vec{v} \rangle}{n} \right] \stackrel{!}{=} \frac{1}{2} m \sum_{i=1}^{3} \int_{Bz} S^{-1}_{v_i v_i}(\vec{k}_1, \vec{r}) \vec{v}(\vec{k}_1) \cdot \nabla_{\mathbf{r}} f(\vec{k}_1, \vec{r}) \mathrm{d}^3 k_1. \tag{2.19}$$

Although the terms in the rhs of (2.11)–(2.13) and (2.17)–(2.19) look rather cumbersome, this is not the case. As soon as we know $S_g^{-1}(\vec{k}, \vec{r})$ for all required g, we can easily compute such terms for a given device with an MC simulation and then locally compare with the terms of the TM. It should be noted that by construction the following holds: If the transport coefficients computed by the MC method are reinserted into the TMs as a function of position only (not as a function of the density or the mean energy), then the TMs will exactly reproduce the MC density and the MC current density in the case of the DD and EB models, as well as the MC energy current density in the case of the EB model. Another way to verify this statement is to reinsert electric field, density, and mean energy from the MC solution of the BE into the TMs that contain the exact transport coefficients, and to observe that they indeed solve these equations. Therefore, the restrictions imposed on these TMs to be valid (restrictions arising from the models for the transport coefficients) become obsolete as soon as the exact expressions (2.11)–(2.13) and (2.17)–(2.19) are used.

Based on this motivation for the computation of MISO, we show how to actually compute them in a general way.

3. Existence and computation of the MISO.

3.1. Derivation of an equation for the MISO. We start with the general form of the BE,

$$\partial_t f(\vec{r},t,\vec{k},b) + \dot{\vec{r}}\cdot\nabla_{\mathbf{r}} f(\vec{r},t,\vec{k},b) + \dot{\vec{k}}\cdot\nabla_{\mathbf{k}} f(\vec{r},t,\vec{k},b) \tag{3.1}$$
$$= \sum_{b_0}\int_{V_{b_0}} f(\vec{r},t,\vec{k}_0,b_0)w(\vec{r},t)(\vec{k}_0,b_0|\vec{k},b)\mathrm{d}^3k_0 - \sum_{b_0}\int_{V_{b_0}} f(\vec{r},t,\vec{k},b)w(\vec{r},t)(\vec{k},b|\vec{k}_0,b_0)\mathrm{d}^3k_0,$$

where $\vec{r}$ is the position in space, $\vec{k}$ the position in k-space, b a band-valley index, V_b the k-space of band-valley b, and $w(\vec{r},t)(\vec{k},b|\vec{k}_0,b_0)$ is the scattering rate from point $(\vec{k},b)$ to $(\vec{k}_0,b_0)$ (at time t and space position $\vec{r}$, respectively). Note that the Pauli blocking factors $(1-f)$ are included in the scattering rates w. Since $0<1-f\leq 1$, they will never increase the magnitude of w. This will be of some importance below. In the following we will work under the assumption that the V_b are compact pairwise disjoint subsets of $\mathbb{R}^3$, and $w(\vec{r},t)(\vec{k},b|\vec{k}_0,b_0) : V_b \times V_{b_0} \to \mathbb{R}$ is a continuous function of $\vec{k}$ and $\vec{k}_0$. We define $K := \bigcup_{i=0}^{N} V_{b_i}$.

The scattering operator S is defined as

$$S(\vec{r},t)(\vec{k},b|\vec{k}_0,b_0) := w(\vec{r},t)(\vec{k}_0,b_0|\vec{k},b) - \delta^3(\vec{k}-\vec{k}_0)\delta_{b,b_0}W_{tot}(\vec{r},t)(\vec{k},b), \tag{3.2}$$

with $W_{tot}(\vec{r},t)(\vec{k},b) := \sum_{b'}\int_{V_{b'}} w(\vec{r},t)(\vec{k},b|\vec{k}',b')\,\mathrm{d}^3k' > 0$.

By definition, $w(\vec{r},t)(\vec{k}_0,b_0|\vec{k},b)$ is a bound continuous compact operator on the Banach space $C^0(K)$ with $||\cdot||_\infty$ (see, e.g., [14, p. 70]).

In the remainder, the argument $(\vec{r},t)$ will be omitted, and the Dirac notation sometimes will be used for better readability (e.g., $|f\rangle := f$). We will also sometimes use the "$\circ$" notation

$$(A\circ B)(\vec{k},b|\vec{k}_1,b_1) := \sum_{b_0}\int_{V_{b_0}} A(\vec{k},b|\vec{k}_0,b_0)B(\vec{k}_0,b_0|\vec{k}_1,b_1)\mathrm{d}^3k_0 \tag{3.3}$$

to avoid confusion. Using the Dirac notation, the lhs of (3.1) can be written as

$$S|f\rangle := \sum_{b_0}\int_{V_{b_0}} S(\vec{k},b|\vec{k}_0,b_0)f(\vec{k}_0,b_0)\mathrm{d}^3k_0 \tag{3.4}$$
$$= \sum_{b_0}\int_{V_{b_0}} \Big(f(\vec{k}_0,b_0)w(\vec{k}_0,b_0|\vec{k},b) - f(\vec{k},b)w(\vec{k},b|\vec{k}_0,b_0)\Big)\mathrm{d}^3k_0.$$

Now we want to define an inverse operator H for the SO, i.e., an ISO. First of all, one cannot define the ISO naively as $H\circ S|f\rangle = |f\rangle$ for all $f\in C^0(K)$, because S has an eigenvector with eigenvalue 0 (indeed only one, as we will show later), namely, the Boltzmann function f_{eq}.[1] Therefore, we have to invert the SO on the space $Ker^\perp$ perpendicular to its kernel $Ker := \{\lambda|f_{eq}\rangle \mid \lambda\in\mathbb{R}\}$. To do so, we define explicitly

$$Ker^\perp := \{P_{f_{eq}}|g\rangle \mid g\in C^0(K)\}, \tag{3.5}$$

[1] Also in the case of degenerate systems and/or systems with two-particle scattering (e.g., e-e collisions) the eigenvector with eigenvalue 0 exists and is unique, but it will depend on the solution f of the BE, because the scattering operator S depends on f. This will not impact the validity of our approach.

where

$$P_{f_{eq}} := \mathbb{1} - |f_{eq}\rangle\langle f_{eq}|, \tag{3.6}$$

f_{eq} is chosen such that

$$\langle f_{eq}|f_{eq}\rangle = 1, \tag{3.7}$$

and the scalar product is given by

$$\langle f|g\rangle := \sum_b \int_{V_b} f(\vec{k},b)g(\vec{k},b)\mathrm{d}^3k. \tag{3.8}$$

A trivial but important property of S is

$$S|g\rangle = S \circ P_{f_{eq}}|g\rangle. \tag{3.9}$$

Now, the ISO H, if it exists on $Ker^{\perp}$, can be unequivocally defined on $C^0(K)$ based on the properties it must fulfill:

1. $H \circ S|g\rangle \stackrel{!}{=} |g\rangle$ for all $g \in Ker^{\perp}$,
2. $H|f_1\rangle = 0$,

with $f_1(\vec{k},b) := \frac{1}{\sqrt{\sum_{b'}|V_{b'}|}} = \text{const}$, where $|V_{b'}|$ is the volume of $V_{b'}$. Using (3.9), condition 1 can be rewritten as

$$H \circ S|g\rangle = H \circ S \circ P_{f_{eq}}|g\rangle \stackrel{!}{=} P_{f_{eq}}|g\rangle \;\; \forall g \in C^0(K). \tag{3.10}$$

The appropriateness of condition 2 will now be explained in detail. First, note that $S^T|f_1\rangle = 0$ ($\langle f_1|S = 0$) by definition of W_{tot}. Without condition 2 we could define an infinite number of ISOs, because if H satisfies condition 1, then $H + |v\rangle\langle f_1|$ fulfills the same condition for any $|v\rangle$. Let H^* be an ISO fulfilling condition 1, and $|h^*\rangle := H^*|f_1\rangle$. We can always rewrite H^* as $H^* = H^{\perp} + |h^*\rangle\langle f_1|$, where $H^{\perp} := H^* - |h^*\rangle\langle f_1|$. (Note here that because $H^{\perp}$ fulfills conditions 1 and 2, it is unambiguously defined and, therefore, independent of H^*.) By multiplying (3.1) by $\langle f_1|$ we obtain $\partial_t n + \nabla_{\mathbf{r}}\langle\dot{\vec{r}}\rangle = 0$, which is nothing but the current continuity equation. (In the case of semiconductors, the boundary term $\oint_{\partial V_{b_0}} f\dot{\vec{k}} \cdot \vec{n}\mathrm{d}a$ always disappears due to the inversion symmetry of the V_{b_0}.) Thus, by multiplying (3.1) by H^* and using (3.10), we obtain

$$\begin{aligned} H^*\partial_t|f\rangle + H^*\dot{\vec{r}}\cdot\nabla_{\mathbf{r}}|f\rangle + H^*\dot{\vec{k}}\cdot\nabla_{\mathbf{k}}|f\rangle &= |f\rangle - \langle f|f_{eq}\rangle\,|f_{eq}\rangle \\ &= H^{\perp}\partial_t|f\rangle + H^{\perp}\dot{\vec{r}}\cdot\nabla_{\mathbf{r}}|f\rangle + H^{\perp}\dot{\vec{k}}\cdot\nabla_{\mathbf{k}}|f\rangle. \end{aligned} \tag{3.11}$$

The last equation shows that $H^{\perp}$ already contains the full physical information, and that it is reasonable to define $H := H^{\perp}$, i.e., $H|f_1\rangle = 0$.

We can now write an equation for the ISO:

$$H \circ S|g\rangle \stackrel{!}{=} |g\rangle - |f_{eq}\rangle\,\langle f_{eq}|g\rangle \;\; \forall g, \tag{3.12}$$

and finally write the operator equations for H:

$$H \circ S \stackrel{!}{=} \delta^3(\vec{k} - \vec{k}_0)\delta_{b,b_0} - f_{eq}(\vec{k},b)f_{eq}(\vec{k}_0,b_0) = \mathbb{1} - |f_{eq}\rangle\langle f_{eq}|, \tag{3.13}$$

$$H|f_1\rangle \stackrel{!}{=} 0. \tag{3.14}$$

Next we have to solve for (3.13), (3.14). From (3.2) and (3.13) we obtain

$$H \circ S = \sum_{b_2} \int_{V_{b_2}} H(\vec{k}, b|\vec{k}_2, b_2) w(\vec{k}_0, b_0|\vec{k}_2, b_2) \mathrm{d}^3 k_2 - W_{tot}(\vec{k}_0, b_0) H(\vec{k}, b|\vec{k}_0, b_0) \stackrel{!}{=} \delta^3(\vec{k} - \vec{k}_0)\delta_{b,b_0} - f_{eq}(\vec{k}, b) f_{eq}(\vec{k}_0, b_0). \tag{3.15}$$

By rearranging the terms, this equation can be written as

$$H(\vec{k}, b|\vec{k}_0, b_0) = \sum_{b_2} \int_{V_{b_2}} H(\vec{k}, b|\vec{k}_2, b_2) \frac{w(\vec{k}_0, b_0|\vec{k}_2, b_2)}{W_{tot}(\vec{k}_0, b_0)} \mathrm{d}^3 k_2 - \frac{\delta^3(\vec{k} - \vec{k}_0)\delta_{b,b_0}}{W_{tot}(\vec{k}_0, b_0)} + f_{eq}(\vec{k}, b) \frac{f_{eq}(\vec{k}_0, b_0)}{W_{tot}(\vec{k}_0, b_0)} . \tag{3.16}$$

Now remember that we are interested only in MISO,

$$H_g(\vec{k}_0, b_0) := \sum_{b} \int_{V_b} g(\vec{k}, b) H(\vec{k}, b|\vec{k}_0, b_0) \mathrm{d}^3 k = \langle g|H, \tag{3.17}$$

for which we can rewrite (3.16) as

$$|H_g\rangle := H_g(\vec{k}_0, b_0) = \sum_{b_2} \int_{V_{b_2}} H_g(\vec{k}_2, b_2) \frac{w(\vec{k}_0, b_0|\vec{k}_2, b_2)}{W_{tot}(\vec{k}_0, b_0)} \mathrm{d}^3 k_2 - \left(\frac{g(\vec{k}_0, b_0) - \langle g \rangle_{eq} f_{eq}(\vec{k}_0, b_0)}{W_{tot}(\vec{k}_0, b_0)} \right), \tag{3.18}$$

where $\langle g \rangle_{eq} := \langle f_{eq}|g \rangle$.

If we define

$$A(\vec{k}_0, b_0|\vec{k}_2, b_2) := \frac{w(\vec{k}_0, b_0|\vec{k}_2, b_2)}{W_{tot}(\vec{k}_0, b_0)},$$

$$A^T(\vec{k}_0, b_0|\vec{k}_2, b_2) := \frac{w(\vec{k}_2, b_2|\vec{k}_0, b_0)}{W_{tot}(\vec{k}_2, b_2)},$$

$$|l_g\rangle := l_g(\vec{k}_0, b_0) := \left(\frac{g(\vec{k}_0, b_0) - \langle g \rangle_{eq} f_{eq}(\vec{k}_0, b_0)}{W_{tot}(\vec{k}_0, b_0)} \right), \tag{3.19}$$

we can express (3.18) as

$$|H_g\rangle = A|H_g\rangle - |l_g\rangle \tag{3.20}$$

and (3.14) as

$$\langle H_g|f_1\rangle = 0. \tag{3.21}$$

Note that

$$S(\vec{k}|\vec{k}_0) = (A^T(\vec{k}|\vec{k}_0) - \mathbb{1}) W_{tot}(\vec{k}_0). \tag{3.22}$$

So far we found that if, for a given g, there exists one and only one solution to (3.20) and (3.21), then this solution will have the properties we are looking for.

3.2. Computation of the solution. One could think that by iteratively inserting the lhs of (3.20) into the rhs, we could solve the problem, i.e.,

$$|H_g\rangle = A^n|H_g\rangle - \sum_{k=0}^{n-1} A^k|l_g\rangle. \tag{3.23}$$

However, the way to solve (3.20) is somewhat more involved.

Before we go to the solution we discuss an additional condition we have to impose on A. The term $A(k,b|k_0,b_0)$ is nothing but the probability for a particle to end in (k_0,b_0) after one scattering event, having started in (k,b). $A^m(k,b|k_0,b_0)$ is then the probability of going to (k_0,b_0) after m scattering events. In the following we will work under the assumption that there exists an $M \in \mathbb{N}$ such that $A^M(k,b|k_0,b_0) > 0$ for all $(k,b),(k_0,b_0)$. It means that it is possible, starting from any (k,b), to reach any (k_0,b_0) after M scattering events. Under this assumption, A^M is a strong positive compact operator.

PROPOSITION 3.1. *The first Krein–Rutman theorem* [9] *ensures the existence and uniqueness of a stationary solution in thermodynamic equilibrium* $(\exists! f \in C^0(K) \mid S|f\rangle = 0)$.

Proof. We first prove $r(A^M) = 1$. By construction, $||A|| := \sup_{x\in C^0(K)\setminus\{0\}} \frac{||Ax||_\infty}{||x||_\infty} \leqslant 1$, and $A|f_1\rangle = |f_1\rangle$. Therefore, $r(A) \geqslant 1$, and because by definition $r(A) \leqslant ||A||$, we find $\sqrt[M]{r(A^M)} = r(A) = ||A|| = 1$. The first Krein–Rutman theorem ensures that there is only one strict positive function $u \in C^0(K)$ such that $A^M u = r(A^M)u = u$. Of course this function is nothing but f_1. Now we are interested in $(A^T)^M$. By construction, $(A^T)^M$ is a strong positive compact operator. Using again the Krein–Rutman theorem we find a unique v with $0 < v \in C^0(K)$ such that $(A^T)^M v = r((A^T)^M)v$. Because $u > 0$ and $v > 0$, we find $0 < \langle u|v\rangle = \langle A^M u|v\rangle = \langle u|(A^T)^M v\rangle = r((A^T)^M)\langle u|v\rangle$. Thus, $r((A^T)^M) = 1$ and $(A^T)^M v = v$. Then, because $(A^T)^{M+1} v = A^T v \iff A^T v = c * v$ with c real and $c^M = 1$, $A^T v = v$. Using (3.22) gives $S f_{eq} = 0$ with $f_{eq} := \frac{v}{W_{tot}}$. The function f_{eq} is the only solution. ∎

To solve (3.20), we construct a sub-Banach space $Q \subset C^0(K)$, where the solution is unique. Using two important properties of A,

1. $A|f_1\rangle = |f_1\rangle$ by definition of W_{tot},
2. $A^T|W_{tot} f_{eq}\rangle = |W_{tot} f_{eq}\rangle$ because $S|f_{eq}\rangle = 0$,

we can define

$$P_L := \mathbb{1} - \frac{|f_1\rangle\langle f_{eq} W_{tot}|}{\langle f_1|f_{eq} W_{tot}\rangle}, \tag{3.24}$$

which has the following important properties:

$$P_L^2 = P_L, \tag{3.25}$$

$$P_L \circ A = A \circ P_L = P_L \circ A \circ P_L. \tag{3.26}$$

Using P_L, we define the space $Q := \{P_L x | x \in C^0(K)\}$. Because $W_{tot} f_{eq}$ and f_1 are continuous functions, Q is a Banach space.

PROPOSITION 3.2. *A is a linear compact operator on Q.*

Proof. Let $x \in Q$. By definition of Q and property (3.25), $P_L x = x$. Therefore, using property (3.26), $Ax = A \circ P_L x = P_L \circ Ax \in Q$. ∎

Note that $|l_g\rangle \in Q$ because

$$\langle f_{eq} W_{tot} | l_g \rangle = 0. \tag{3.27}$$

By multiplying (3.20) by P_L we reformulate the problem on Q:

$$P_L |H_g\rangle = P_L \circ A \circ P_L |H_g\rangle - |l_g\rangle. \tag{3.28}$$

Defining $|H_g^\perp\rangle := P_L |H_g\rangle$ we obtain

$$(\mathbb{1} - A)|H_g^\perp\rangle = -|l_g\rangle. \tag{3.29}$$

PROPOSITION 3.3. *$(\mathbb{1} - A^M)$ is invertible on Q, and, therefore, $(\mathbb{1} - A)$ is also invertible on Q.*

Proof. We first prove $||A^M||_\infty < 1$ on Q. Let x be in Q. x is a continuous function and $\langle f_{eq} W_{tot} | x \rangle = 0$ because $\langle f_{eq} W_{tot} | P_L = 0$. By definition, $f_{eq} W_{tot}$ is a strict positive function, and therefore x must have a positive part and a negative part. We define $x_+(k) := x(k)$ if $x(k) \geqslant 0$, $x_+(k) = 0$ else, and $x_-(k) := x(k)$ if $x(k) < 0$, $x_-(k) = 0$ else. By definition, $x(k) = x_+(k) + x_-(k)$ for all $k \in K$. Without loss of generality, $||x||_\infty = ||x_+||_\infty$. Then, remembering that A^M is strictly positive, we have $|A^M x| = |\,|A^M x_+| - |A^M x_-|\,| < |A^M x_+| \leqslant ||x_+||_\infty$. Therefore, $||A^M||_\infty < 1$. It means that $(\mathbb{1} - A^M)$ has an inverse on Q that can be written as a Neumann series: $(\mathbb{1} - A^M)^{-1} = \sum_{i=0}^\infty A^{M^i}$. Rewriting $(\mathbb{1} - A^M)$ as $(\mathbb{1} - A) \circ \sum_{j=0}^{M-1} A^j$ and multiplying by $(\mathbb{1} - A^M)^{-1}$, we find $\mathbb{1} = \sum_{i=0}^\infty A^{M^i} \circ (\mathbb{1} - A) \circ \sum_{j=0}^{M-1} A^j = (\mathbb{1} - A) \circ \sum_{i=0}^\infty A^{M^i} \circ \sum_{j=0}^{M-1} A^j$. Thus, clearly $(\mathbb{1} - A)^{-1} = \sum_{i=0}^\infty A^{M^i} \circ \sum_{j=0}^{M-1} A^j = \sum_{i=0}^\infty A^i$. ∎

Multiplying (3.29) by $(\mathbb{1} - A)^{-1}$ gives the solution we were looking for:

$$|H_g^\perp\rangle = -\sum_{i=0}^\infty A^i |l_g\rangle. \tag{3.30}$$

Thus, problem (3.20) has a unique solution on Q, but infinitely many of the form $H_g^\lambda := H_g^\perp + \lambda |f_1\rangle$, $\lambda \in \mathbb{R}$, on $C^0(K)$. In (3.21) we imposed the condition $\langle H_g | f_1 \rangle = 0$ to obtain a unique solution. The only H_g^λ fulfilling this condition is $H_g^{\lambda_g}$ with

$$\lambda_g := -\langle f_1 | H_g^\perp \rangle, \tag{3.31}$$

which is the unique solution we are looking for.

Equation (3.30) represents an iterative method for computing $H_g^\perp$, i.e., for finding an exact solution of (3.13), (3.14) for any $g \in C^0(K)$.

As $\langle H_g | = \langle g | H$ exists for all $g \in C^0(K)$, H also exists and is unique. Thus, (3.13) together with (3.14) also has a unique solution.

3.3. Connection between H and S^{-1}. We show the connection between H and S^{-1}, as well as between H_g and S_g^{-1}. If we let the operator H act on both sides of (3.1), we find with (3.13)

$$H \partial_t |f\rangle + H \dot{\vec{r}} \cdot \nabla_{\mathbf{r}} |f\rangle + H \dot{\vec{k}} \cdot \nabla_{\mathbf{k}} |f\rangle = |f\rangle - \langle f | f_{eq} \rangle \, |f_{eq}\rangle. \tag{3.32}$$

We want to dispose of the term $\langle f | f_{eq} \rangle$ and replace it with a term containing the density. By computing the 0th moment of (3.32), we obtain

$$\begin{aligned} &\langle H_1 | \partial_t | f \rangle + \langle H_1 | \dot{\vec{r}} \cdot \nabla_{\mathbf{r}} | f \rangle + \langle H_1 | \dot{\vec{k}} \cdot \nabla_{\mathbf{k}} | f \rangle = n - \langle f | f_{eq} \rangle \, n_{eq} \\ &\quad \Leftrightarrow \langle f | f_{eq} \rangle = \frac{n}{n_{eq}} - \frac{1}{n_{eq}} \langle H_1 | \partial_t | f \rangle - \frac{1}{n_{eq}} \langle H_1 | \dot{\vec{r}} \cdot \nabla_{\mathbf{r}} | f \rangle - \frac{1}{n_{eq}} \langle H_1 | \dot{\vec{k}} \cdot \nabla_{\mathbf{k}} | f \rangle, \end{aligned} \tag{3.33}$$

where $H_1(\vec{k}_0, b_0) := \sum_b \int_{V_b} H(\vec{k}, b|\vec{k}_0, b_0)\mathrm{d}^3k$ and $n_{eq} := \sum_b \int_{V_b} f_{eq}(\vec{k}, b)\mathrm{d}^3k$. Inserting (3.33) into (3.32) results in

(3.34)
$$\left(H - \frac{|f_{eq}\rangle\langle H_1|}{n_{eq}}\right)\partial_t|f\rangle + \left(H - \frac{|f_{eq}\rangle\langle H_1|}{n_{eq}}\right)\dot{\vec{r}}\cdot\nabla_{\mathbf{r}}|f\rangle + \left(H - \frac{|f_{eq}\rangle\langle H_1|}{n_{eq}}\right)\dot{\vec{k}}\cdot\nabla_{\mathbf{k}}|f\rangle = |f\rangle - |f_{eq}\rangle\frac{n}{n_{eq}}.$$

With the definition $S^{-1}(\vec{k}, b|\vec{k}_0, b_0) := H(\vec{k}, b|\vec{k}_0, b_0) - f_{eq}(\vec{k}, b)H_1(\vec{k}_0, b_0)/n_{eq}$ we find (2.4). Finally, one obtains for the g-moment

$$S_g^{-1}(\vec{k}_0, b_0) := H_g(\vec{k}_0, b_0) - \frac{\langle g\rangle_{eq}\, H_1(\vec{k}_0, b_0)}{n_{eq}}. \tag{3.35}$$

Note that $S_g^{-1}(\vec{k}_0, b_0)$ fulfills by definition the equations

$$AS_g^{-1} = S_g^{-1}(\vec{k}, b) + \left(\frac{g(\vec{k}, b) - \frac{\langle g\rangle_{eq}}{n_{eq}}}{W_{tot}(\vec{k}, b)}\right) \Leftrightarrow |S_g^{-1}\rangle = A|S_g^{-1}\rangle - |h_g\rangle, \tag{3.36}$$

$$\langle S_g^{-1}|f_1\rangle = 0, \tag{3.37}$$

where $h_g(\vec{k}, b) := (g(\vec{k}, b) - \frac{\langle g\rangle_{eq}}{n_{eq}})/W_{tot}(\vec{k}, b)$ (compare with (3.20) and (3.21)).

The proof of the existence and of the uniqueness of S_g^{-1} up to a constant is herewith completed.

Putting (3.30) and (3.35) together leads to

(3.38)
$$\left|S_g^{-1}\right\rangle = -\left(\mathbb{1} - |f_1\rangle\langle f_1|\right)\circ\sum_{i=0}^{\infty}\left(P_L\circ A\circ P_L\right)^i|h_g\rangle = -\left(\mathbb{1} - |f_1\rangle\langle f_1|\right)\circ\sum_{i=0}^{\infty}A^i|h_g\rangle,$$

which is an iterative method for computing S_g^{-1} for any given operator S and function g.

3.4. The Dirac delta distribution. Before going to the applications we would like to discuss the hypothesis on $w(\vec{k}, b|\vec{k}', b')$. At the beginning of this section we chose $w(\vec{k}, b|\vec{k}', b')$ to be a continuous function of its arguments $\vec{k}$ and $\vec{k}'$. In many "physical" models, however, the function $w(\vec{k}, b|\vec{k}', b')$ is replaced with a distribution (usually a sum of Dirac's delta "functions" of a continuous function of $\vec{k}$ and $\vec{k}'$). In the following we argue against delta distributions (functions) based on the concept of regularization of the delta distribution. A function belonging to a family of continuous functions $\delta_\gamma(\varepsilon)$ depending on a continuous parameter γ is called a *regularization of the delta function* iff $\lim_{\gamma\to\infty}\int_{-\infty}^{\infty}\delta_\gamma(\varepsilon)f(\varepsilon)d\varepsilon = f(0)$ for all continuous functions of the energy $f(\varepsilon)$, and $\int_{-\infty}^{\infty}\delta_\gamma(\varepsilon)d\varepsilon = 1$ for all γ. Now, remember that the delta distributions contained in the usual scattering rates arise from Fermi's golden rule, i.e., from the regularization $\delta_\gamma^{(F)}(\varepsilon) := (\sin(\varepsilon\gamma)^2)/(\gamma\pi\varepsilon^2)$. Replacing in the scattering rate w the delta distributions with $\delta_\gamma^{(F)}(\varepsilon)$ for any $\gamma < \infty$, the hypotheses on w are again fulfilled. This fact can be interpreted physically, mathematically, and numerically.

Physically, a simple argument can be found for why γ should be smaller than ∞: From the Heisenberg uncertainty principle there is no exact energy conservation in a finite amount of time. Thus, the delta distribution is just a useful approximation. Mathematically, the situation is quite different if one considers the limit $\gamma \to \infty$ before or after computing the MISO. Taking the limit before the computation of the MISO leads to an operator A, which is not compact and, even worse, leads to infinitely many solutions to the BE at thermodynamic equilibrium (see [11]). Taking the limit after the computation of the MISO trivially leads to a single well-defined solution for each $g \in C^0(K)$. Numerically, a computer using double arithmetic cannot digitize the difference between a delta distribution and, e.g., the regularization $\delta_\gamma^{(a)}(\varepsilon)$ such that $\mathrm{supp}(\delta_\gamma^{(a)}) = [-\gamma^{-1}, \gamma^{-1}]$ and $\gamma > 10^{307}$. This means that when solving the BE on a computer, one is actually working with a delta distribution. For these reasons we conclude that a delta distribution contained in the scattering rates can and should be replaced with a well-chosen regularization. Metaphorically speaking, delta distributions give birth to operators which are a bit like monsters (due to their noncompactness) and, although they are interesting objects from a mathematical point of view (see, e.g., [2] and [11]), they create artificial problems from a physical point of view.

4. Applications.

4.1. Introduction. The knowledge of the MISO and of the solution f of the BE is necessary and sufficient to compute all transport parameters next to and far from thermodynamic equilibrium. In this section, five important applications are presented.

4.2. Low-field solution to the BE. The space-homogeneous, stationary BE

$$-\frac{q}{\hbar}\vec{E}\cdot\nabla_{\mathbf{k}}|f\rangle - q(\vec{v}\wedge\vec{B})\cdot\nabla_{\mathbf{k}}|f\rangle = S|f\rangle \tag{4.1}$$

can be solved for small electric and magnetic fields using the ansatz (see, e.g., [10])

$$f(\vec{k}) = f_{eq}(\varepsilon(\vec{k})) + q\frac{\partial f_{eq}}{\partial\varepsilon}(\varepsilon(\vec{k}))\vec{E}\cdot\vec{\Lambda}^a(\vec{k}) + q\frac{\partial f_{eq}}{\partial\varepsilon}(\varepsilon(\vec{k}))(\vec{v}\wedge\vec{B})\cdot\vec{\Lambda}^b(\vec{k}). \tag{4.2}$$

Inserting (4.2) into (4.1) and taking into account only the first order terms in the magnetic and electric fields leads to

$$(\vec{v}\wedge\vec{B})\cdot\vec{\Lambda}^b = 0, \tag{4.3}$$

because $(\vec{v}\wedge\vec{B})\cdot\vec{v} = 0$, and

$$\langle(1-f_{eq})(\vec{v})_i| = -\langle(\vec{\Lambda}^a)_i)|S. \tag{4.4}$$

It is important that (4.4) be derived only by using the principle of detailed balance. The solution to (4.4) is trivially

$$(\vec{\Lambda}^a)_i(\vec{k}) = -S^{-1}_{(1-f_{eq})v_i}(\vec{k}). \tag{4.5}$$

Therefore, the solution to the low-field BE is

$$f(\vec{k}) = f_{eq}\left(1 + \frac{q}{k_BT}(1-f_{eq})\vec{E}\cdot S^{-1}_{(1-f_{eq})\vec{v}}\right) \tag{4.6}$$

in the case of Fermi–Dirac statistics, and

$$f(\vec{k}) = f_{eq}\left(1 + \frac{q}{k_B T}\vec{E}\cdot S_{\vec{v}}^{-1}\right) \tag{4.7}$$

in the case of Boltzmann statistics.

In the low-field case the solution of the BE is therefore determined only by the equilibrium distribution f_{eq} and by $S^{-1}_{(1-f_{eq})\vec{v}}$ (resp., $S_{\vec{v}}^{-1}$).

4.3. Transport parameters. As already mentioned in section 2, tensorial transport parameters can be exactly computed using MISO. For example, the mobility is given by

$$\mu_{ij} := \frac{q}{n\hbar}\int_K S^{-1}_{v_i}\partial_{k_j} f d^3k \tag{4.8}$$

and the diffusivity tensor by

$$D_{ij} := -\frac{1}{n}\int_K S^{-1}_{v_i} v_j f d^3k. \tag{4.9}$$

Note that in the case of Boltzmann statistics, setting $f = f_{eq}$ in (4.8) and (4.9) yields the well-known Einstein relation for all components of the tensors

$$\frac{k_B T}{q}\mu_{ij} = D_{ij}. \tag{4.10}$$

These transport coefficients are exact and unique. Their definition does not require any restrictions except those already contained in the BE. If they are used in the associated TM, its solution will reproduce the corresponding moment(s) of the BE. In two and three dimensions this is, to the authors' knowledge, the first generally valid scheme ever described which can be used to compute tensorial transport coefficients for all possible geometries and configurations.

A straightforward application is the customization of the model for the transport coefficients for a given device. Using (4.8) and (4.9), transport coefficients can be computed in a device for different bias points (using, e.g., the MC method), especially in the parts of the device where the usual bulk models for the transport coefficients are no longer valid. Then, a customized model for the device can be extracted by choosing a proper local parametrization. In the case of the DD model the transport coefficients can be parametrized using, e.g., the local electric field or the local current density. This custom model will be, of course, valid only for the considered device, but nevertheless, it will enable us to compute the direct current (DC), alternating current (AC), and noise characteristics of the device in a much simpler way than by directly working with the BE.

4.4. Hall factor. When a constant voltage is applied between A and B (see Figure 4.1), and a constant magnetic field $B_{z'}$ is present in the z' direction, then two Hall factors can be defined,

$$R_H := \frac{V_{21}}{d_1 J_{x'} B_{z'}}, \tag{4.11}$$

$$R_H^* := \frac{V_{43}}{d_2 J_{x'} B_{z'}}, \tag{4.12}$$

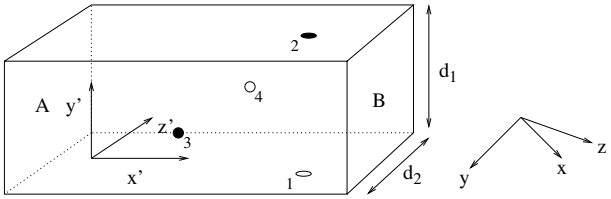


FIG. 4.1. *Piece of bulk material.*

where V_{21} is the voltage between the points 1 and 2, V_{43} is the voltage between the points 3 and 4, and $J_{x'}$ is the current density in x'-direction. Under the assumption that the electric fields $E_{y'}$ in the y'-direction and $E_{z'}$ in the z'-direction are constant, the definitions can be rewritten as

$$R_H := \frac{E_{y'}}{J_{x'} B_{z'}}, \tag{4.13}$$

$$R_H^* := \frac{E_{z'}}{J_{x'} B_{z'}}. \tag{4.14}$$

To obtain expressions for R_H and R_H^*, the current density equation in the space-homogeneous case can be written as

$$\frac{\vec{J}}{nq} = \mu \vec{E} + \alpha \vec{B}, \tag{4.15}$$

where μ is defined in (4.8), and α is defined as

$$\alpha_{ij} := \frac{1}{n} \int_K S_{v_i}^{-1} \left(\vec{v} \wedge \nabla_{\mathbf{k}} f\right)_j d^3k. \tag{4.16}$$

For better readability, only the case of Boltzmann statistics will be considered.

In the low-field case, f can be replaced with (4.7), leading to

$$\alpha_{ij} = \frac{1}{n} \frac{q}{k_B T} \int_K S_{v_i}^{-1} \left(\vec{v} \wedge \nabla_{\mathbf{k}} (f_{eq} \vec{E} S_{\vec{v}}^{-1})\right)_j d^3k, \tag{4.17}$$

where the field-independent term disappeared because $\vec{v} \wedge \vec{v} = 0$. Taking advantage of the linearity of (4.17) in $\vec{E}$, (4.15) can be rewritten as

$$\frac{\vec{J}}{nq} = \mu \vec{E} + B_x \gamma_x \vec{E} + B_y \gamma_y \vec{E} + B_z \gamma_z \vec{E}, \tag{4.18}$$

with

$$(\gamma_l)_{ij} := \frac{1}{n} \frac{q}{k_B T} \int_K S_{v_i}^{-1} \left(\vec{v} \wedge \nabla_{\mathbf{k}} (f_{eq} S_{v_j}^{-1})\right)_l d^3k. \tag{4.19}$$

Therefore,

$$\vec{E} = \left(\mu + B_x \gamma_x + B_y \gamma_y + B_z \gamma_z\right)^{-1} \frac{\vec{J}}{nq}. \tag{4.20}$$

If R is the matrix that transforms the (x, y, z)-coordinate system into the (x', y', z')-coordinate system, then the Hall factors can be written as

$$R_H = \frac{\left(R\left(\mu + B_x\gamma_x + B_y\gamma_y + B_z\gamma_z\right)^{-1} R^{-1}\right)_{yx}}{qnB_{z'}}, \tag{4.21}$$

$$R_H^* = \frac{\left(R\left(\mu + B_x\gamma_x + B_y\gamma_y + B_z\gamma_z\right)^{-1} R^{-1}\right)_{zx}}{qnB_{z'}}. \tag{4.22}$$

Equation (4.21) is more general than the formula given in [7] and reduces to the formula given in [10] in special cases. In the case of unstrained bulk silicon, e.g., because of the symmetries of the crystal, (4.18) takes the form

$$\frac{\vec{J}}{n_{eq}q} = \mu_{eq}\vec{E} - \gamma_{eq}\vec{B} \wedge \vec{E}, \tag{4.23}$$

where

$$\mu_{eq} := \frac{q}{n_{eq}\hbar}\int_K S_{v_x}^{-1}\partial_{k_x} f_{eq} d^3k, \tag{4.24}$$

$$\gamma_{eq} := \frac{1}{n_{eq}}\frac{q}{k_B T}\int_K S_{v_x}^{-1}\left(\vec{v} \wedge \nabla_{\mathbf{k}}(f_{eq}S_{v_y}^{-1})\right)_z d^3k. \tag{4.25}$$

The Hall factors are then

$$R_H = \frac{1}{qn_{eq}}\frac{\gamma_{eq}}{\mu_{eq}^2 + \gamma_{eq}^2 B_{z'}^2}, \tag{4.26}$$

$$R_H^* = 0, \tag{4.27}$$

where R_H and R_H^* are independent of the transformation matrix R, i.e., of the crystal orientation.

4.5. RTs. The RT for the g-moments of the space-homogeneous BE is usually computed using the formula (see [8, p. 136])

$$\tau_g = -\frac{\langle g|f - f_{eq}\rangle}{\langle g|S|f\rangle}. \tag{4.28}$$

At least in all semiconductors (strained and unstrained), (4.28) is fully inappropriate, because in the low-field limit (4.28) reduces to the singular expression $\frac{0}{0}$ in the case of even functions g, such as, e.g., ε, ε^2, v^2, and v^4. To solve this problem, the alternative expression

$$\tau_g = -\frac{\vec{n} \cdot \int (\nabla_{\mathbf{k}} f) S_g^{-1} d^3k}{\vec{n} \cdot \int (\nabla_{\mathbf{k}} f) g d^3k} \tag{4.29}$$

can be used, where $\vec{n}$ is the vector pointing in the direction of the electric field. Equation (4.29) never becomes singular in the limit of vanishingly small electric fields.

An extensive application of this theory to silicon can be found in [4].

4.6. Langevin noise sources. We are interested in the BE with an additional Langevin term, the so-called "Boltzmann–Langevin" equation (BLE)

$$\partial_t f(\vec{r},t,\vec{k},b) + \dot{\vec{r}}\cdot\nabla_{\mathbf{r}} f(\vec{r},t,\vec{k},b) + \dot{\vec{k}}\cdot\nabla_{\mathbf{k}} f(\vec{r},t,\vec{k},b) = Sf + \delta s(\vec{r},t,\vec{k},b). \tag{4.30}$$

The BE describes the average state of an infinite number of systems with identical initial conditions, whereas the BLE describes the evolution of one of these systems. The Langevin source term is responsible for the deviation from the average state.

By multiplying (4.30) by S_g^{-1}, we obtain

$$\langle S_g^{-1}|\partial_t f\rangle + \langle S_g^{-1}|\dot{\vec{r}}\cdot\nabla_{\mathbf{r}} f\rangle + \langle S_g^{-1}|\dot{\vec{k}}\cdot\nabla_{\mathbf{k}} f\rangle = \langle g\rangle - \langle g\rangle_{eq}\frac{n}{n_{eq}} + \langle S_g^{-1}|\delta s\rangle. \tag{4.31}$$

We want to derive an expression for the Fourier transform of correlation functions of $\langle S_g^{-1}|\delta s\rangle$ around a stationary state for the homogeneous BLE with constant density n; i.e., we want to compute

$$C_{gg'}(\omega) := \int_{-\infty}^{\infty} \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} (\langle S_g^{-1}|\delta s\rangle)(t)(\langle S_{g'}^{-1}|\delta s\rangle)(t+s)dt e^{-i\omega s}ds \text{ for constant } n. \tag{4.32}$$

We define the correlation function

$$C(\omega)(\vec{k},b|\vec{k}_0,b_0) := \int_{-\infty}^{\infty} \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} \delta s(t)(\vec{k},b)\delta s(t+s)(\vec{k}_0,b_0)dt e^{-i\omega s}ds. \tag{4.33}$$

Note that

$$C_{gg'}(\omega) = \sum_{b,b_0}\int_{V_b}\int_{V_{b_0}} S_g^{-1}(\vec{k},b)S_{g'}^{-1}(\vec{k}_0,b_0)C(\omega)(\vec{k},b|\vec{k}_0,b_0)\mathrm{d}^3k\mathrm{d}^3k_0. \tag{4.34}$$

We know from [5, eq. (1.55a), p. 21], that for a homogeneous nondegenerate system with given density, we obtain

$$\begin{aligned} C(\omega)(\vec{k},b|\vec{k}_0,b_0) &= \delta^3(\vec{k}-\vec{k}_0)\delta_{b,b_0}\sum_{b_1}\int_{V_{b_1}} w(\vec{k},b|\vec{k}_1,b_1)f(\vec{k},b)\mathrm{d}^3k_1 \\ &+ \delta^3(\vec{k}-\vec{k}_0)\delta_{b,b_0}\sum_{b_1}\int_{V_{b_1}} w(\vec{k}_1,b_1|\vec{k},b)f(\vec{k}_1,b_1)\mathrm{d}^3k_1 \\ &\quad - w(\vec{k}_0,b_0|\vec{k},b)f(\vec{k}_0,b_0) - w(\vec{k},b|\vec{k}_0,b_0)f(\vec{k},b), \end{aligned} \tag{4.35}$$

where $f(\vec{k},b)$ is the stationary homogeneous solution to (3.1).[2]

By plugging (4.35) into (4.34) and rearranging terms, we obtain

$$C_{gg'}(\omega) = \sum_b\int_{V_b} K_{gg'}(\vec{k},b)f(\vec{k},b)\mathrm{d}^3k = \langle K_{gg'}\rangle, \tag{4.36}$$

[2]In the case of particle-particle scattering the corresponding additional contribution to the correlation function has to be added (see [5, eq. (1.55b), p. 21]).

where

$$K_{gg'}(\vec{k},b) := \sum_{b_0} \int_{V_{b_0}} w(\vec{k},b|\vec{k}_0,b_0) \begin{bmatrix} S_g^{-1}(\vec{k},b)S_{g'}^{-1}(\vec{k},b) \\ -S_g^{-1}(\vec{k},b)S_{g'}^{-1}(\vec{k}_0,b_0) \\ -S_g^{-1}(\vec{k}_0,b_0)S_{g'}^{-1}(\vec{k},b) \\ +S_g^{-1}(\vec{k}_0,b_0)S_{g'}^{-1}(\vec{k}_0,b_0) \end{bmatrix} \mathrm{d}^3k_0. \tag{4.37}$$

Note that (4.37) is invariant under the transformation $S_g^{-1}(\vec{k},b) \to S_g^{-1}(\vec{k},b) + \alpha_g f_1$. Therefore, $C_{gg'}(\omega)$ is independent of condition (3.14), as it should be.

Thus, the function $C_{gg'}(\omega)$ is nothing but the expectancy of $K_{gg'}$. Since we can compute $K_{gg'}$, we can also compute $\langle K_{gg'} \rangle$ in a very efficient way with an MC simulation. We call $C_{gg'}(\omega)$ the Langevin noise source of the functions g, g'. It describes white noise because it does not depend on ω.

5. Conclusion. The formalism developed in section 3 cannot only be used for studying interesting systems like strained semiconductors, where the SO is fully dependent on the band-valley index, but also for studying electron-hole systems. To do so we have only to formally replace the distribution function f_h of the holes in the valence bands by $f_e := 1 - f_h$, i.e., the distribution function for the electrons in the valence bands.

We have described a method based on exact S_g^{-1} moments of the inverse scattering operator (MISO) of the Boltzmann equation (BE). This formalism is therefore free of any relaxation time approximation (RTA). We have shown under what sufficient conditions the S_g^{-1} exist, and we gave an explicit algorithm to compute them.

We have demonstrated that the knowledge of the S_g^{-1} enables the exact computation of transport parameters, correlation functions, and Langevin noise sources. Moreover, the important assumptions underlying the transport models (TMs) and method such as, e.g., the impedance field method (IFM) can be critically examined by our approach.

In forthcoming papers, we will give a general discretization scheme to numerically compute any MISO and extend the method to time-dependent scattering operators (SOs).

Acknowledgments. The authors would like to thank Dr. F. Geelhaar, Dr. B. Schmithüsen, Priv.-Doz. Dr. F. M. Bufler, and T. Bühler of the ETH Zurich for helpful discussions.

REFERENCES

[1] A. M. Anile, V. Romano, and G. Russo, *Extended hydrodynamical model of carrier transport in semiconductors*, SIAM J. Appl. Math., 61 (2000), pp. 74–101.

[2] J. Banasiak, *On well-posedness of a Boltzmann-like semiconductor model*, Math. Models Methods Appl. Sci., 13 (2003), pp. 875–892.

[3] K. Bløtekjær, *Transport equations for electrons in two-valley semiconductors*, IEEE Trans. Electron Devices, ED–17 (1970), pp. 38–47.

[4] S. C. Brugger and A. Schenk, *First-principle computation of relaxation times in semiconductors for low and high electric fields*, in Proceedings of the International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), IEEE, Piscataway, NJ, 2005, pp. 151–154.

[5] S. V. Gantsevich, V. L. Gurevich, and R. Katilius, *Theory of fluctuations in nonequilibrium electron gas*, Riv. Nuovo Cimento (3), 2 (1979), pp. 1–87.

[6] T. Grasser, H. Kosina, M. Gritsch, and S. Selberherr, *Using six moments of Boltzmann's transport equation for device simulation*, J. Appl. Phys., 90 (2001), pp. 2389–2396.

[7] C. Jungemann, M. Bartels, S. Keith, and B. Meinerzhagen, *Efficient methods for Hall factor and transport coefficient evaluation for electrons and holes in Si and SiGe based on a full-band structure*, in Extended Abstracts of the Sixth International Workshop on Computational Electronics, IEEE, Piscataway, NJ, 1998, pp. 104–107.

[8] C. Jungemann and B. Meinerzhagen, *Hierarchical Device Simulation. The Monte-Carlo Perspective*, Computational Microelectronics, Springer, Vienna, New York, 2003.

[9] M. G. Krein and M. A. Rutman, *Linear operators leaving invariant a cone in a Banach space*, Uspehi Matem. Nauk (N.S.), 3 (1948), pp. 3–95 (in Russian); Amer. Math. Soc. Translation, 1950, No. 26 (in English).

[10] O. Madelung, *Introduction to Solid-State Theory*, Solid-State Sciences, Springer, Berlin, Heidelberg, New York, 1996.

[11] A. Majorana, *Space homogeneous solutions of the Boltzmann equation describing electron-phonon interactions in semiconductors*, Transport Theory Statist. Phys., 20 (1991), pp. 261–279.

[12] A. Schenk, *Advanced Physical Models for Silicon Device Simulation*, Computational Microelectronics, Springer, Vienna, New York, 1998.

[13] R. Stratton, *Diffusion of hot and cold electrons in semiconductor barriers*, Phys. Rev. 126 (1962), pp. 2002–2014.

[14] D. Werner, *Funktionalanalysis*, 3rd ed., Springer-Verlag, Berlin, 2000.

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1227–1260

ON WEAK PLANE COUETTE AND POISEUILLE FLOWS OF RIGID ROD AND PLATELET ENSEMBLES*

ZHENLU CUI†, M. GREGORY FOREST‡, QI WANG§, AND HONG ZHOU¶

Abstract. Films and molds of nematic polymer materials are notorious for heterogeneity in the orientational distribution of the rigid rod or platelet macromolecules. Predictive tools for structure length scales generated by shear-dominated processing are vitally important: both during processing because of flow feedback phenomena such as shear thinning or thickening, and postprocessing since gradients in the rod or platelet ensemble translate to nonuniform composite properties and to residual stresses in the material. These issues motivate our analysis of two prototypes for planar shear processing: drag-driven Couette and pressure-driven Poiseuille flows. Hydrodynamic theories for high aspect ratio rod and platelet macromolecules in viscous solvents are well developed, which we apply in this paper to model the coupling between short-range excluded volume interactions, anisotropic distortional elasticity (unequal elasticity constants), wall anchoring conditions, and hydrodynamics. The goal of this paper is to generalize scaling properties of steady flow molecular structures in slow Couette flows with equal elasticity constants [M. G. Forest et al., *J. Rheol.*, 48 (2004), pp. 175–192] in several ways: to contrast isotropic and anisotropic elasticity; to compare Couette versus Poiseuille flow; and to consider dynamics and stability of these steady states within the asymptotic model equations.

Key words. liquid crystals, nematic polymers, asymptotic expansions, partial differential equations, instability

AMS subject classifications. 76A15, 82D60

DOI. 10.1137/04061934x

1. Introduction.

Shear dominated flows of nematic liquid crystal polymers (NLCPs) generate anisotropy and spatial heterogeneity in the orientational distribution of the rigid rod or platelet ensemble. These phenomena are well documented in light scattering textures [9, 1, 24, 25, 31]. A characterization of the lengthscales in the molecular distribution responsible for the scattering patterns, and whether they are due to changes in the direction of peak orientation (nematic elasticity) or due to focusing and defocusing of the orientational distribution (molecular elasticity), are the subject of numerous modeling and computational studies (cf. [32, 30, 21, 29]). Molecular orientation features in different flow regimes are of extreme importance for materials design, as they impart anisotropic and nonuniform material properties [34, 18, 19]. Another issue typical of non-Newtonian fluids is flow feedback, where

*Received by the editors November 21, 2004; accepted (in revised form) August 16, 2005; published electronically March 31, 2006. This research was sponsored by Air Force Office of Scientific Research, Air Force Materials Command, grants F49620-02-1-0086 and F49620-03-1-0098; National Science Foundation grants DMS-0204243 and DMS-0308019; and the Army Research Office, Materials Division.

http://www.siam.org/journals/siap/66-4/61934.html

†Department of Mathematics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3250 (zcui@email.unc.edu).

‡Department of Mathematics; Institute for Advanced Materials, Nanoscience & Technology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3250 (forest@amath.unc.edu). The research of this author was supported in part by NASA University Research, Engineering and Technology Institute on Bio Inspired Materials (BIMat) award NCC-1-02037.

§Department of Mathematics, Florida State University, Tallahassee, FL 32306-4510 and Nankai University, Tianjin, 300071, People's Republic of China (wang@math.fsu.edu).

¶Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA 93943 (hzhou@nps.navy.mil).

elastic stresses alter apparent viscosity. This paper is a continuation of our systematic studies of mesostructures, both from free space elasticity patterns absent of external fields and boundary anchoring conditions [11, 12], and from planar Couette cells moving at prescribed slow speeds [13, 14]. We defer to these articles, where a detailed account of analytical results on continuum Leslie–Ericksen–Frank (LEF) models is given, notably by [26, 4, 5, 7, 27, 23, 28].

In [14], the authors considered a Doi–Marrucci–Greco (DMG) mesoscopic orientation tensor model, allowing a full coupling between flow structure, director (nematic) and order parameter (molecular) distortions, and with imposed plate motion and molecular anchoring conditions. The model has been benchmarked in the longwave, monodomain regime with resolved simulations of the Doi kinetic theory [15, 16]. Indeed, the motivation for an analytical study of structure properties is to provide guidance for structure simulations of mesoscopic [17] and kinetic [15, 16] models, where the parameter space is too large to assimilate any kind of collapse of the numerical data through scaling laws.

In this paper, we extend our previous asymptotic scaling analysis in several ways. First, we consider a more general physical model to admit anisotropic distortional elasticity (unequal bend, splay, twist elasticity constants). The second-moment orientation model is derived from a recent generalization of the Doi–Hess–Marrucci–Greco kinetic theory [33] and guides our numerical studies [15, 16], for which there are no preceding numerical or analytical results. Second, the asymptotic analysis is extended from plate-driven Couette cell properties to pressure-driven Poiseuille flows. The boundary conditions consist of molecular orientational anchoring conditions at solid walls, where the degree of order is set by the concentration of the nematic liquid and the principal orientation axis is a free parameter, together with no-slip conditions for the velocity field. We further assume an in-plane orientation tensor (restricting the principal orientation axes of the molecular distribution to the flow-flow gradient plane), and posit that the velocity field varies only transverse to the primary flow direction. These assumptions are not easily lifted, in that the fortuitous diagonalization of the flow-nematic steady balance equations is apparently lost for higher dimensional orientational and spatial degrees of freedom. Finally, we extend the asymptotic analysis to time-dependent model equations.

From this formulation, we develop a formal asymptotic analysis in the slow-plate (so-called small Deborah number) and weak pressure gradient limits, which yield exactly solvable, steady flow-nematic model equations. From the explicit solutions, lengthscale selection criteria and scaling properties become explicit, parameterized in terms of *molecular parameters* (nematic concentration N, molecule aspect ratio r, persistence length $\mathcal{L}$ of distortional elasticity, persistence length L of the anisotropic distortional elasticity) [33], and *experimental conditions* (gap width ($2h$), plate speeds $\pm v_0$ for plane Couette flows and pressure gradient $\frac{\partial p}{\partial x}$ for the plane Poiseuille flow, and plate anchoring conditions on the molecular field). From the time-dependence in the asymptotic equations, we explore transient solutions at the first and second order in the asymptotic scheme to infer stability of the steady states within the asymptotic balance equations. We first consider plane Couette flow, followed in the next section by plane Poiseuille flow.

2. Spatial structures and their stability in plane Couette flows. We consider plane Couette flow between two parallel plates located at $y = \pm h$ and moving with velocity $\mathbf{v} = (\pm v_0, 0, 0)$, respectively, in Cartesian coordinates (x, y, z). Figure 1 depicts the cross section of the flow geometry on the (x, y) plane.

Here we consider flow-orientation interactions in weak plane Couette flow, char-

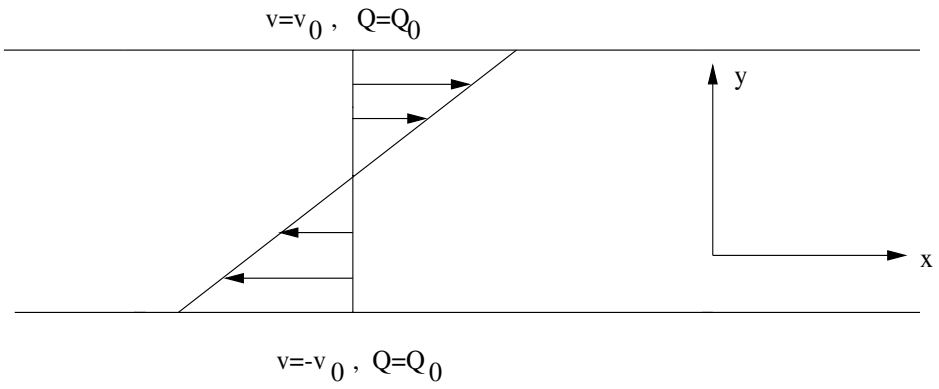


FIG. 1. *Geometry of plane Couette flow. The gap width in the shear cell is* $2h$. *The LCP in the cell is sheared by moving the upper plate with a constant speed* v_0 *and the lower one with the same speed in the opposite direction. At the bounding surfaces, the orientation tensor is equal to its equilibrium value.*

acterized by a small effective or averaged shear rate. We nondimensionalize using the gap half-width (h) between the shearing plates and the nematic polymer mean relaxation time $t_0 = \frac{1}{D_r^0}$, where D_r^0 is the rotary diffusivity for the rigid rod or platelet [33]. We denote the position vector by $\mathbf{x}$, the velocity by $\mathbf{v}$, the extra stress tensor by τ, and the pressure by p, respectively. The dimensionless flow and stress variables are defined by:

$$\tilde{\mathbf{v}} = \frac{t_0}{h}\mathbf{v}, \quad \tilde{\mathbf{x}} = \frac{1}{h}\mathbf{x}, \quad \tilde{t} = \frac{t}{t_0}, \quad \tilde{\tau} = \frac{h^2}{f_0}\tau, \quad \tilde{p} = \frac{h^2}{f_0}p, \tag{1}$$

where $f_0 = \rho h^4/t_0^2$ is a mesophase bulk force and ρ is the nematic polymer (NLCP) density. Let c be the NLCP number density, k the Boltzmann constant, T absolute temperature, N a dimensionless concentration, η_s the solvent viscosity, and $\zeta_i, i = 1, 2, 3$ three friction coefficients related to NLCP-solvent interactions. $\mathcal{L}$ measures the range of isotropic elastic interaction while L does so for the anisotropic elastic interaction [33]. The following eight dimensionless parameters arise:

$$Re = \frac{\rho h^2}{t_0 \eta_s}, \quad \alpha = \frac{3ckTt_0^2}{h^2\rho}, \quad Er = \frac{8h^2}{N\mathcal{L}^2}, \quad \mu_i = \frac{3ckT\zeta_i t_0}{h^2\rho}, \quad i = 1, 2, 3, \quad \theta = \frac{L^2}{\mathcal{L}^2}. \tag{2}$$

α measures the strength of elastic energy relative to kinetic energy; Re is the solvent Reynolds number; Er is the Ericksen number which measures the relative strength of the short-range nematic potential and the isotropic distortional elasticity potential; θ measures the degree of anisotropy in the distortional elasticity, with values limited to $[-1, \infty)$; $1/\mu_i, i = 1, 2, 3$ are three nematic Reynolds numbers. We drop the tilde ˜ on all variables from now on so that all equations and figures in the following correspond to normalized variables, length, and time scales.

The dimensionless forms of the balance of linear momentum, stress constitutive equation, and the continuity equation (dimensional forms are in [33]) take the following form.

Linear momentum balance:

$$\frac{d}{dt}\mathbf{v} = \nabla \cdot (-p\mathbf{I} + \tau), \tag{3}$$

where external forces are neglected.

Continuity equation:

$$\nabla \cdot \mathbf{v} = 0. \tag{4}$$

Constitutive equation for the extra stress:

$$\begin{aligned}
\tau = {} & 2\eta\mathbf{D} + a\alpha[\mathbf{M} - \tfrac{\mathbf{I}}{3} - \tfrac{N}{2}((\mathbf{I} + \tfrac{1}{3NEr}\Delta)\mathbf{M}\cdot\mathbf{M} + \mathbf{M}\cdot(\mathbf{I} + \tfrac{1}{3NEr}\Delta)\mathbf{M} \\
& -2(\mathbf{I} + \tfrac{1}{3NEr}\Delta)\mathbf{M} : \mathbf{M}_4)] - \tfrac{\alpha}{6Er}(\Delta\mathbf{M}\cdot\mathbf{M} - \mathbf{M}\cdot\Delta\mathbf{M}) - \tfrac{\alpha}{12Er}[\nabla\mathbf{M} : \nabla\mathbf{M} \\
& -(\nabla\nabla\mathbf{M}) : \mathbf{M}] + \tfrac{a\alpha\theta}{12Er}[4\mathbf{M}_6 :: \nabla\nabla\mathbf{M} + 2\mathbf{M}_4\nabla\nabla :: \mathbf{M}_4 - \nabla\nabla\mathbf{M}\dddot{:}\mathbf{M}_4 - (\nabla\nabla\mathbf{M}\dddot{:}\mathbf{M}_4)^T \\
& -\mathbf{M}_4\dddot{:}\nabla\nabla\mathbf{M} - (\mathbf{M}_4\dddot{:}\nabla\nabla\mathbf{M})^T - (\mathbf{M}\nabla\nabla\dddot{:}\mathbf{M}_4)^T - \mathbf{M}\nabla\nabla\dddot{:}\mathbf{M}_4] - \tfrac{\alpha\theta}{12Er}[\nabla\nabla\mathbf{M}\dddot{:}\mathbf{M}_4 \\
& -(\nabla\nabla\mathbf{M}\dddot{:}\mathbf{M}_4)^T - \mathbf{M}_4\dddot{:}\nabla\nabla\mathbf{M} + (\mathbf{M}_4\dddot{:}\nabla\nabla\mathbf{M})^T - \mathbf{M}\nabla\nabla\dddot{:}\mathbf{M}_4 + (\mathbf{M}\nabla\nabla\dddot{:}\mathbf{M}_4)^T] \\
& +[\mu_1(a)(\mathbf{D}\mathbf{M} + \mathbf{M}\mathbf{D}) + \mu_2(a)\mathbf{D} : \mathbf{M}_4],
\end{aligned} \tag{5}$$

where $\mathbf{M}$ is the second moment of the orientational probability density function of the kinetic theory (called the structure tensor), $\mathbf{M}_4$ and $\mathbf{M}_6$ are the fourth and sixth moment of the probability density function, respectively, $\eta = 1/Re + \frac{1}{2}\mu_3(a)$ and $a = \frac{r^2+1}{r^2-1}$ parameterizes the aspect ratio r of the spheroidal molecules, where $0 < a \le 1$ corresponds to a rod-like molecule and $-1 \le a < 0$ for platelets [33].

The boundary conditions on velocity $\mathbf{v}$ are scaled to

$$\mathbf{v}|_{y=\pm 1} = (\pm De, 0, 0), \tag{6}$$

where

$$De = \frac{t_0 v_0}{h}, \tag{7}$$

the *Deborah number*, is the ratio of the relaxation time relative to the time scale set by the moving plates in the shear experiment. Weak shear is defined by a small value of De indicating the time scale set by the shear experiment is much larger than the molecular relaxation time scale. Following previous studies [8, 11, 12], we assume strong molecular anchoring at the plates given by the quiescent nematic equilibrium of the orientation tensor (the deviatoric part of the structure tensor) $\mathbf{Q}_0 = \mathbf{M}_0 - \frac{\mathbf{I}}{3} = s_0(\mathbf{nn} - \frac{\mathbf{I}}{3})$. The rest state equilibrium of $\mathbf{Q}$ at sufficiently high concentrations is a uniaxial nematic phase, with unique order parameter,

$$s_0 = \frac{1}{4}\left[1 + 3\sqrt{1 - \frac{8}{3N}}\right]. \tag{8}$$

The uniaxial director $\mathbf{n}$ is arbitrary for quiescent phases; this degeneracy is broken experimentally by mechanical or chemical plate preparations. We model a uniform plate anchoring condition, either parallel to the flow direction, called tangential anchoring, or perpendicular to the shearing plates, called normal (or homeotropic) anchoring.

The time evolution equation of $\mathbf{M}$ (in dimensionless form) is given by [33]:

$$
(9)\quad \begin{cases} \frac{d}{dt}\mathbf{M} - \Omega\cdot\mathbf{M} + \mathbf{M}\cdot\Omega - a[\mathbf{D}\cdot\mathbf{M} + \mathbf{M}\cdot\mathbf{D}] = -2a\mathbf{D} : \mathbf{M}_4 \\ -6[\mathbf{Q} - N(\mathbf{M}\cdot\mathbf{M} - \mathbf{M} : \mathbf{M}_4)] + \frac{1}{Er}[\Delta\mathbf{M}\cdot\mathbf{M} + \mathbf{M}\cdot\Delta\mathbf{M} - 2\Delta\mathbf{M} : \mathbf{M}_4] \\ +\frac{\theta}{2Er}[(\nabla\nabla\mathbf{M})\dddot{:}\mathbf{M}_4 + ((\nabla\nabla\mathbf{M})\dddot{:}\mathbf{M}_4)^T + \mathbf{M}_4\dddot{:}\nabla\nabla\mathbf{M} + (\mathbf{M}_4\dddot{:}\nabla\nabla\mathbf{M})^T \\ +\mathbf{M}\nabla\nabla\dddot{:}\mathbf{M}_4 + (\mathbf{M}\nabla\nabla\dddot{:}\mathbf{M}_4)^T - 4\mathbf{M}_6 :: \nabla\nabla\mathbf{M} - 2\mathbf{M}_4\nabla\nabla :: \mathbf{M}_4]. \end{cases}
$$

In order to arrive at a closed system of governing equations at the level of second order tensors, we approximate fourth ($\mathbf{M}_4$)and sixth ($\mathbf{M}_6$) order tensors in the above governing system of equations using the following simple closure rules:

$$
(10)\qquad \mathbf{M}_4 \approx \mathbf{MM}, \qquad \mathbf{M}_6 \approx \mathbf{MMM}.
$$

These simple closure approximations respect the traceless property of the orientational dynamic equation, and have been shown to yield a good approximation of kinetic theory in the dynamics of monodomains at the nematic concentrations of interest here [10, 15, 16]. These closures are exact when the molecules are aligned perfectly.

We remark that the distortional elastic free energy reduces to the Oseen–Frank energy after the closure approximation, in which the three Frank elastic constants are given by

$$
(11)\qquad k_1 = k_2 = \tfrac{2kT}{Er}s^2(1 + \tfrac{\theta}{3}(1 - s)), \qquad k_3 = \tfrac{2kT}{Er}s^2(1 + \tfrac{\theta}{6}(1 + 4s)).
$$

For rod-like NLCPs,

$$
(12)\qquad 0 < k_1 = k_2 < k_3;
$$

whereas for discotic NLCPs (platelets),

$$
(13)\qquad 0 < k_3 < k_1 = k_2.
$$

2.1. Asymptotic solutions in weak plane Couette flows. We seek asymptotic solutions of the governing system of equations with the boundary conditions given by (6) and (8). We employ a biaxial representation of the orientation tensor [11]

$$
(14)\qquad \mathbf{Q} = s\left(\mathbf{nn} - \frac{1}{3}\mathbf{I}\right) + \beta\left(\mathbf{n}^\perp\mathbf{n}^\perp - \frac{1}{3}\mathbf{I}\right),
$$

where (s, β) are two order parameters measuring the birefringence relative to the optical axes (also called directors) $\mathbf{n}$ and $\mathbf{n}^\perp$ confined to the shearing plane (x, y) and parameterized by a director angle ψ,

$$
(15)\qquad \mathbf{n} = (\cos\psi, \sin\psi, 0), \qquad \mathbf{n}^\perp = (-\sin\psi, \cos\psi, 0),
$$

and $\mathbf{I}$ is the 3×3 identity matrix. We propose the solution ansatz

$$
(16)\qquad v_x = \sum_{k=1}^{\infty} De^k v_x^{(k)}, \qquad (\bullet) = \sum_{k=0}^{\infty} (\bullet)_k De^k, \qquad \psi = \psi_0 + \sum_{k=1}^{\infty} \psi^{(k)} De^k,
$$

where $(\bullet)$ represents the order parameters s, β, respectively. The solution is sensitive to the choice of boundary conditions, so we present tangential ($\psi_0 = 0$) and normal ($\psi_0 = \frac{\pi}{2}$) anchoring conditions separately.

2.2. Tangential anchoring ($\psi_0 = 0$). First, we note that

$$v_x^{(2k)} = \psi^{(2k)} = s_{2k-1} = \beta_{2k-1} = 0, \qquad k = 1, \dots, \infty, \tag{17}$$

demanded by the boundary conditions and the governing equations at the respective orders. This also applies to the case of normal anchoring, but not to tilted anchoring ($\psi_0 \neq 0, \frac{\pi}{2}$) [14]. The governing equations at order $O(1)$ give the equilibrium solution of $\mathbf{Q}$ consistent with the boundary anchoring condition; the equations at order $O(De)$ are obtained by solving the following equations for $\psi^{(1)}$ and $v_x^{(1)}$:

$$\begin{aligned} \frac{\partial \psi^{(1)}}{\partial t} &= A\frac{\partial^2 \psi^{(1)}}{\partial y^2} + B\frac{\partial v_x^{(1)}}{\partial y}, \\ \frac{\partial v_x^{(1)}}{\partial t} &= \frac{\partial \tau_{xy}}{\partial y}, \\ \tau_{xy} &= C\frac{\partial^2 \psi^{(1)}}{\partial y^2} + D\frac{\partial v_x^{(1)}}{\partial y}, \end{aligned} \tag{18}$$

where

$$\begin{aligned} &A = \tfrac{1}{9Er}(s_0+2)(3+\theta(1-s_0)), \qquad B = \tfrac{1}{2}(\lambda_L - 1), \\ &C = -\tfrac{\alpha s_0^2}{18Er}[\theta(1-s_0)\lambda_L + 3(\lambda_L - 1)], \qquad D = \tfrac{1}{3}(\mu_1 s_0 + 3\eta), \end{aligned} \tag{19}$$

where the "tumbling parameter" λ_L is defined by

$$\lambda_L = \frac{a(2+s_0)}{3s_0}; \tag{20}$$

$|\lambda_L| > 1$ corresponds to flow aligning and $|\lambda_L| < 1$ yields director tumbling in monodomain shear flows ($Er \to \infty$) [14].

2.2.1. Steady state features of the major director $\psi^{(1)}$ and primary flow $v_x^{(1)}$. The nonzero leading order steady solution for the velocity, order parameters s and β, and the director angle ψ can be solved explicitly:

$$v_x^{(1)}(y) = y, \quad s_0 = s_0, \quad \beta_0 = 0, \quad \psi^{(1)}(y) = MEr(y^2 - 1), \tag{21}$$

$$M = \tfrac{9}{4(s_0+2)(3+(1-s_0)\theta)}(1 - \lambda_L). \tag{22}$$

Note that, as in the isotropic elasticity limit [14], these solvability conditions imply simple shear flow at leading order in De, and yield that the orientational distribution is dominated by nematic (director) distortions. The prefactor (22) yields that the winding number of the major director between the plates is proportional to the Ericksen number, as with the isotropic elasticity limit [14]. The formula (22) yields the scaling law for elastic distortions which are nonuniform across the gap with length-scale proportional to M^{-1}, which in turn is proscribed by three material parameters: (a, N, θ). For fixed (a, N), $|M|$ decreases as θ increases ($\theta \in [-1, \infty)$). We summarize the dependence of M on θ for given material parameters (a, N) as follows:

- The sign of M governs the "chirality" of nematic distortion, or direction of director rotation from the plates. M is negative for flow-aligning rods ($a > 0, \lambda_L > 1$) and positive for tumbling rods ($a > 0, 0 < \lambda_L < 1$) and discs or platelets ($a < 0$). $|M|$ decreases with respect to all $\theta \in [-1, \infty)$.

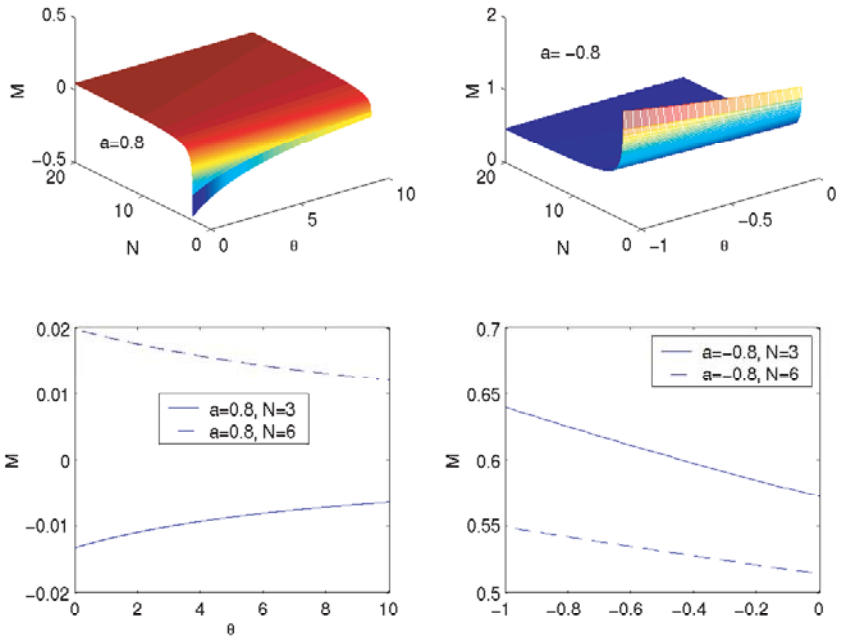


FIG. 2. *M as a function of N and θ with tangentially anchored boundary conditions in weak Couette flow: The top left panel is for rods, with $a = 0.8$, whereas the top right panel is for platelets, with $a = -0.8$. The bottom panels show M for two distinct concentrations of rods (left) and platelets (right).*

- Physically, the director angle winds counterclockwise for rods in the flow-aligning regime $\lambda_L > 1$ from the lower plate to the midplane, then unwinds from the midplane to the upper plane; the orientation reverses for rods in the tumbling regime and for platelets in all regimes.
- At fixed Ericksen number, anisotropic elasticity tends to reduce the magnitude of the director winding for rods while enhancing director distortion for platelets.

Figure 2 depicts M as a function of (N, θ) two values of a corresponding to rods ($a = 0.8$) and platelets ($a = -0.8$). Next, we consider a limited notion of stability of this steady-state structure, by studying transients of the first order governing system of equations in the presence of superimposed spatial disturbances.

2.2.2. Transient behavior of $(v_x^{(1)}, \psi^{(1)})$ near steady states. The transient solution for $v_x^{(1)}$ and $\psi^{(1)}$ (the difference between the time-dependent solution and the steady state) obeys the same homogeneous linear partial differential equations but satisfies a zero boundary condition. Its behavior dictates the stability of the steady state within the asymptotic balance model: the steady state is asymptotically stable if the transient solution vanishes as $t \to \infty$.

PROPOSITION 1. *The steady solution* (21, 22) *of system* (18) *is stable for $AD - BC > 0$ and unstable for $AD - BC < 0$ with respect to zero boundary conditions on $v_x^{(1)}$ and $\phi^{(1)}$.*

Proof. We first consider the case of $BC \neq 0$ and prove the steady solution is stable provided $AD - BC > 0$. In the following proof, we drop the superscripts on ψ and v_x. Extending $(18)_1$ to the boundary and accounting for the boundary condition $\psi(-1,t) = \psi(1,t) = 0$, we have

$$(A\tfrac{\partial^2\psi}{\partial y^2} + B\tfrac{\partial v_x}{\partial y})|_{y=\pm 1} = 0. \tag{23}$$

We introduce a nonnegative functional

$$I(t) = \int_{-1}^{1} [\delta_1 \psi_y^2 + \delta_2 v_x^2] dy \tag{24}$$

with $\delta_1 > 0$ and $\delta_2 > 0$. We note that $A > 0$ and $D > 0$ from (19). ∎

Case 1. $BC < 0$. Choosing $\delta_1 = |C|$ and $\delta_2 = |B|$ and integrating by parts, the time derivative of the nonnegative functional can be estimated:

$$\begin{aligned} \frac{dI(t)}{dt} &= -2\int_{-1}^{1} [\delta_1 A\psi_{yy}^2 + (\delta_1 B + \delta_2 C)\psi_{yy} v_{x,y} + \delta_2 D v_{x,y}^2] dy \\ &= -2\int_{-1}^{1} [|C| A\psi_{yy}^2 + |B| D v_{x,y}^2] dy < 0. \end{aligned} \tag{25}$$

This shows that the steady solution of the system is stable.

Case 2. Choose $\delta_1 = \max(\frac{C^2}{AD}, 1) \geq 1, \delta_2 = \max(\frac{B^2}{AD}, 1) \geq 1$. We have

$$\begin{aligned} \frac{dI(t)}{dt} &= 2\int_{-1}^{1} [\delta_1 \psi_y \psi_{ty} + \delta_2 v_x v_{xt}] dy \\ &= -2\int_{-1}^{1} [\delta_1 A\psi_{yy}^2 + (\delta_1 B + \delta_2 C)\psi_{yy} v_{x,y} + \delta_2 D v_{x,y}^2] dy. \end{aligned} \tag{26}$$

The integrand is quadratic and the discriminant is

$$\begin{aligned} &(\delta_1 B + \delta_2 C)^2 - 4\delta_1 A \delta_2 D \\ &= \delta_1 AD\left(\frac{B^2}{AD}\delta_1 - \delta_2\right) + \delta_2 AD\left(\frac{C^2}{AD}\delta_2 - \delta_1\right) - 2\delta_1\delta_2(AD - BC) \\ &= \delta_1 AD\left[\frac{B^2}{AD}\max\left(\frac{C^2}{AD}, 1\right) - \max\left(\frac{B^2}{AD}, 1\right)\right] \\ &\quad + \delta_2 AD\left[\frac{C^2}{AD}\max\left(\frac{B^2}{AD}, 1\right) - \max\left(\frac{C^2}{AD}, 1\right)\right] \\ &-2\delta_1\delta_2(AD - BC) < -2\delta_1\delta_2(AD - BC) < 0. \end{aligned} \tag{27}$$

The first inequality is based on $\frac{B^2C^2}{(AD)^2} < 1$ because of $AD - BC > 0$ and $BC > 0$. Since $\delta_1 A > 0$, the integrand is always positive; thus $\frac{dI(t)}{dt} < 0$. Hence the steady solution of the system is stable.

The proof for $BC = 0$ is far simpler and omitted. To prove instability when $AD - BC < 0$, we only need to find one unstable mode. Let $\phi(y,t) = \psi_y(y,t)$; the system (18) with the boundary condition becomes

$$\begin{aligned} &\phi_t = A\phi_{yy} + Bv_{x,yy}, \\ &v_{x,t} = C\phi_{yy} + Dv_{x,yy}, \\ &(A\phi_y + Bv_{x,y})|_{y=\pm 1} = 0, \quad v_x(1,t) = 1, \quad v_x(-1,t) = -1. \end{aligned} \tag{28}$$

To find an unstable mode, we seek normal modes of the form

$$\begin{pmatrix} \phi \\ v_x \end{pmatrix} = e^{\gamma t} \begin{pmatrix} \tilde{\phi}(y) \\ \tilde{v_x}(y) \end{pmatrix} \tag{29}$$

and consider the resultant eigenvalue problem

$$P \begin{pmatrix} \tilde{\phi} \\ \tilde{v_x} \end{pmatrix}_{yy} = \gamma \begin{pmatrix} \tilde{\phi} \\ \tilde{v_x} \end{pmatrix}, \qquad P = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \tag{30}$$

where γ is the growth rate. The steady solution is unstable if $\gamma > 0$.

The matrix P has two distinct eigenvalues given by

$$r_1 = \frac{A+D-\sqrt{(A+D)^2-4(AD-BC)}}{2} < 0, \qquad r_2 = \frac{A+D+\sqrt{(A+D)^2-4(AD-BC)}}{2} > 0. \tag{31}$$

Let $\xi = \arctan \frac{A-r_1}{B}$, then

$$\begin{pmatrix} \tilde{\phi} \\ \tilde{v_x} \end{pmatrix} = \begin{pmatrix} \cos\xi & \sin\xi \\ -\sin\xi & \cos\xi \end{pmatrix} \begin{pmatrix} \cos\xi \dfrac{\sin\sqrt{\frac{\gamma}{-r_1}}y}{\sin\sqrt{\frac{\gamma}{-r_1}}} \\ \sin\xi \dfrac{\sinh\sqrt{\frac{\gamma}{r_2}}y}{\sinh\sqrt{\frac{\gamma}{r_2}}} \end{pmatrix} \tag{32}$$

is a solution of (28) satisfying $v_x(1,t) = v_x(-1,t) = 0$ and γ is determined by $(A\phi_y + Bv_{x,y})|_{y=\pm 1} = 0$, which yields

$$(A - r_1)(1 + \tfrac{A(A-r_1)}{B^2}) \coth\sqrt{\tfrac{\gamma}{r_2}} - \sqrt{-r_1 r_2}\cot\sqrt{\tfrac{\gamma}{-r_1}} = 0. \tag{33}$$

As $\gamma \to +\infty$, the first term goes to a finite value while the second one is periodic and varies between $-\infty$ and $+\infty$ within one period. Consequently, the equation has infinitely many positive solutions for γ, which completes the proof, and indicates that the diagnostic $AD - BC$ signals catastrophic instability when it is negative.

We note that for discs ($a < 0$) and flow-aligning rods ($a > 0, \lambda_L > 1$),

$$\begin{aligned} AD - BC = \tfrac{1}{27Er}[&3(2+s_0)(\mu_1 s_0 + 3\eta) + \tfrac{\alpha}{4}((\lambda_L - 1)3s_0)^2 \\ &+\theta(1-s_0)(2+s_0)(\tfrac{\alpha}{12}a((\lambda_L-1)3s_0) + \mu_1 s_0 + 3\eta)] > 0. \end{aligned} \tag{34}$$

Hence, the steady state is always stable for discotic LCPs and flow-aligning rods, and may be unstable only for tumbling rods ($0 < a, 0 < \lambda_L < 1$).

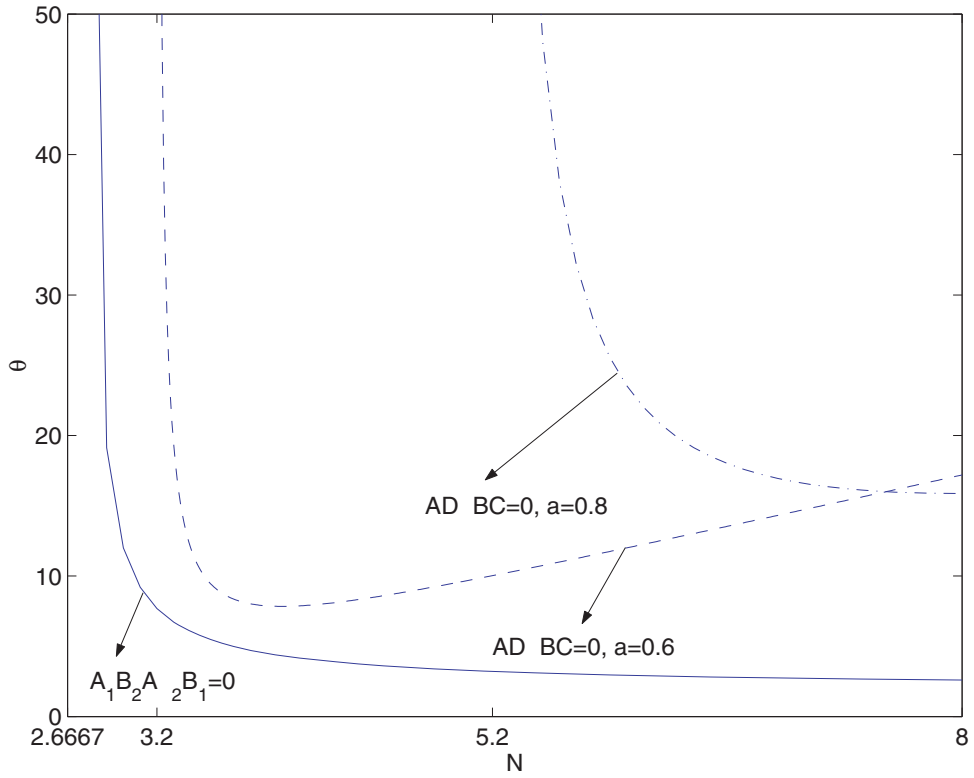


FIG. 3. *The neutral stability curves defined by* $A_1B_2 - A_2B_1 = 0$ *and* $AD - BC = 0$ *with tangentially anchored boundary conditions in weak shear. The arrows point to the positive directions of the discriminant. The systems for* (s_2, β_2) *and* $(\psi^{(1)}, v_x^{(1)})$ *are stable when the discriminant is positive, respectively, unstable otherwise. The parameter values are* $Er = 100, \alpha = 1, \mu_1 = 0.001$ *and* $\eta = 0.002$.

The neutral stability curve ($AD - BC = 0$) depends on the values of the parameters $(\theta, a, N, \alpha, \mu_1, \eta)$. From (34), $AD - BC > 0$ can be translated into two separate constraints on the energy parameter α and θ,

$$\begin{aligned} &\alpha \le \alpha_c = \tfrac{4(\mu_1 s_0 + 3\eta)}{a(1-\lambda_L)s_0}; \text{ or} \\ &\alpha > \alpha_c \quad \text{and} \quad \theta < \theta_c = \tfrac{3(2+s_0)(\mu_1 s_0 + 3\eta) + \frac{\alpha}{4}((\lambda_L - 1)3s_0)^2}{(1-s_0)(2+s_0)(\frac{\alpha}{12}a((1-\lambda_L)3s_0) - \mu_1 s_0 + 3\eta)}. \end{aligned} \tag{35}$$

Figure 3 depicts the stability transition curve in the parameter space (N, θ) at a few selected values of other parameters. We observe that the values of N, θ that yield instability tend to be large and out of the practical range for nematic polymer materials. We also note that the flow-aligning region and the stable region versus (N, θ) both grow significantly as the shape parameter a increases, i.e., as the aspect ratio becomes more extreme. The instability region vanishes as $a \to \frac{3s_0}{2+s_0}$. The stable region also grows as $\mu_1/\alpha, \eta/\alpha$ increase.

We summarize this more precise statement of Proposition 1 in the following corollary.

COROLLARY 1. *The steady asymptotic solution* (21, 22) *is asymptotically stable within the leading order balance equations if either of conditions* (35) *is satisfied. If*

$\alpha > \alpha_c$ and $\theta > \theta_c$, where α_c, θ_c are defined in (35)*, the steady solution is unstable and the leading order system of equations is ill-posed.*

2.2.3. Steady state features of the order parameters s_2 and β_2 at $O(De^2)$. The order parameters (s, β) vanish at leading order in De, with order $O(De^2)$ behavior governed by the equations

$$\begin{aligned}
\frac{\partial s_2}{\partial t} &= \frac{-1}{9Er}\Big(A_1\frac{\partial^2 s_2}{\partial y^2} + B_1\frac{\partial^2 \beta_2}{\partial y^2} + C_1 s_2 + D_1\beta_2 + E_1\big(\frac{\partial \psi^{(1)}}{\partial y}\big)^2 + \\
&F_1\psi_1\frac{\partial^2 \psi^{(1)}}{\partial y^2} + G_1\psi^{(1)}\frac{\partial v_x^{(1)}}{\partial y}\Big) + 2s_0\psi^{(1)}\frac{\partial \psi^{(1)}}{\partial t}, \\
\frac{\partial \beta_2}{\partial t} &= \frac{-1}{9Er}\Big(A_2\frac{\partial^2 s_2}{\partial y^2} + B_2\frac{\partial^2 \beta_2}{\partial y^2} + C_2 s_2 + D_2\beta_2 + E_2\big(\frac{\partial \psi^{(1)}}{\partial y}\big)^2 + \\
&F_2\psi^{(1)}\frac{\partial^2 \psi^{(1)}}{\partial y^2} + G_2\psi^{(1)}\frac{\partial v_x^{(1)}}{\partial y}\Big) - 2s_0\psi^{(1)}\frac{\partial \psi^{(1)}}{\partial t},
\end{aligned} \tag{36}$$

where the coefficients are lengthy and provided in Appendix A. This system of equations is linear in (s_2, β_2) but driven by nonlinear functions of $v_x^{(1)}$ and $\psi^{(1)}$.

The steady solution, with Λ and Γ defined in the appendix, is

$$\begin{aligned}
\beta_2(y) &= K_1\Big(\frac{\cosh(\sqrt{Er}\Lambda y)}{\cosh(\sqrt{Er}\Lambda)} - 1\Big) + K_2\Big(\frac{\cosh(\sqrt{Er}\Gamma y)}{\cosh(\sqrt{Er}\Gamma)} - 1\Big) + R_1 Er(y^2 - 1), \\
s_2(y) &= K_3\Big(\frac{\cosh(\sqrt{Er}\Lambda y)}{\cosh(\sqrt{Er}\Lambda)} - 1\Big) + K_4\Big(\frac{\cosh(\sqrt{Er}\Gamma y)}{\cosh(\sqrt{Er}\Gamma)} - 1\Big) + S_1 Er(y^2 - 1),
\end{aligned} \tag{37}$$

where the coefficients are given in Appendix B.

We denote $\lambda^2 = 1/(Er\Lambda)$ and $\mu^2 = 1/(Er\Gamma)$; λ^2 is always positive and concave up as a function of θ, whereas μ^2 is monotonically decreasing and may change sign as θ varies. For example, μ^2 goes through zero at a critical degree of anisotropy $\theta_c = 2.93$ for $N = 6$. For $\theta > \theta_c$, the steady state becomes highly oscillatory; we show in the study of transient solutions that this behavior coincides with the onset of ill-posedness in the governing system of equations. The behavior of θ_c versus concentration N can be gleaned from Figure 3.

Since the dominating terms in the order parameters near $y = \pm 1$ are

$$\frac{e^{\sqrt{Er}\Lambda y}}{e^{\sqrt{Er}\Lambda}}, \qquad \frac{e^{\sqrt{Er}\Gamma y}}{e^{\sqrt{Er}\Gamma}}, \tag{38}$$

the order parameters have a boundary layer near the wall, whose width is proportional to

$$\frac{1}{\sqrt{Er}\Lambda}, \qquad \frac{1}{\sqrt{Er}\Gamma}, \tag{39}$$

respectively. These are the penetration depths of the wall layer for tangential anchoring, which agrees with the asymptotic analysis of the DMG model [14] in the single elastic constant limit. One finds the order parameters are coupled with anisotropic elasticity, i.e., the orientational distribution is strongly biaxial (birefringent in any plane). For both rods and discs, by comparing the two exponential terms in β_2 and s_2, we notice that the boundary layer in s_2 is governed by $\frac{1}{\sqrt{Er}\Lambda}$, whereas in β_2 by $\frac{1}{\sqrt{Er}\Gamma}$, so that their scaling behavior is incommensurate with the leading order wall layer scaling.

TABLE 1

Steady state features of the order parameter morphology for Couette flow with tangential anchoring (BL denotes boundary layer).

	FA/rods	FA/discs	T/rods	T/discs
$s - s_0$	Concave down	Concave up	Concave up	Concave down & Concave up in BL
β	Concave up	Concave up & Concave down in BL	Concave down	Concave down

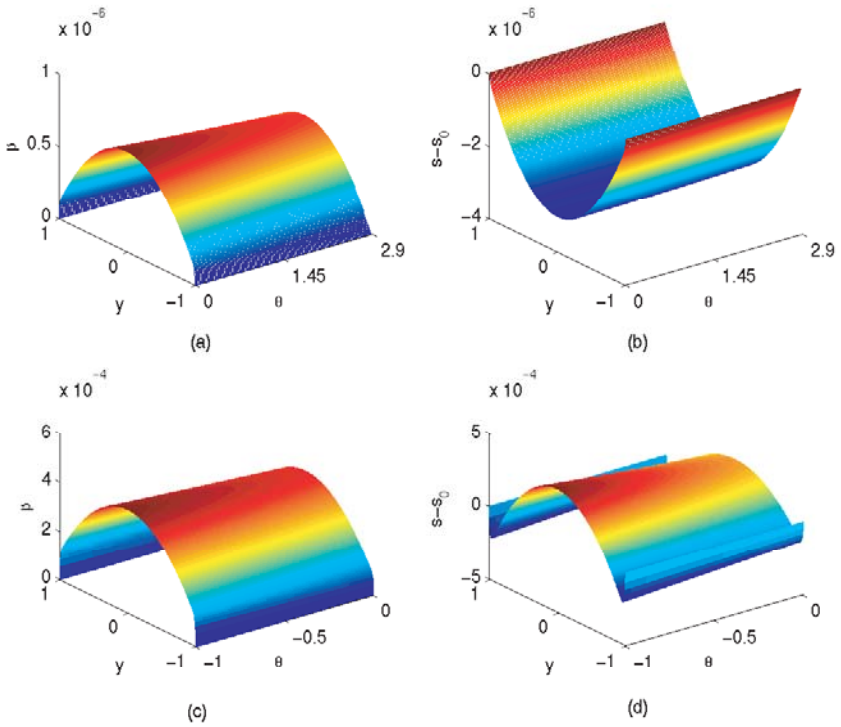


FIG. 4. *The steady-state asymptotic solutions β (left column) and $s - s_0$ (right column) in the tumbling regime as functions of (θ, y) with tangentially anchored boundary conditions in weak Couette flows.* (a) *and* (b) *depict the solution for rods with parameter values* $a = 0.8, N = 6, De = 0.01, Er = 100$. (c) *and* (d) *depict the solution for discs with parameter values* $a = -0.8, N = 6, De = 0.01, Er = 100$.

Table 1 tabulates the general behavior of the steady states in the regime of flow-aligning and tumbling for both rods and discs, in which FA stands for flow-aligning, T denotes tumbling, and BL denotes a boundary layer. The concavity switching phenomenon between tumbling and flow-aligning materials, noted in [14], is observed in both rods and discs. The steady states are in general insensitive to changes in θ. In addition, the order parameter correction s_2 goes through a similar transition for rods versus discs, whereas β has the same concavity; this is another indication of strong biaxiality for anisotropic elasticity. Figure 4 depicts a typical steady state asymptotic solution of (s_2, β_2) for tumbling rods $(a = 0.8)$ and tumbling discs $(a = -0.8)$.

2.2.4. Transient behavior of (s_2, β_2) near steady states. The transient behavior of the order parameters s_2 and β_2 obey:

$$\begin{aligned}\frac{\partial \tilde{s_2}}{\partial t} &= \frac{-1}{9Er}\left(A_1\frac{\partial^2 \tilde{s_2}}{\partial y^2} + B_1\frac{\partial^2 \tilde{\beta}_2}{\partial y^2} + C_1\tilde{s_2} + D_1\tilde{\beta}_2\right),\\ \frac{\partial \tilde{\beta}_2}{\partial t} &= \frac{-1}{9Er}\left(A_2\frac{\partial^2 \tilde{s_2}}{\partial y^2} + B_2\frac{\partial^2 \tilde{\beta}_2}{\partial y^2} + C_2\tilde{s_2} + D_2\tilde{\beta}_2\right),\end{aligned} \tag{40}$$

with zero boundary conditions.

After tedious but straightforward Fourier analysis, the growth ($\sigma' > 0$) or decay ($\sigma' < 0$) of solutions $\tilde{s_2}(y,t)$ and $\tilde{\beta}_2(y,t)$, due to the time-dependent factor $e^{\sigma' t}$, is determined by $sgn(\sigma')$

$$\sigma'_{\pm} = -\tfrac{1}{2}[C_1 + D_2 - A_1k^2 - B_2k^2 \pm$$

$$\sqrt{(C_1 + D_2 - A_1k^2 - B_2k^2)^2 - 4((A_2k^2 - C_2)(D_1 - B_1k^2) + (C_1 - A_1k^2)(D_2 - B_2k^2))}]. \tag{41}$$

We now analyze $sgn(\sigma'_{\pm})$ to deduce stability. For long waves ($|k| << 1$), asymptotic formulae can be derived:

$$\sigma'_{\pm} \sim -\frac{1}{2}[C_1 + D_2 \pm \sqrt{(C_1 + D_2)^2 + 4(C_2D_1 - C_1D_2)}]. \tag{42}$$

Using the formulae in Appendix B, $C_1, D_2 > 0, C_2 = 0$, which implies $\sigma'_{\pm} < 0$. On the other hand, the growth rate for short waves ($|k| >> 1$) is dominated by

$$\sigma'_{\pm} \sim \frac{1}{2}[A_1 + B_2 \pm \sqrt{(A_1 + B_2)^2 + 4(A_2B_1 - A_1B_2)}]k^2. \tag{43}$$

We find $\sigma'_{\pm} < 0$ only for sufficiently small $|\theta|$. Otherwise, $A_1B_2 - A_2B_1$ may be negative, leading to a positive growth rate proportional to k^2, an ill-posed behavior. The transition to ill-posedness, $A_1B_2 - A_2B_1 = 0$, simplifies dramatically to the condition $\theta = \frac{6}{5s_0-2}$, where s_0 is given in (8). This neutral stability curve is plotted in Figure 3.

PROPOSITION 2. *The steady state solution* (s_2, β_2) *of* (36) *is catastrophically unstable if and only if the degree of anisotropic elasticity satisfies*

$$\theta > \frac{6}{5s_0 - 2}. \tag{44}$$

2.2.5. Rheological features of steady structures. The shear viscosity (shear stress divided by local shear rate) at the plates is identical to the averaged shear viscosity over the shear cell; it is a nonzero constant at $O(De)$, given by

$$\eta_{wall} = \frac{\tau_{xy}}{\frac{dv_x^{(1)}}{dy}} = \frac{\alpha a^2(1-\frac{1}{\lambda_L})[\theta(1-s_0)+3(1-\frac{1}{\lambda_L})]}{36(3+\theta(1-s_0))} + \tfrac{1}{3}(\mu_1 s_0 + 3\eta). \tag{45}$$

It can be readily shown that η_{wall} is a slowly varying, decreasing function of the degree of anisotropic elasticity θ for tumbling rods and all platelet nematic liquids; however, η_{wall} increases versus θ for flow-aligning rods. Nonzero first normal stress difference N_1 and second normal stress difference N_2 show up at order $O(De^2)$ and

are given in Appendix C. These non-Newtonian effects are measurable physically, e.g., $N_1 > 0$ corresponds to pushing the parallel plates apart and $N_1 < 0$ corresponds to pulling the plates together. N_1 is a linear combination of α and μ_2 and N_2 is a linear combination of α and $\mu_1 + \mu_2$. At walls, the terms containing $\mu_{1,2}$ drop out so that N_1 and N_2 are proportional to α:

$$\begin{aligned}
N_1 = [&\tfrac{G}{54Er}(K_3\lambda^2 + K_4\mu^2 + 2\,S) \\
&+\tfrac{(-4a\theta+3s_0+16a\theta s_0^3-12a+12as_0^2-12a\theta s_0)}{108Er}\left(K_1\lambda^2 + K_2\mu^2 + 2\,R\right) \\
&+\tfrac{4K^2 s_0}{27Er}(-4a\theta {s_0}^2 - 4a\theta + 24a\theta {s_0}^3 - 6as_0 + 9s_0 + 18a{s_0}^2 - 12a - 16a\theta s_0)]\alpha, \\
N_2 = [&-\tfrac{G}{54Er}(K_3\lambda^2 + K_4\mu^2 + 2\,S) \\
&+\tfrac{(8a\theta+12a-3+8as_0-16as_0^2-12s_0)s_0}{108Er}\left(K_1\lambda^2 + K_2\mu^2 + 2\,R\right) \\
&+\tfrac{4K^2 as_0}{27Er}(8\theta {s_0}^2 + 2\theta - 24\theta {s_0}^3 + 12s_0 - 9s_0 - 18a{s_0}^2 + 6 + 14\theta s_0)]\alpha,
\end{aligned} \tag{46}$$

where G, K, R, S, λ, μ and $K_i (i = 1, 2, 3, 4)$ are given in Appendix C.

Figure 5 depicts N_1 and N_2 for tumbling rods as well as discs as functions of y at some parameter values; Figure 6 shows N_1 and N_2 for flow-aligning rods and discs. Table 3 lists the averaged normal stress differences calculated in four representative cases. We summarize the noticeable features in the stress differences below.

- For flow-aligning rods, $N_1 < 0$ and $N_2 > 0$ across the gap; their signs change in plate boundary layers for small degree θ of elastic anisotropy.
- These properties reverse for flow-aligning or tumbling discs, which may experience sign changes in N_1 and N_2 in the middle of the plate gap for small $|\theta|$.
- For tumbling rods and all discotic NLCPs, $N_1 > 0$ and $N_2 < 0$, with sign changes for rods in the wall boundary layer for large θ, and sign changes for platelets in the midgap at small $|\theta|$.
- The averages across the gap yield $N_1 > 0$ and $N_2 < 0$ for flow-aligning rods, and $N_1 < 0$ and $N_2 > 0$ in all other cases.

2.3. Homeotropic anchoring ($\psi_0 = \frac{\pi}{2}$). The boundary anchoring condition affects only the coefficients of the governing system of partial differential equations at each order. The structure of the equations at $O(De)$ for $(v_x^{(1)}, \psi^{(1)})$ are identical to (18) with the following new coefficients:

$$\begin{aligned}
&A = \tfrac{1}{9Er}(s_0 + 2)(3 + \theta(2s_0 + 1)), \qquad B = -\tfrac{1}{2}(1 + \lambda_L), \\
&C = \tfrac{\alpha s_0}{18Er}[\theta(2s_0 + 1)\lambda_L + 3(\lambda_L + 1)], \qquad D = \tfrac{1}{3}(\mu_1 s_0 + 3\eta).
\end{aligned} \tag{47}$$

The nonzero leading order solution for the order parameters, primary velocity component, and major director angle are

$$s_0 = s_0, \quad \beta_0 = 0, \quad v_x^{(1)}(y) = y, \quad \psi^{(1)}(y) = MEr(y^2 - 1), \tag{48}$$

$$M = \frac{9}{2(s_0 + 2)(3 + \theta(2s_0 + 1))}(1 + \lambda_L). \tag{49}$$

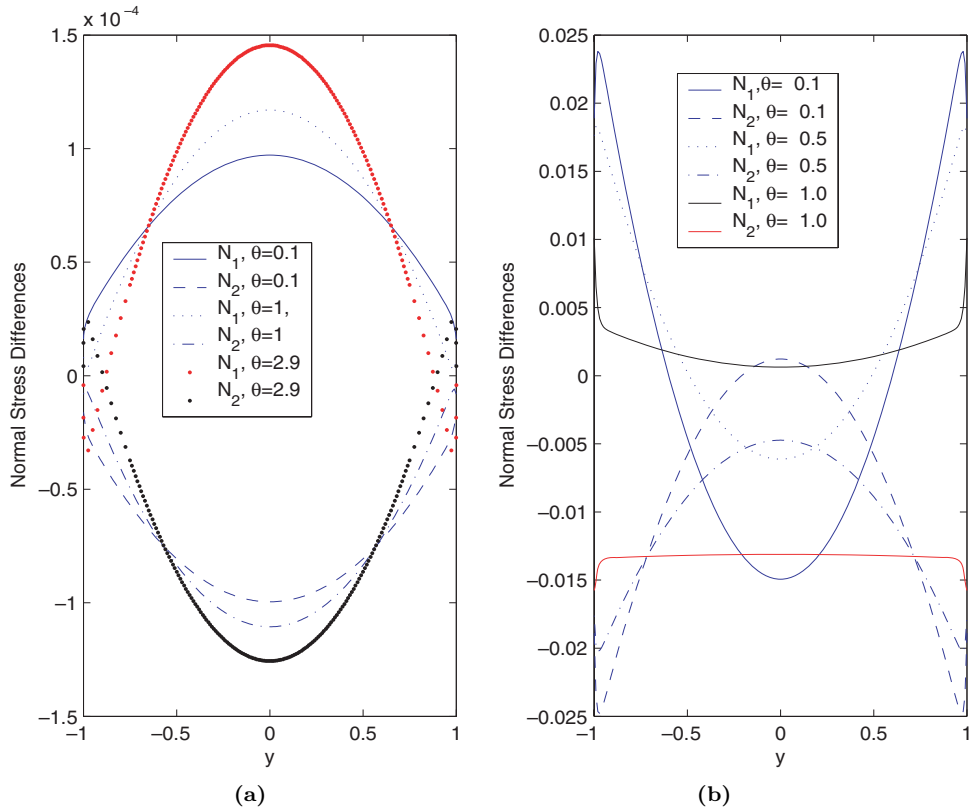


FIG. 5. *The normal stress differences N_1 and N_2 in tumbling regimes as functions of y at selected values of θ with tangentially anchored boundary conditions in weak shear.* (a) *Rods:* $a = 0.8, N = 6, De = 0.01, Er = 100, \alpha = 10, \mu_1 = 0.01, \eta = 0.02$. (b) *Discs:* $a = -0.8, N = 6, De = 0.01, Er = 100, \alpha = 10, \mu_1 = -0.01, \eta = 0.02$.

It is easy to see that $|M|$ decreases with respect to θ; M is negative for flow-aligning discs ($a < 0, \lambda_L < -1$) and positive for all other cases. Thus, directors wind counterclockwise from the lower shearing plate to the midplane, and unwind from the midplane to the upper plate for flow-aligning discs. The winding reverses direction in the other cases.

Analogous to tangential anchoring, *steady states may be catastrophically unstable for discs in the flow-aligning regime $a < 0, \lambda_L < -1$ if $AD - BC < 0$, yet are stable in the other cases, where*

$$AD - BC = \tfrac{1}{27Er}\,[(s+2)(\mu_1 s_0 + 3\eta)(\theta(2s_0+1)+3) + \tfrac{\alpha s_0}{4}(1+\lambda_L)(a\theta(2s_0+1)(s_0+2) + 9s_0(1+\lambda_L))]. \tag{50}$$

PROPOSITION 3. *The steady-state solution is stable so long as*

$$\theta > \theta_c = -\frac{12(s_0+2)(\mu_1 s_0 + 3\eta) + 9\alpha s_0^2(1+\lambda_L)^2}{4(s_0+2)(2s_0+1)(\mu_1 s_0 + 3\eta) + a\alpha s_0(1+\lambda_L)(1+2s_0)(2+s_0)}. \tag{51}$$

The governing system of equations for the order parameters (s, β) at order $O(De^2)$ is of the same form as tangential anchoring, with different coefficients given in Appendix C.

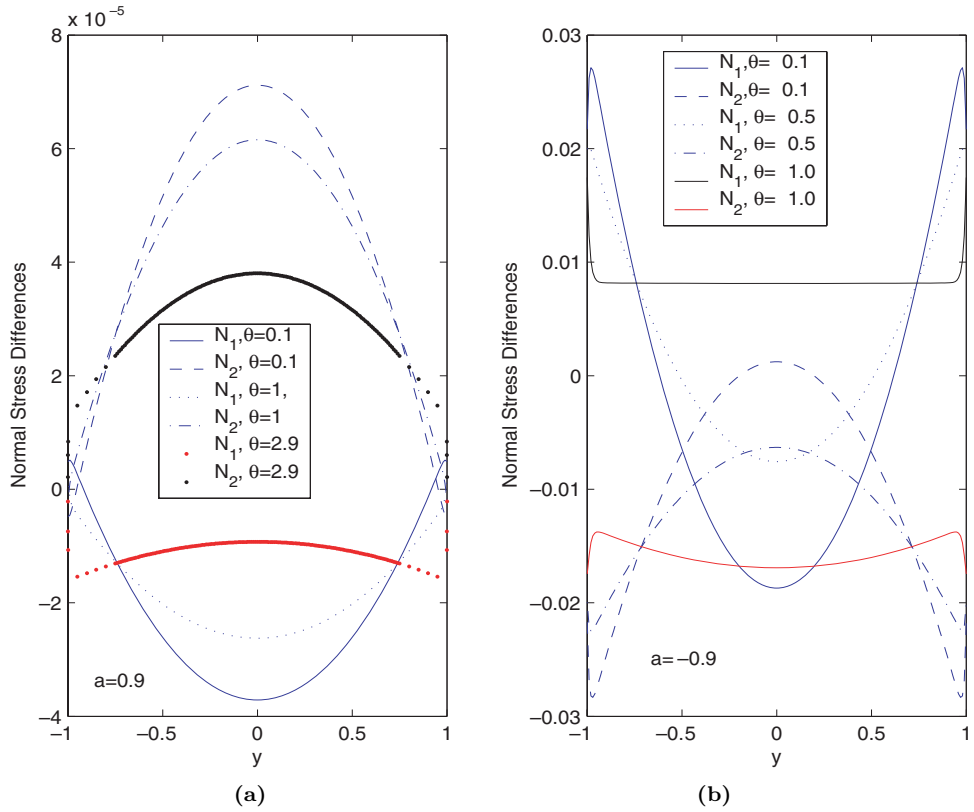


FIG. 6. *The normal stress differences N_1 and N_2 in flowing-aligning regimes as functions of y at selected values of θ with tangentially anchored boundary conditions in weak shear.* (a) *Rods:* $a = 0.9, N = 6, De = 0.01, Er = 100, \alpha = 10, \mu_1 = 0.01, \eta = 0.02$. (b) *Discs:* $a = -0.9, N = 6, De = 0.01, Er = 100, \alpha = 10, \mu_1 = -0.01, \eta = 0.02$.

The steady solutions are

$$\beta_2(y) = K_{11}\left(\frac{\cosh(\sqrt{Er}\Lambda y)}{\cosh(\sqrt{Er}\Lambda)} - 1\right) + R_2 Er(y^2 - 1),$$
$$s_2(y) = K_{21}\left(\frac{\cosh(\sqrt{Er}\Lambda y)}{\cosh(\sqrt{Er}\Lambda)} - 1\right) + K_{22}\left(\frac{\cosh(\sqrt{Er}\Gamma y)}{\cosh(\sqrt{Er}\Gamma)} - 1\right) + S_2 Er(y^2 - 1). \tag{52}$$

In this solution, μ^2 (defined earlier) changes sign only as θ varies below a threshold value θ_d for platelets. Again, the change of sign in μ^2 coincides with ill-posedness in the governing system of equations.

Notice that in s_2, there are two cosh terms, whereas there is only one in β_2. For rods, the second term in s_2 dominates in the boundary layer while the first term dominates for discs. Table 2 tabulates features of the steady states. Compared with results above for tangential anchoring, the steady states with normal anchoring are more sensitive to the degree θ of elastic anisotropy. The order parameter variation versus θ decreases for rods and increases for platelets. The solution profiles switch concavity in the boundary layer for flow-aligning versus tumbling discotics, whereas the concavity remains the same for rods. Figure 7 depicts typical steady solutions for tumbling rods and discs.

TABLE 2
Steady-state features of the order parameter morphology for Couette flow with normal anchoring.

	FA/rods	FA/discs	T/rods	T/discs
$s - s_0$	Concave down	Concave down & Concave up in BL	Concave down & Concave up in BL	Concave up & Concave down in BL
β	Concave up	Concave up	Concave up & Concave down in BL	Concave down

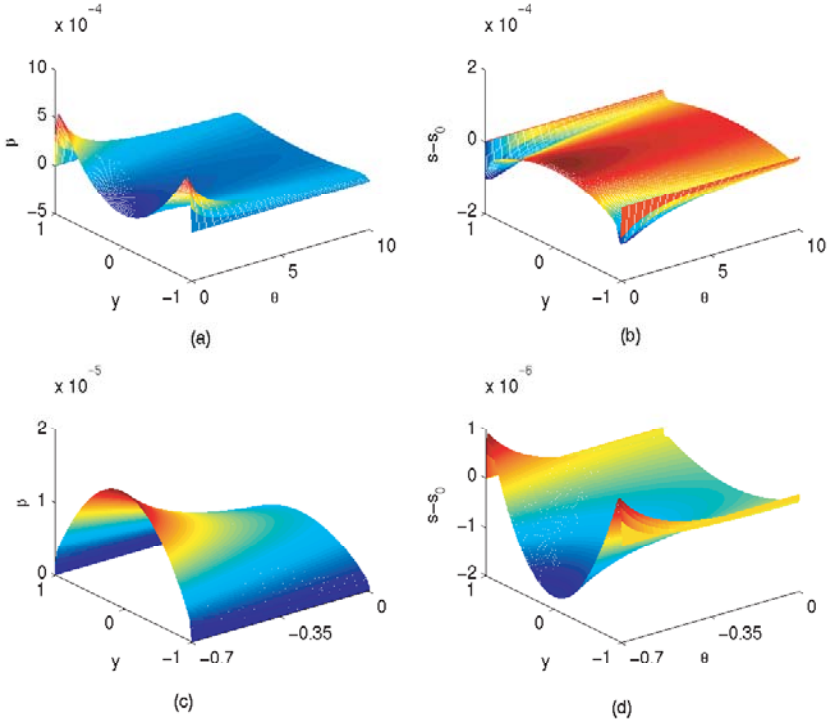


FIG. 7. *The steady-state asymptotic solution as functions of* (θ, y) *in tumbling regime with normally anchored boundary condition in weak shear.* (a) *and* (b) *depict the solution for rods with parameter values* $a = 0.8, N = 6, De = 0.01, Er = 100$. (c) *and* (d) *depict the solution for discs with parameter values* $a = -0.8, N = 6, De = 0.01, Er = 100$.

Again the governing system can become ill-posed. The growth rate formulae for the steady states are identical to (41), but with new coefficients. Notice that the discriminant $A_1B_2 - A_2B_1 = 4(1 - s_0)^2(1 + 2s_0)[3 + \theta(1 + 2s_0)][3 + \theta(1 + 4s_0)]$. For rods, $\theta > 0$ and $A_1B_2 - A_2B_1 > 0$ indicating stability. For discs and $\theta < -\frac{3}{1+4s_0}$, the discriminant is negative so that the steady state is unstable.

PROPOSITION 4. *The system for the two order parameters* (s_2, β_2) *is locally ill-posed if and only if* $\theta < \theta_d$, *where*

$$\theta_d = -\frac{3}{1 + 4s_0}. \tag{53}$$

The shear viscosity at the walls in this case is given by

$$\eta_{wall} = \frac{\tau_{xy}}{\frac{dv_x^{(1)}}{dy}} = \frac{\alpha a^2(1+\frac{1}{\lambda_L})(2\theta(2s_0+1)+3(1+\frac{1}{\lambda_L}))}{36(\theta(2s_0+1)+3)} + \frac{1}{3}(\mu_1 s_0 + 3\eta). \tag{54}$$

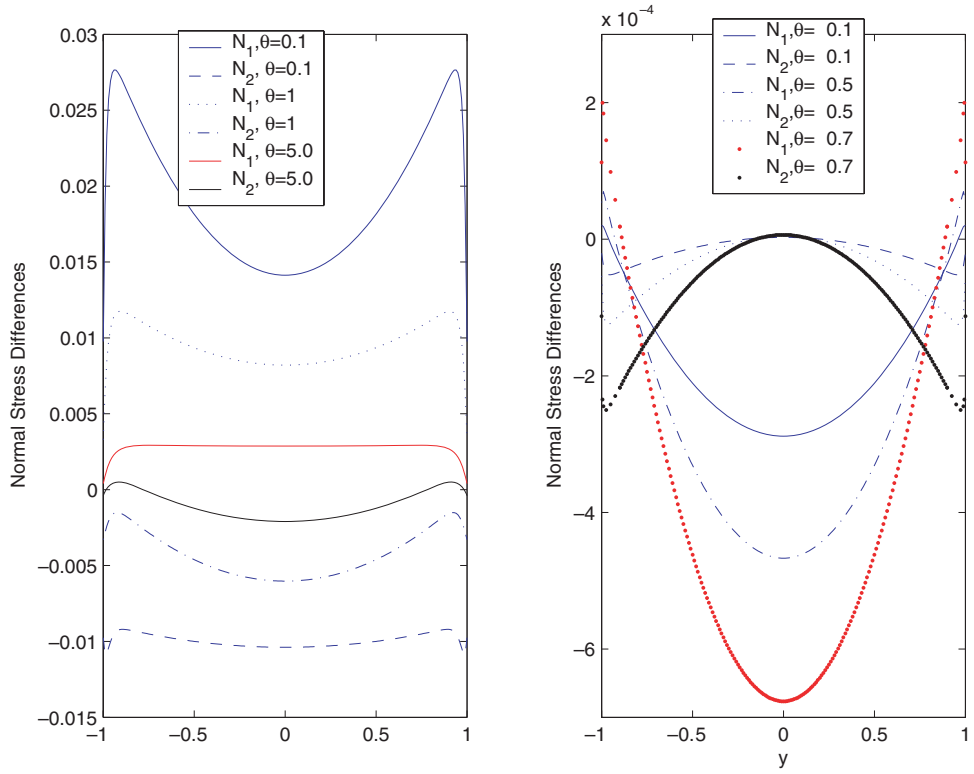


FIG. 8. *The normal stress differences N_1 and N_2 as functions of y in tumbling regime at selected values of θ with normally anchored boundary conditions in weak shear. The parameter values are $a = \pm 0.8, N = 6, De = 0.01, Er = 100, \alpha = 10, \mu_1 = \pm 0.01, \mu_2 = 0.02$ for rods and discs, respectively.*

Similar to the case of tangential anchoring, the viscosity decreases with respect to θ for rods and flow-aligning discs, but increases with respect to θ for tumbling discs. The leading order normal stress differences again show up at $O(De^2)$ and are given in Appendix D.

Figures 8 *and* 9 *depict representative plots of the first and second normal stress differences in tumbling and flow-aligning regimes, respectively, and demonstrate the following properties:*

- *Rods are more sensitive to the variation in degree θ of elastic anisotropy than discs.*
- *For tumbling rods, $N_1 > 0$ and $N_2 < 0$ except there may be a sign change in a boundary layer near the wall at large θ. For tumbling discs, N_1 and N_2 are both negative except that N_1 is positive in a boundary layer near the wall.*
- *The behavior of the normal stress differences does not change much for flow-aligning rods compared with tumbling rods. For flow-aligning discs, $N_1 > 0$ and $N_2 > 0$ except that N_2 is negative in boundary layer near the wall.*
- *The average values across the gap obey $N_1 > 0$ and $N_2 < 0$ for rods, but they may change signs for discs.*

The averaged normal stress differences are tabulated in Table 3.

In summary, the salient predictions from this analysis are:

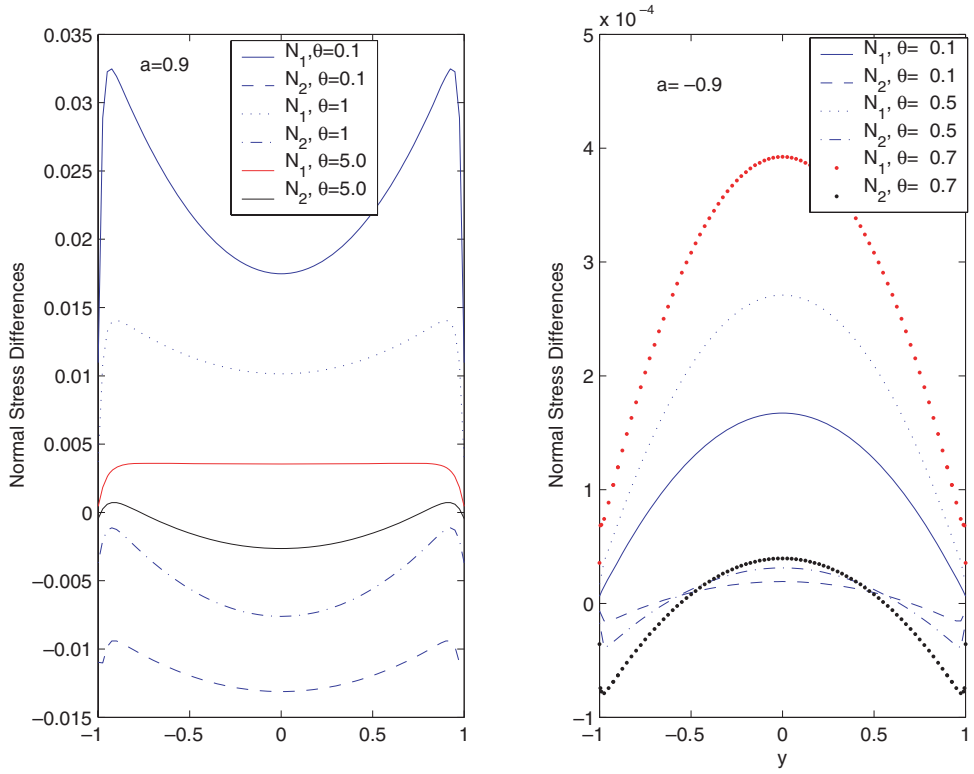


FIG. 9. *The normal stress differences N_1 and N_2 as functions of y in flowing-aligning regime at selected values of θ with normally anchored boundary conditions in weak shear. The parameter values are $a = \pm 0.9, N = 6, De = 0.01, Er = 100, \alpha = 10, \mu_1 = \pm 0.01, \mu_2 = 0.02$ for rods and discs, respectively.*

TABLE 3
The averaged normal stress differences (Couette).

	FA/rods	FA/discs	T/rods	T/discs
Tangential	$N_1 < 0, N_2 > 0$	$N_1 > 0, N_2 < 0$	$N_1 > 0, N_2 < 0$	$N_1 > 0, N_2 < 0$
Normal	$N_1 > 0, N_2 < 0$	$N_1 > 0$	$N_1 > 0, N_2 < 0$	$N_1 < 0, N_2 < 0$

- The major director winds counterclockwise from the bottom to top plates, for both flow-aligning rods in tangential anchoring and flow-aligning discs in homeotropic anchoring. Remarkably, the principal orientation axis rotates clockwise if the nematic polymer tumbles in weak shear rather than flow aligns. The magnitude of winding of the orientation axis, which sets the number of bands of nematic distortion, reduces with the degree of elastic anisotropy θ.
- The order parameters are relatively insensitive to the degree of elastic anisotropy in tangential anchoring, and more sensitive in normal anchoring.
- Ill-posedness may occur within each order of asymptotic equations depending on the values of the parameters, although the full equations are well posed. This transition implies a breakdown in the asymptotic ordering which allows explicit solution and scaling properties, and suggests a physical transition away from these asymptotic structures.

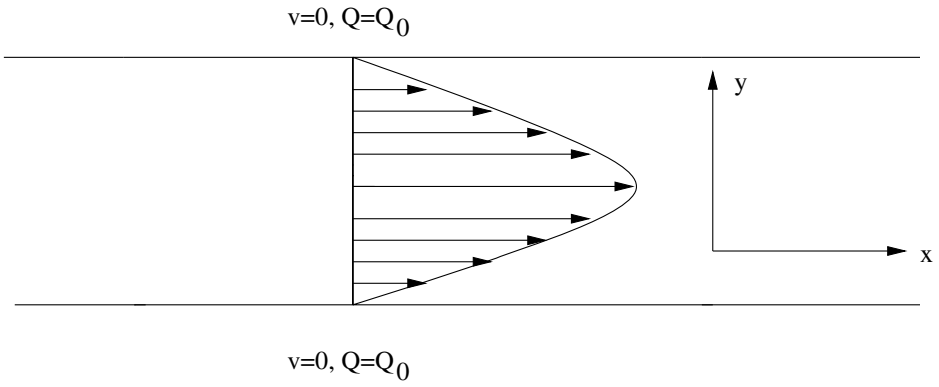


FIG. 10. *The geometry of the plane Poiseuille flow. A pressure gradient* $\frac{\partial p}{\partial x} = -De^2$ *is imposed across the channel. At the bounding surfaces, the orientation tensor is assumed to equal to its equilibrium value.*

- The averaged shear viscosity varies weakly with respect to the strength of anisotropic elasticity.
- The averaged normal stress differences may take on all possible signs depending on the parameter regime.

3. Spatial structures in weak Poiseuille flows. In this section, we study steady structures in the direction of the velocity gradient under an imposed, small pressure gradient $\frac{\partial p}{\partial x}$ in plane Poiseuille flow. Figure 10 depicts the cross section of the plane Poiseuille flow on the (x, y) plane. The boundary condition for the orientation tensor is identical to that used for shear flows while the velocity boundary condition is no-slip $\mathbf{v}(\pm 1) = \mathbf{0}$. As before, we use h and $t_0 = t_n = \frac{1}{D_r}$ as the characteristic length and time scale, respectively. We adopt the same dimensionless symbols used in weak plane Couette and assume $\frac{\partial p}{\partial x} = -\epsilon = -De^2$ in the dimensionless form, where the Deborah number is defined by

$$De = \sqrt{-\frac{\partial p}{\partial x}\frac{t_0^2}{\rho h}}, \qquad \epsilon = De^2. \tag{55}$$

We seek asymptotic solutions in powers of ϵ. The momentum equation yields at order $O(\epsilon)$:

$$\frac{\partial v_x^{(1)}}{\partial t} = -1 + \frac{\partial \tau_{xy}^{(1)}}{\partial y}. \tag{56}$$

The other governing equations are identical to those derived for plane Couette flows. Hence, the transient solution and the stability of steady states are identical to the corresponding problems in plane Couette flows. We will not repeat them here; instead, we only present the asymptotic steady states with respect to the two anchoring conditions.

3.1. Tangential anchoring ($\psi_0 = 0$). The steady solutions up to order $O(\epsilon)$ are

$$s_0 = s_0, \quad \beta_0 = 0, \quad v_x^{(1)}(y) = H_1(1 - y^2), \quad \psi^{(1)}(y) = H_2 Ery(1 - y^2), \tag{57}$$

where H_1 and H_2 are given in Appendix E. The positivity of H_1 coincides with the stability of the steady state, giving rise to a parabolic velocity profile. H_2 is positive for flow aligning rods and negative otherwise in stable steady states. H_2 behaves more or less like the diagnostic M in the plane Couette flow. Notice that $\psi^{(1)}$ is an odd function of y leading to an asymmetric major director pattern, known as a chevron pattern, with respect to the midplane [6, 2, 3].

For flow-aligning rods in stable steady states, H_1 (H_2) decreases (increases) with respect to θ, and increases (decreases) with respect to θ in all other cases. The rotational pattern of the major director (a function of $\psi^{(1)}$) is dictated by λ_L. For flow-aligning rods, $\lambda_L > 1$, the major director rotates counterclockwise from the lower plate to the $\frac{\sqrt{3}}{6}$ of the shear cell and then reverses its rotation to the midplane. The orientation pattern in the top half of the cell is the mirror image of that in the lower half. The rotation reverses for the other cases where $\lambda_L < 1$.

The steady solutions of the order parameters at order $O(\epsilon^2)$ are given by

$$\beta_2(y) = K_{11}\left(\frac{\cosh(\sqrt{Er}\Lambda y)}{\cosh(\sqrt{Er}\Lambda)} - 1\right) + K_{12}\left(\frac{\cosh(\sqrt{Er}\Gamma y)}{\cosh(\sqrt{Er}\Gamma)} - 1\right) + R_{11}(y^4 - 1) + R_{12}(y^2 - 1),$$

$$s_2(y) = K_{21}\left(\frac{\cosh(\sqrt{Er}\Lambda y)}{\cosh(\sqrt{Er}\Lambda)} - 1\right) + K_{22}\left(\frac{\cosh(\sqrt{Er}\Gamma y)}{\cosh(\sqrt{Er}\Gamma)} - 1\right) + R_{21}(y^4 - 1) + R_{22}(y^2 - 1), \tag{58}$$

where the coefficients are given in Appendix E. The order parameters behave like a quartic polynomial with respect to y in most part of the cell except at the boundary layers near the plates. The velocity $v_x^{(1)}$, angle variable $\psi^{(1)}$ and the order parameter β_2 are insensitive to the variation of θ. The sensitivity of the order parameter correction s_2 is the most pronounced at $\theta = 0$, i.e., in the one-constant approximation. Figure 11 depicts typical steady solutions for tumbling rods. Table 4 tabulates all the steady state behavior:

- The thickness of the boundary layers in this flow are narrower than those in the weak plane Couette flows, suggesting a mollifying effect of stronger velocity gradients near the walls.
- The two order parameter corrections (at small $|\theta|$) and the angle variable change their signs and concavity in the tumbling versus flow-aligning regime, but only for rods; this predicts a profile concavity flip in the focusing and defocusing of the orientation distribution occurs as rods pass through the flow-aligning to tumbling transition; the profile of β_2 is either W-shaped or M-shaped.
- The velocity profile is concave down.
- The parameter $s - s_0$ is very sensitive around $\theta = 0$ for both rods and discs.
- The angle profile is a rotated-S shape.

Figure 11 depicts typical steady solutions for tumbling discs as functions of (θ, y).

The shear viscosity is given by

$$\eta_{app} = \frac{\tau_{xy}}{\frac{dv_x^{(1)}}{dy}} = -\frac{\alpha H_2 s_0}{18 H_1}\left[a(s_0 + 2)(\theta(1 - s_0) + 3(1 - \tfrac{1}{\lambda_L}))\right] + \tfrac{1}{3}(\mu_1 s_0 + 3\eta), \tag{59}$$

a decreasing function with respect to θ for discs and tumbling rods but an increasing function for flow-aligning rods.

The normal stress differences in this case are given in Appendix E. Unlike weak Couette flows, they are rational functions of μ_1, μ_2 and α. Figures 12 and 13

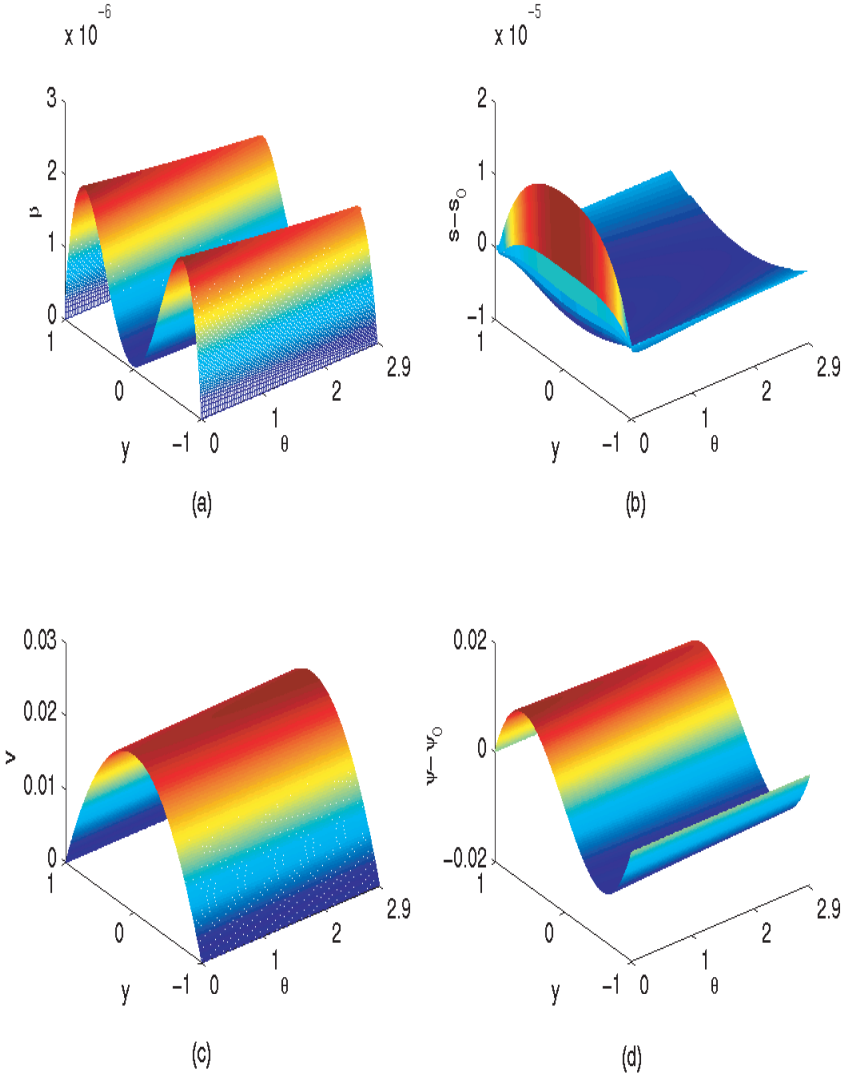


FIG. 11. *Steady solutions of* $\beta, s-s_0, v$ *and* $\psi-\psi_0$ *as functions of* (θ, y) *in the regime of tumbling rods with tangentially anchored boundary conditions in plane Poiseuille flow. The parameter values are* $N = 6, a = 0.8, De = 0.01, Er = 100, \mu_1 = 0.1, \eta = 0.2, \alpha = 10$.

depict some representative plots of the normal stress differences for tumbling and flow-aligning nematics, respectively. In summary, they have the following properties:

- For flow-aligning rods, N_1 is positive and N_2 is negative. The signs are reversed for flow-aligning discs.
- For tumbling rods and discs, N_1 is negative, but N_2 is positive.
- In both stable tumbling and flow-aligning regimes, for rods and discs, the absolute values of N_1 and N_2 increase and decrease, respectively, as α increases.

TABLE 4
Steady states (Poiseuille) in tangential anchoring.

	FA/rods	FA/discs	T/rods	T/discs
$s - s_0$	Concave up	Concave up	Concave down at small θ & Concave up at large θ	Concave up
β	W-shape	M-shape	M-shape	M-shape
$\psi - \psi_0$	S	Rotated-S	Rotated-S	Rotated-S

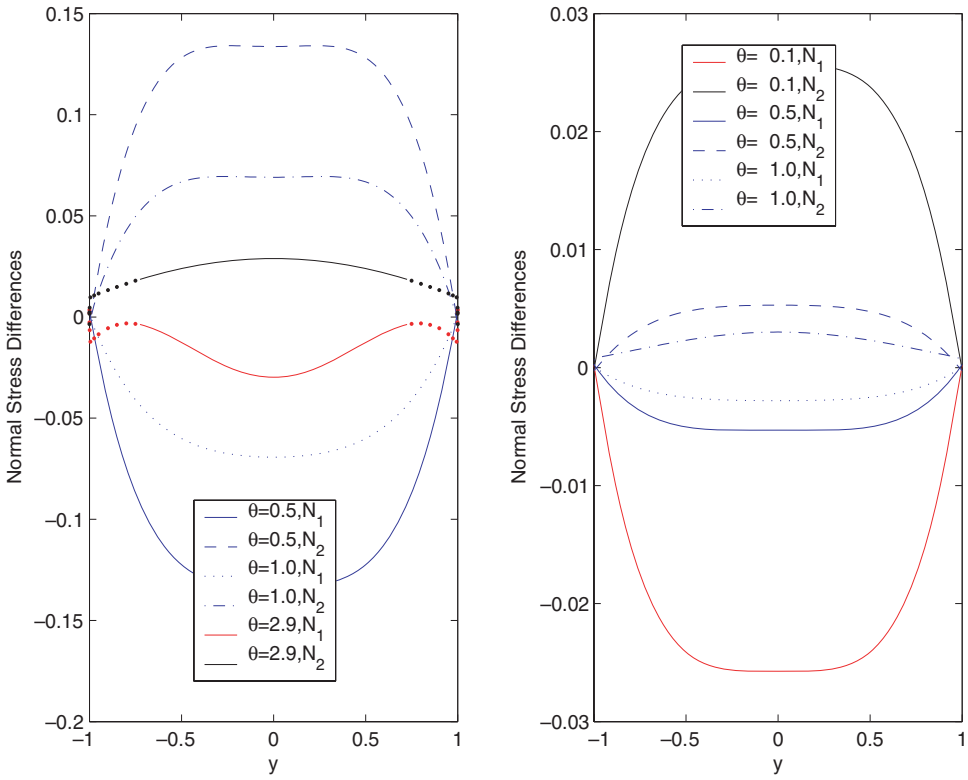


FIG. 12. *The normal stress differences N_1 and N_2 as functions of y in tumbling regime at selected values of θ with tangentially anchored boundary conditions in plane Poiseuille flows. The other parameter values are $N = 6, De = 0.01, Er = 100, \alpha = 10, \mu_2 = 0.02$, for rods $a = 0.8, \mu_1 = 0.01$, and for discs, $a = -0.8, \mu_1 = -0.01$.*

- The gap averages satisfy $N_1 > 0$ and $N_2 < 0$ for flow-aligning rods while $N_1 < 0, N_2 > 0$ in all other regimes.

The behavior of the averaged normal stress differences is tabulated in Table 6.

3.2. Homeotropic anchoring ($\psi = \frac{\pi}{2}$). The steady solutions up to order $O(\epsilon)$ are given by (57) with new H_1 and H_2 given in Appendix F. As in the tangential anchoring case, H_1 is positive in all stable steady states. H_2 is negative for flow-aligning discs and positive otherwise for other stable steady states. For flow-aligning discs in stable steady states, H_1 (H_2) decreases (increases) with respect to θ for flow-aligning discs, yet increases (decreases) with respect to θ for all other stable steady states.

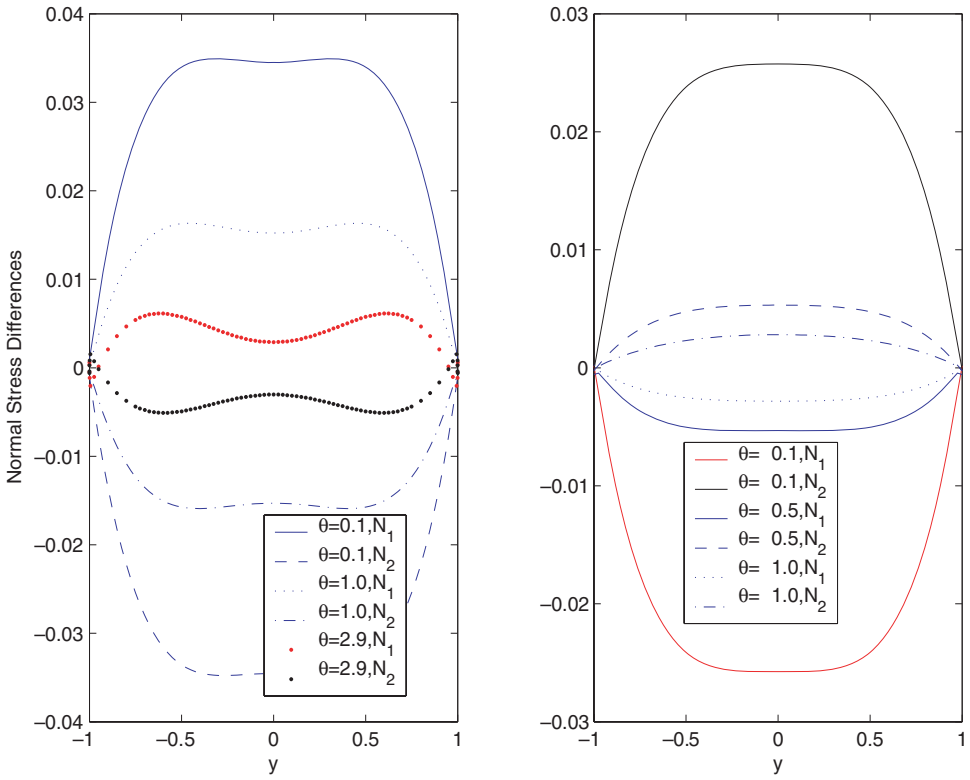


FIG. 13. *The normal stress differences N_1 and N_2 as functions of y in flow-aligning regime at some values of θ with tangentially anchored boundary conditions in plane Poiseuille flows. The other parameter values are $N = 6, De = 0.01, Er = 100, \alpha = 10, \mu_2 = 0.02$, for rods $a = 0.9, \mu_1 = 0.01$, and for discs, $a = -0.9, \mu_1 = -0.01$.*

TABLE 5
Steady states (Poiseuille) in homeotropic anchoring.

	FA/rods	FA/discs	T/rods	T/discs
$s - s_0$	W-shape	M-shape	W-shape	W-shape
β	W-shape	M-shape	W-shape	W-shape
$\psi - \psi_0$	Rotated-S	S	Rotated-S	Rotated-S

The steady solutions of the order parameters at $O(\epsilon^2)$ are

$$\beta_2(y) = K_{11}\left(\frac{\cosh(\sqrt{Er}\Lambda y)}{\cosh(\sqrt{Er}\Lambda)} - 1\right) + R_1(y^4 - 1) + S_1(y^2 - 1),$$

$$s_2(y) = K_{21}\left(\frac{\cosh(\sqrt{Er}\Lambda y)}{\cosh(\sqrt{Er}\Lambda)} - 1\right) \tag{60}$$

$$+K_{22}\left(\frac{\cosh(\sqrt{Er}\Gamma y)}{\cosh(\sqrt{Er}\Gamma)} - 1\right) + R_2(y^4 - 1) + S_2(y^2 - 1).$$

We summarize the features of stable steady states in Table 5:

- The two order parameters and the angle parameter change their signs and shapes in tumbling and flow-aligning regime for discs but not for rods, indicating the solutions are more sensitive for platelet molecules than for rods.

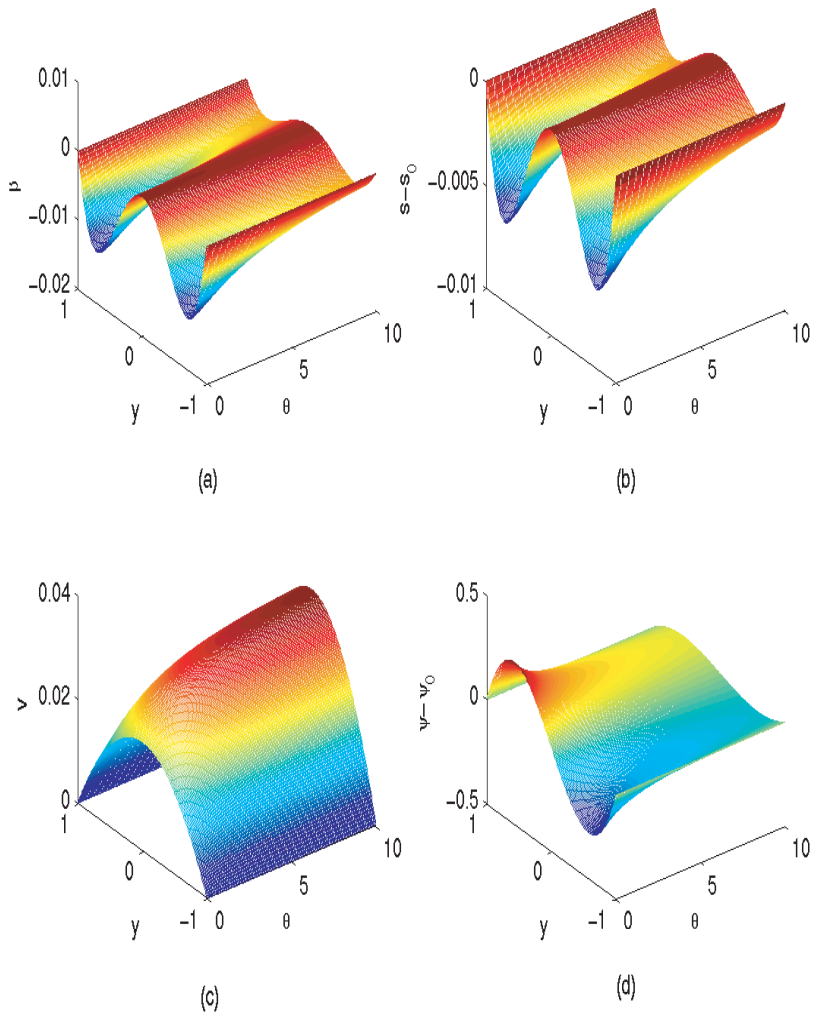


FIG. 14. *Steady solutions* $\beta, s - s_0, v$ *and* $\psi - \psi_0$ *for tumbling rods as functions of* (θ, y) *with normally anchored boundary conditions in plane Poiseuille flows. The parameter values are* $a = 0.8, N = 6, De = 0.01, Er = 100, \mu_1 = 0.01, \eta = 0.02, \alpha = 1$.

- As the anisotropic elasticity enhances, the order parameter variations and the angle variation reduces for rods yet amplifies for discs.
- The velocity profile has fixed concavity in all regimes. As the anisotropic elasticity increases, the velocity increases for rods.
- The orientational variables decrease with respect to θ while the flow variable v increases.
- Again, the thickness of the boundary layers are smaller compared to weak plane Couette flows.

Figure 14 depicts a typical steady solution for tumbling discs as functions of (θ, y).

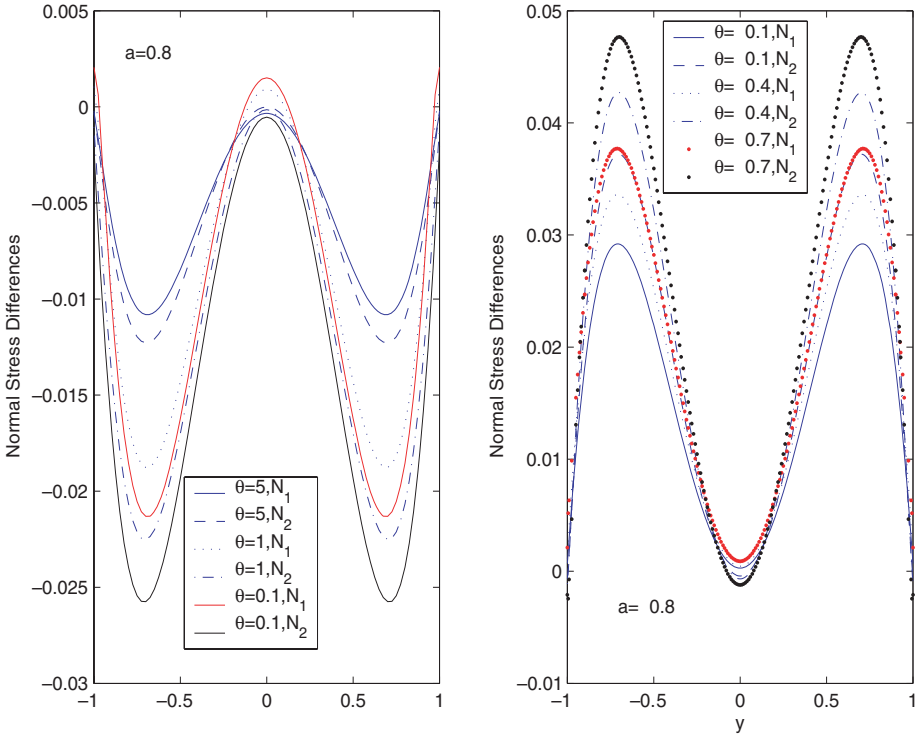


FIG. 15. *The normal stress differences N_1 and N_2 as functions of y in tumbling regime at selected values of θ with normal anchoring boundary conditions in plane Poiseuille flows. The parameter values are $N = 6, De = 0.01, Er = 100, \alpha = 10,$, $a = 0.8, \mu_1 = 0.01, \mu_2 = 0.02$ for rods, and $a = -0.8, \mu_1 = -0.01, \mu_2 = 0.02$ for discs.*

The wall shear viscosity is given by

$$\eta_{wall} = \frac{\tau_{xy}}{\frac{dv_x^{(1)}}{dy}} = \frac{\alpha H_2}{18 Er H_1}(a(2+s_0)(\theta(2s_0+1)+3(1+\tfrac{1}{\lambda_L}))) + \tfrac{1}{3}(\mu_1 s_0 + 3\eta), \tag{61}$$

which decays with respect to θ for all rods and flow-aligning discs, but increases for tumbling discs. The first and second normal stress differences in this case are given in Appendix F. Figures 15 and 16 depict the normal stress differences for flow-aligning and tumbling nematics, respectively. In summary, they exhibit the following features:

- For flow-aligning rods and disks, N_1 and N_2 are negative except for a small region at the midplane, where some of the first normal stress difference may be positive.
- For tumbling rods, N_1 and N_2 are negative. For tumbling discs ($-1 < \lambda_L < 0$), N_1 and N_2 are negative except for a tiny region at the midplane at small θ. For tumbling discs, the behavior reverses completely, i.e., the normal stress differences are positive except for a small region at the midplane.
- In both stable tumbling and flow-aligning regimes, for rods and discs, the absolute values of N_1 and N_2 will decrease and increase as α increases, respectively.
- The gap averages obey $N_1 > 0$ and $N_2 > 0$ for tumbling rods and both become negative in all other regimes.

The behavior of averaged normal stress differences is tabulated in Table 6.

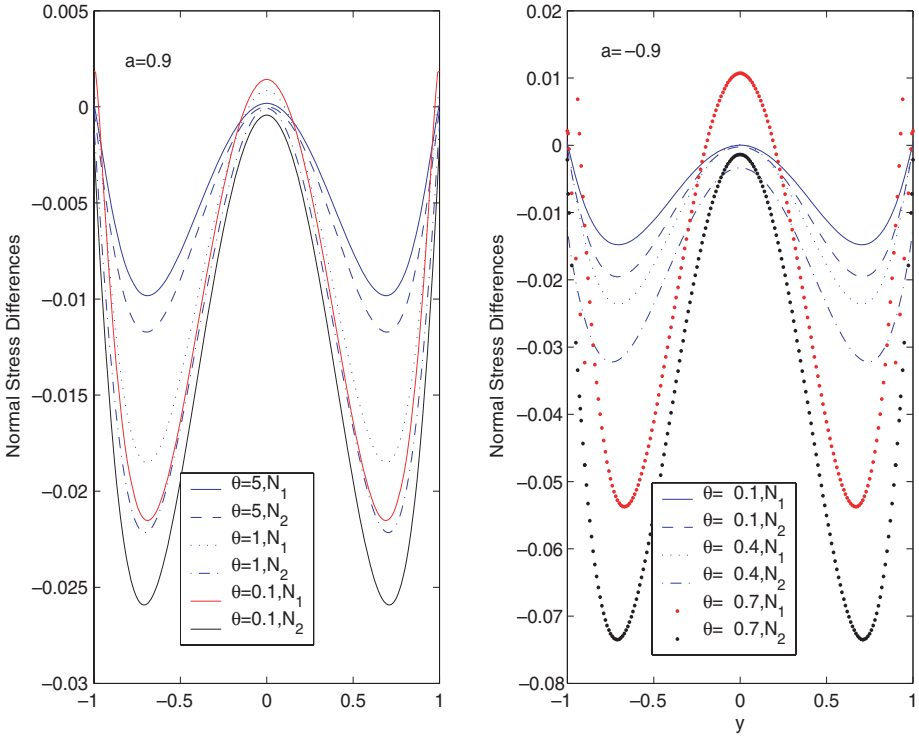


FIG. 16. *The normal stress differences N_1 and N_2 as functions of y in flow-aligning regime at selected values of θ with normal anchoring boundary conditions in plane Poiseuille flows. The parameter values are $N = 6, De = 0.01, Er = 100, \alpha = 10, a = 0.9, \mu_1 = 0.01, \mu_2 = 0.02$ for rods, and $N = 6, De = 0.01, Er = 100, \alpha = 1, a = -0.9, \mu_1 = -0.01, \mu_2 = 0.02$ for discs.*

TABLE 6
The normal stress differences (Poiseuille).

	FA/rods	FA/discs	T/rods	T/discs
Tangential	$N_1 > 0, N_2 > 0$	$N_1 < 0, N_2 > 0$	$N_1 < 0, N_2 > 0$	$N_1 < 0, N_2 > 0$
Normal	$N_1 < 0, N_2 < 0$	$N_1 < 0, N_2 < 0$	$N_1 < 0, N_2 < 0$	$N_1 > 0, N_2 > 0$

4. Conclusion. We have derived explicit asymptotic structures for weakly sheared nematic polymers in both plate-driven and pressure-driven experimental conditions. The goal of this analysis is to predict scaling properties in the orientational distribution of the rigid rod ensemble from the strong elasticity, weak flow regime, which then guide numerical continuation studies of heterogeneous films and molds across a multi-parameter space of material properties and processing conditions. We have explored the effect of anisotropic elasticity for both flow conditions, using a second-moment model for the orientational distribution derived from Doi–Hess–Marrucci–Greco kinetic theory. These results extend previous work of the authors for steady structures with equal elasticity constants in plate-driven flow in several ways: anisotropic elasticity, pressure-driven flows, and transient asymptotic equations followed by stability predictions within the asymptotic ordering of the flow-nematic system. The leading order flow structure is simple linear shear versus Poiseuille profiles for the respective driving conditions, with elastic hydrodynamic feedback contributions characterized at next order. These results confirm the consistency of imposing

the kinematics (and thereby suppressing flow feedback) in special asymptotic parameter regimes, yet also predict breakdown of this decoupling of the momentum equation when either of several conditions is relaxed: weak flow, strong elasticity, or sufficiently isotropic elasticity.

The orientational structures for both flow conditions convey scaling properties of nematic (director dominated) elastic distortions as well as molecular elasticity (dominated by focusing or defocusing of the orientational distribution). The structure scaling laws are similar for plane Couette and Poiseuille flows, with plate boundary layer thicknesses proportional to $1/\sqrt{Er}$ and nonuniform structures spanning the plates with mean lengthscale proportional to $1/Er$. The prefactors of the structures capture the roles of material properties: flow-aligning versus tumbling nematics, degree of anisotropy in the elasticity potential, strength of the short-range nematic potential, and molecular aspect ratio. These subtleties are detailed in the body of the paper, where the amplitude of structure variations, convexity of profiles, and stability of the steady structures all depend strongly on these molecular parameters as well as plate anchoring conditions.

The particular results are less important than the overall insight into the sensitivity, flow-nematic feedback and processing-generated structures on material and device properties. The instability within the asymptotic equations is catastrophic, similar to backward heat flow instabilities, so there is no mistaking the breakdown of these steady profiles in the weak flow model system. The role of anisotropic elasticity is shown to be greater for normal versus tangential plate anchoring, and greater in pressure-driven than plate-driven flows. In Poiseuille flow, the transition to catastrophic instability coincides with a rapidly growing midplane axial velocity, confirming a breakdown in the asymptotic analysis. Finally, anisotropic elasticity is shown to contribute to either shear thinning or shear thickening behavior as other parameters are modified, and signs of normal stress differences (which determine whether the plates are pushed or pulled by the stresses generated between) are likewise sensitive to various material parameters.

Appendix A. The coefficients in the second-order equations in tangential anchoring.

$$
\begin{aligned}
&A_1 = (1-s_0)[-6(1+2s_0) + (8s_0^2 - s_0 - 2)\theta], \qquad B_1 = (1-s_0)s_0(6 - 7s_0\theta),\\
&C_1 = 18ErNs_0(-1+4s_0), \qquad D_1 = 36ErNs_0(1-s_0),\\
&E_1 = 2s_0(1-s_0)(1+3s_0)[6 + (2-5s_0)\theta],\\
&F_1 = 2s_0[3(2+s_0) + 2(3s_0^2 + s_0 - 1)\theta], \qquad G_1 = 9Ers_0(-1-a+2as_0),\\
&A_2 = -2s_0\theta(1-s_0), \quad B_2 = (1-s_0)[-6 + (3s_0-2)\theta], \quad C_2 = 0,\\
&D_2 = 54ErNs_0, E_2 = 2s_0(s_0-1)[6 + (2-5\theta)],\\
&F_2 = 2s_0[3(2-s_0) + (s_0+2)(s_0-1)\theta], \qquad G_2 = 9Ers_0(1-a).
\end{aligned} \tag{62}
$$

Appendix B. The coefficients in the second-order equations in normal anchoring.

$$
\begin{aligned}
&A_1 = -2(1-s_0)(2s_0+1)[3 + (4s_0+1)\theta],\\
&B_1 = 2s_0(1-s_0)[3 + 2(2s_0+1)\theta],\\
&C_1 = 18ErNs_0(-1+4s_0), \qquad D_1 = 36ErNs_0(1-s_0),\\
&E_1 = 4s_0(1-s_0)[3(1+3s_0) - (2s_0+1)(6s_0+1)\theta)],\\
&F_1 = 2s_0[3(2+s_0) - (6s_0^3 - 5s_0^2 - 8s_0 - 2)\theta], \qquad G_1 = 9Ers_0(-1+a-2as_0),\\
&A_2 = 0, \quad B_2 = 2(s_0-1)[3 + (2s_0+1)\theta], \quad C_2 = 0, \quad D_2 = 54ErNs_0,\\
&E_2 = 4s_0[3(-1+s_0) + (2s_0^2 + 3s_0 - 1)\theta],\\
&F_2 = 2s_0[-3(2+s_0) + (2s_0+1)(s_0-2)\theta], \qquad G_2 = 9Ers_0(1+a).
\end{aligned} \tag{63}
$$

Appendix C. The coefficients of the steady solutions for tangential anchoring in weak plane Couette flows.

$$\lambda,\mu=\sqrt{Er}\{18[Ns_0(s_0-1)((16{s_0}^2-5s_0-2)\theta-6(5s_0+1))\mp$$
$$\sqrt{2}((68{s_0}^4+49{s_0}^3-24{s_0}^2-20s_0+8)\theta^2-6(8{s_0}^3+21{s_0}^2+6s_0-8)\theta+$$
$$18(s_0+2)^2)^{1/2}s_0(1-s_0)N]/[(s_0-1)^2(2s_0+1)(5\theta s_0-2\theta-6)(\theta s_0-2\theta-6)]\}^{\frac{1}{2}},$$
$$R_1=\tfrac{(as_0-3s_0+2a)(s_0-1)[(-5s_0^2+2as_0^2+3as_0+2s_0-2a)\theta-3(as_0-2s_0+2a)]}{4Ns_0^2(s_0+2)^2(\theta s_0-\theta-3)^2},$$
$$S_1=(1-s_0)[(-88s_0^3+34as_0+77as_0^2-31s_0^2+8as_0+20s_0-20a)\theta-6(7as_0^2-14s_0^2$$
$$-10s_0+19as_0+10a)](as_0-3s_0+2a)/[8s_0^2N(-1+4s_0)(s_0+2)^2(\theta s_0-\theta-3)^2],$$
$$\Lambda=\tfrac{\lambda}{\sqrt{Er}},\Gamma=\tfrac{\mu}{\sqrt{Er}},R=R_1Er,S=S_1Er,K=\tfrac{3(as_0-3s_0+2a)Er}{4s_0(s_0+2)[(s_0-1)\theta-3]},$$
$$A=-(s_0-1)^2(2s_0+1)(5\theta s_0-2\theta-6)(\theta s_0-2\theta-6),$$
$$B=36ErNs_0(s_0-1)[(16s_0^2-5s_0-2)\theta-6(5s_0+1)],\quad C=972(ErNs_0)^2(1-4s_0),$$
$$T=\tfrac{2A_1(4K^2E_2+2K^2F_2+KG_2)-2A_2(4K^2E_1+2K^2F_1+KG_1)-C_1(2K^2F_2+KG_2)-2BR}{C},$$
$$K_1=\tfrac{1}{\lambda^2-\mu^2}[\tfrac{4K^2A_1E_2-4K^2A_2E_1}{A}+\mu^2(R+T)-2R],\qquad K_3=-\tfrac{B_2\lambda^2+D_2}{A_2\lambda^2}K_1,$$
$$K_2=\tfrac{1}{-\lambda^2+\mu^2}[\tfrac{4K^2A_1E_2-4K^2A_24E_1}{A}+\lambda^2(R+T)-2R],\qquad K_4=-\tfrac{B_2\mu^2+D_2}{A_2\mu^2}K_2,$$
$$N_1=\tfrac{G\alpha}{54Er}\left(\tfrac{K_3\lambda^2\cosh(\lambda y)}{\cosh(\lambda)}+\tfrac{K_4\mu^2\cosh(\mu y)}{\cosh\mu}+2S\right)$$
$$+\tfrac{aNs_0(1-4s_0)\alpha}{3}\left[K_3\left(\tfrac{\cosh(\lambda y)}{\cosh(\lambda)}-1\right)+K_4\left(\tfrac{\cosh(\mu y)}{\cosh(\mu)}-1\right)+S\left(y^2-1\right)\right]$$
$$+\tfrac{(-4a\theta+3s_0+16a\theta s_0^3-12a+12as_0^2-12a\theta s_0)\alpha}{108Er}\left(\tfrac{K_1\lambda^2\cosh(\lambda y)}{\cosh(\lambda)}+\tfrac{K_2\mu^2\cosh(\mu y)}{\cosh(\mu)}+2R\right)$$
$$+\tfrac{Ns_0(1+2s_0)a\alpha}{3}\left[K_1\left(\tfrac{\cosh(\lambda y)}{\cosh(\lambda)}-1\right)+K_2\left(\tfrac{\cosh(\mu y)}{\cosh(\mu)}-1\right)+R\left(y^2-1\right)\right]$$
$$+\tfrac{4K^2s_0\alpha}{27Er}[4a(s_0-1)(2s_0+1)(3s_0+1)\theta-3(2as_0-3s_0-6as_0{}^2+4a]y^2$$
$$+\tfrac{2as_0K^2\alpha}{27Er}[(2s_0+1)(3s_0{}^2-5s_0-4))\theta-6(s_0+2)](y^2-1)+2\mu_2Ks_0{}^2\left(y^2-1\right),$$
$$N_2=-\tfrac{G\alpha}{54Er}\left(\tfrac{K_3\lambda^2\cosh(\lambda y)}{\cosh(\lambda)}+\tfrac{K_4\mu^2\cosh(\mu y)}{\cosh\mu}+2S\right)$$
$$+\tfrac{aNs_0(-1+4s_0)\alpha}{3}\left[K_3\left(\tfrac{\cosh(\lambda y)}{\cosh(\lambda)}-1\right)+K_4\left(\tfrac{\cosh(\mu y)}{\cosh(\mu)}-1\right)+S\left(y^2-1\right)\right]$$
$$+\tfrac{(8a\theta+12a-3+8as_0-16as_0^2-12s_0)s_0\alpha}{108Er}\left(\tfrac{K_1\lambda^2\cosh(\lambda y)}{\cosh(\lambda)}+\tfrac{K_2\mu^2\cosh(\mu y)}{\cosh(\mu)}+2R\right)$$
$$+\tfrac{2Ns_0(1-s_0)a\alpha}{3}\left[K_1\left(\tfrac{\cosh(\lambda y)}{\cosh(\lambda)}-1\right)+K_2\left(\tfrac{\cosh(\mu y)}{\cosh(\mu)}-1\right)+R\left(y^2-1\right)\right]$$
$$+\tfrac{4K^2s_0\alpha}{27Er}[-2a(s_0-1)(2s_0+1)(6s_0+1)\theta+3(4as_0-3s_0-6as_0{}^2+2a)]y^2$$
$$+\tfrac{2as_0K^2\alpha}{27Er}[-(2s_0+1)(3s_0{}^2-4s_0-2))\theta+3(s_0+2)](y^2-1)$$
$$-2\left(\mu_1+\mu_2\right)Ks_0{}^2\left(y^2-1\right),$$
$$G=2a\theta+4a\theta s_0^2+6as_0-3s_0-12as_0^2-16a\theta s_0^3+10a\theta s_0+6a.$$
(64)

Appendix D. The coefficients of the steady solutions in normal anchoring in weak plane Couette flows.

$$\lambda = \sqrt{Er}[\tfrac{3Ns_0}{(1-s_0)(2\theta s_0+3+\theta)}]^{1/2}, \mu = \sqrt{Er}[\tfrac{9Ns_0(-1+4s_0)}{(1-s_0)(2s_0+1)(4\theta s_0+3+\theta)}]^{1/2}, \Lambda = \tfrac{\lambda}{\sqrt{Er}}, \Gamma = \tfrac{\mu}{\sqrt{Er}},$$

$$R = \tfrac{9(1-s_0)(2s_0+as_0+2a)(as_0+3s_0+2a)}{4N(s_0+2)^2(2\theta s_0+3+\theta)s_0^2}Er, \quad R_2 = \tfrac{R}{Er}, \quad K = \tfrac{3(as_0+3s_0+2a)}{4s_0(s_0+2)(2\theta s_0+\theta+3)}Er,$$

$$T = 3(-1+s_0)[2(2s_0+1)(s_0-1)(as_0+2s_0+2a)\theta + ErNs_0a(s_0+2)$$

$$+6(-1+s_0)(2s_0+as_0+2a))](as_0+3s_0+2a)/[4N^2(s_0+2)^2(2\theta s_0+3+\theta)s_0^3],$$

$$S = \{3(as_0+3s_0+2a)(1-s_0)[(2s_0+1)(-3s_0^2+as_0^2-4as_0-12s_0-12a)\theta$$

$$+18(s_0-1)(as_0+2s_0+2a)]/[8Ns_0^2(s_0+2)^2(2\theta s_0+3+\theta)^2(-1+4s_0)]\}Er,$$

$$S_2 = \tfrac{S}{Er}, \qquad T_1 = -\tfrac{A_1S+2B_1R+D_1T+(-2K^2F_1-G_1K)}{C_1},$$

$$K_{11} = -R-T, \quad K_{21} = \tfrac{B_1\lambda^2+D_1}{A_1\lambda^2+C_1}K_{11}, \quad K_{22} = -K_{21}-S-T_1,$$

$$N_1 = \tfrac{G\alpha}{54Er}\left(\tfrac{K_{21}\lambda^2\cosh(\lambda\,y)}{\cosh(\lambda)} + \tfrac{K_{22}\mu^2\cosh(\mu\,y)}{\cosh\mu} + 2\,S\right)$$

$$+\tfrac{aNs_0(1-4\,s_0)\alpha}{3}\left[K_{21}\left(\tfrac{\cosh(\lambda\,y)}{\cosh(\lambda)}-1\right)+K_{22}\left(\tfrac{\cosh(\mu\,y)}{\cosh(\mu)}-1\right)+S\left(y^2-1\right)\right]$$

$$+\tfrac{(-4a\theta+3s_0+16a\theta s_0^3-12a+12as_0^2-12a\theta s_0)\alpha}{108Er}\left(\tfrac{K_{11}\lambda^2\cosh(\lambda\,y)}{\cosh(\lambda)}+2\,R\right)$$

$$+\tfrac{Ns_0(1+2\,s_0)\,a\alpha}{3}\left[K_{11}\left(\tfrac{\cosh(\lambda\,y)}{\cosh(\lambda)}-1\right)+R\left(y^2-1\right)\right]$$

$$+\tfrac{4K^2s_0\alpha}{27Er}[4a(s_0-1)(2s_0+1)(3s_0+1)\theta+3(-2as_0+3s_0+6as_0{}^2-4a)]y^2$$

$$+\tfrac{2as_0K^2\alpha}{27Er}\left[(2s_0+1)(3s_0^2-5s_0-4)\theta-6(s_0+2)\right](y^2-1)+2\,\mu_2\,Ks_0{}^2\left(y^2-1\right),$$

$$N_2 = -\tfrac{G\alpha}{54Er}(\tfrac{K_{21}\lambda^2\cosh\lambda\,y}{\cosh(\lambda)}+\tfrac{K_{22}\mu^2\cosh\mu\,y}{\cosh\mu}+2\,S)$$

$$+\tfrac{aNs_0(-1+4\,s_0)\alpha}{3}\left[K_{21}\left(\tfrac{\cosh(\lambda\,y)}{\cosh(\lambda)}-1\right)+K_{22}\left(\tfrac{\cosh(\mu\,y)}{\cosh(\mu)}-1\right)+S\left(y^2-1\right)\right]$$

$$+\tfrac{(8a\theta+12a-3+8as_0-16as_0^2-12s_0)s_0\alpha}{108Er}\left(\tfrac{K_{11}\lambda^2\cosh(\lambda\,y)}{\cosh(\lambda)}+2\,R\right)$$

$$+\tfrac{2Ns_0(1-s_0)a\alpha}{3}\left[K_{11}\left(\tfrac{\cosh(\lambda\,y)}{\cosh(\lambda)}-1\right)+R\left(y^2-1\right)\right]$$

$$+\tfrac{4K^2s_0\alpha}{27Er}[2a(1-s_0)(2s_0+1)^2)\theta+3(4as_0-9s_0-6as_0{}^2+2a)y^2$$

$$+\tfrac{2as_0K^2\alpha}{27Er}[-(2s_0+1)(3s_0^2-4s_0-2)\theta+3(s_0+2)](y^2-1)$$

$$-2\left(\mu_1+\mu_2\right)Ks_0{}^2\left(y^2-1\right),$$

$$G = 2a\theta+4a\theta s_0^2+6as_0-3s_0-12as_0^2-16a\theta s_0^3+10a\theta s_0+6a.$$

(65)

Appendix E. The coefficients of the steady solutions for tangential anchoring in Poiseuille flows.

$$H_1 = \frac{6(s_0+2)(3+\theta(1-s_0))}{4(\mu_1 s_0+3\eta)(s_0+2)(3+\theta(1-s_0))+\alpha s_0(1-\lambda_L)(-a\theta(1-s_0)(s_0+2)+9s_0(1-\lambda_L))},$$

$$H_2 = \frac{9(\lambda_L-1)}{4(\mu_1 s_0+3\eta)(s_0+2)(3+\theta(1-s_0))+\alpha s_0(1-\lambda_L)(-a\theta(1-s_0)(s_0+2)+9s_0(1-\lambda_L))},$$

$$D = -C_1(9H_2^2E_2 + 6H_2^2F_2 + 2G_2H_1H_2),$$

$$E = -12A_1(9H_2^2E_2 + 6H_2^2F_2 + 2G_2H_1H_2) + C_1(6H_2^2E_2 + 6H_2^2F_2 + 2G_2H_1H_2)$$
$$+12A_2(9H_2^2E_1 + 6H_2^2F_1 + 2G_2H_1H_2),$$

$$F = 2A_1(6H_2^2E_2 + 6H_2^2F_2 + 2G_2H_1H_2) - C_1H_2^2E_2 + 6H_2^2E_2$$
$$-2A_2(6H_2^2E_1 + 6H_2^2F_1 + 2G_2H_1H_2),$$

$$R_{11} = -\frac{D}{C}, \quad R_{12} = -\frac{12BR_{11}+E}{C}, \quad R_{13} = -\frac{24AR_{11}+2BR_{12}+F}{C},$$

$$K_{11} = \frac{1}{(\lambda^2-\mu^2)}[\frac{4H_2^2(E_2A_1-E_1A_2)}{A} - 12.R_{11} - 2R_{12} + (R_{11}+R_{12}+R_{13})\mu^2],$$

$$K_{12} = \frac{1}{(-\lambda^2+\mu^2)}[\frac{4H_2^2(E_2A_1-E_1A_2)}{A} - 12.R_{11} - 2R_{12} + (R_{11}+R_{12}+R_{13})\lambda^2],$$

$$K_{21} = -\frac{K_{11}(B_2\lambda^2+D_2)}{A_2\lambda^2}, \qquad K_{22} = -\frac{K_{12}(B_2\mu^2+D_2)}{A_2\mu^2},$$

$$R_{21} = -\frac{D_2R_{12}-6H_2^2E_2-6H_2^2F_2-2G_2H_1H_2}{12A_2}, \qquad R_{22} = -\frac{2.B_2R_{12}+D_2R_{13}+E_2H_2^2}{2.A_2},$$

$$N_1 = \frac{G\alpha}{54Er}\left(\frac{K_{21}\lambda^2\cosh(\lambda y)}{\cosh(\lambda)} + \frac{K_{22}\mu^2\cosh(\mu y)}{\cosh\mu} + 12R_{21}y^2 + 2\,S_{22}\right)$$
$$+\frac{aNs_0(1-4\,s_0)\alpha}{3}\left[K_{21}\left(\frac{\cosh(\lambda y)}{\cosh(\lambda)}-1\right) + K_{22}\left(\frac{\cosh(\mu y)}{\cosh(\mu)}-1\right) + R_{21}\left(y^4-1\right) + S_{22}\left(y^2-1\right)\right]$$
$$+\frac{(-4a\theta+3s_0+16a\theta s_0^3-12a+12as_0^2-12a\theta s_0)\alpha}{108Er}\left(\frac{K_{11}\lambda^2\cosh(\lambda y)}{\cosh(\lambda)} + \frac{K_{12}\mu^2\cosh(\mu y)}{\cosh(\mu)}\right.$$
$$\left.+12R_{11}y^2 + 2\,R_{12}\right)$$
$$+\frac{Ns_0(1+2\,s_0)\,a\alpha}{3}\left[K_{11}\left(\frac{\cosh(\lambda y)}{\cosh(\lambda)}-1\right) + K_{12}\left(\frac{\cosh(\mu y)}{\cosh(\mu)}-1\right) + R_{11}\left(y^4-1\right) + R_{12}\left(y^2-1\right)\right]$$
$$+\frac{{H_2}^2s_0\alpha}{27}[4a(s_0-1)(2s_0+1)(3s_0+1)\theta + 3(-2as_0+3s_0+6a{s_0}^2-4a)](3y^2-1)^2$$
$$-\frac{2as_0\alpha}{9}\,{H_2}^2[(2s_0+1)(3s_0^2-5s_0-4)\theta - 6(s_0+2)]y^2(y^2-1) + 4\,\mu_2\,H_1H_2{s_0}^2y^2\left(y^2-1\right),$$

$$N_2 = -\frac{G\alpha}{54Er}\left(\frac{K_{21}\lambda^2\cosh\lambda y}{\cosh(\lambda)} + \frac{K_{22}\mu^2\cosh\mu y}{\cosh\mu} + 12R_{21}y^2 + 2\,S_{22}\right)$$
$$+\frac{aNs_0(-1+4\,s_0)\alpha}{3}\left[K_{21}\left(\frac{\cosh(\lambda y)}{\cosh(\lambda)}-1\right) + K_{22}\left(\frac{\cosh(\mu y)}{\cosh(\mu)}-1\right)\right.$$
$$\left.+R_{21}\left(y^4-1\right) + S_{22}\left(y^2-1\right)\right]$$
$$+\frac{(8a\theta+12a-3+8as_0-16as_0^2-12s_0)s_0\alpha}{108Er}\left(\frac{K_{11}\lambda^2\cosh(\lambda y)}{\cosh(\lambda)} + \frac{K_{12}\mu^2\cosh(\mu y)}{\cosh(\mu)} + 12R_{11}y^2 + 2\,R_{11}\right)$$
$$+\frac{2Ns_0(1-s_0)a\alpha}{3}\left[K_{11}\left(\frac{\cosh(\lambda y)}{\cosh(\lambda)}-1\right) + K_{12}\left(\frac{\cosh(\mu y)}{\cosh(\mu)}-1\right) + R_{11}\left(y^4-1\right) + R_{11}\left(y^2-1\right)\right]$$
$$+\frac{H_2^2s_0\alpha}{27Er}[-2a(s_0-1)(2s_0+1)(6s_0+1)\theta + 3(4as_0-3s_0-6a{s_0}^2+2a)](3y^2-1)^2$$
$$-\frac{2as_0{H_2}^2\alpha}{9Er}[-(2s_0+1)(3s_0^2-4s_0-2)\theta + 3(s_0+2)]y^2(y^2-1)$$
$$+4\,(\mu_1+\mu_2)\,H_1s_0y^2\left(y^2-1\right),$$

$$G = 2a\theta + 4a\theta s_0^2 + 6as_0 - 3s_0 - 12as_0^2 - 16a\theta s_0^3 + 10a\theta s_0 + 6a.$$

(66)

Appendix F. The coefficients of the steady solutions in normal anchoring in Poiseuille flows.

$$H_1 = \frac{6(s_0+2)(\theta(2s_0+1)+3)}{4(\mu_1 s_0+3\eta)(s_0+2)(\theta(2s_0+1)+3)+\alpha s_0(1+\lambda_L)(a\theta(s_0+2)(2s_0+1)+9s_0(1+\lambda_L))},$$

$$H_2 = \frac{9(1+\lambda_L)}{4(\mu_1 s_0+3\eta)(s_0+2)(\theta(2s_0+1)+3)+\alpha s_0(1+\lambda_L)(a\theta(s_0+2)(2s_0+1)+9s_0(1+\lambda_L))},$$

$$R_1 = -\frac{2(9H_2^2E_1+6H_2^2F_1+2G_1H_1H_2)+(9H_2^2E_2+6H_2^2F_2+2G_2H_1H_2)}{2D_1+D_2},$$

$$S_1 = \frac{2(6H_2^2E_1+6H_2^2F_1+2G_1H_1H_2)+(6H_2^2E_2+6H_2^2F_2+2G_2H_1H_2)-12(2B_1+B_2)R_1}{2D_1+D_2},$$

$$T_1 = -\frac{2H_2^2E_1+2H_2^2E_2+2(2B_1+B_2)S_1}{2D_1+D_2}, \qquad K_{11} = -(R_1+S_1+T_1),$$

$$K_{21} = -\frac{B_1\lambda^2+D_1}{A_1\lambda^2+C_1}K_{11}, \qquad R_2 = -\frac{D_1R_1+(9H_2^2E_1+6H_2^2F_1+2G_1H_1H_2)}{C_1},$$

$$S_2 = \frac{(6H_2^2E_1+6H_2^2F_1+2G_1H_1H_2)-12A_1R_2-12R_1B_1-D_1S_1}{C_1},$$

$$T_2 = \frac{H_2^2E_1+2A_1S_2+2S_1B_1+D_1T_1}{C_1}, \qquad K_{22} = -(K_{21}+R_2+S_2+T_2),$$

$$N_1 = \frac{G\alpha}{54Er}\left(\frac{K_{21}\lambda^2\cosh(\lambda y)}{\cosh(\lambda)}+\frac{K_{22}\mu^2\cosh(\mu y)}{\cosh\mu}+12R_2y^2+2S_2\right)$$
$$+\frac{aNs_0(1-4s_0)\alpha}{3}\left[K_{21}\left(\frac{\cosh(\lambda y)}{\cosh(\lambda)}-1\right)+K_{22}\left(\frac{\cosh(\mu y)}{\cosh(\mu)}-1\right)\right.$$
$$\left.+R_2\left(y^4-1\right)+S_2\left(y^2-1\right)\right]$$
$$+\frac{(-4a\theta+3s_0+16a\theta s_0^3-12a+12as_0^2-12a\theta s_0)\alpha}{108Er}\left(\frac{K_{11}\lambda^2\cosh(\lambda y)}{\cosh(\lambda)}+12R_1y^2+2S_1\right)$$
$$+\frac{Ns_0(1+2s_0)a\alpha}{3}\left[K_{11}\left(\frac{\cosh(\lambda y)}{\cosh(\lambda)}-1\right)+R_1\left(y^4-1\right)+S_1\left(y^2-1\right)\right]$$
$$+\frac{{H_2}^2s_0\alpha}{27}[4a(s_0-1)(2s_0+1)(3s_0+1)\theta+3(-2as_0+3s_0+6a{s_0}^2-4a)](3y^2-1)^2$$
$$-\frac{2as_0\alpha}{9}{H_2}^2[(2s_0+1)(3s_0^2-5s_0-4)\theta-6(s_0+2)]y^2(y^2-1)+4\mu_2H_1H_2{s_0}^2y^2\left(y^2-1\right),$$

$$N_2 = -\frac{G\alpha}{54Er}\left(\frac{K_{21}\lambda^2\cosh\lambda y}{\cosh(\lambda)}+\frac{K_{22}\mu^2\cosh\mu y}{\cosh\mu}+12R_2y^2+2S_2\right)$$
$$+\frac{aNs_0(-1+4s_0)\alpha}{3}\left[K_{21}\left(\frac{\cosh(\lambda y)}{\cosh(\lambda)}-1\right)+K_{22}\left(\frac{\cosh(\mu y)}{\cosh(\mu)}-1\right)\right.$$
$$\left.+R_2\left(y^4-1\right)+S_2\left(y^2-1\right)\right]$$
$$+\frac{(8a\theta+12a-3+8as_0-16as_0^2-12s_0)s_0\alpha}{108Er}\left(\frac{K_{11}\lambda^2\cosh(\lambda y)}{\cosh(\lambda)}+12R_1y^2+2S_1\right)$$
$$+\frac{2Ns_0(1-s_0)a\alpha}{3}\left[K_{11}\left(\frac{\cosh(\lambda y)}{\cosh(\lambda)}-1\right)+R_1\left(y^4-1\right)+S_1\left(y^2-1\right)\right]$$
$$+\frac{H_2^2as_0\alpha}{27Er}[-2a(s_0-1)(2s_0+1)(6s_0+1)\theta+3(4as_0-3s_0-6a{s_0}^2+2a)](3y^2-1)^2$$
$$-\frac{2as_0{H_2}^2\alpha}{9Er}[-(2s_0+1)(3s_0^2-4s_0-2)\theta+3(s_0+2)]y^2(y^2-1)$$
$$+4\left(\mu_1+\mu_2\right)H_1s_0y^2\left(y^2-1\right),$$
$$G = 2a\theta+4a\theta s_0^2+6as_0-3s_0-12as_0^2-16a\theta s_0^3+10a\theta s_0+6a.$$
(67)

REFERENCES

[1] W. R. Burghardt, *Molecular orientation and rheology in sheared lyotropic liquid crystalline polymers*, Macromol. Chem. Phys., 199 (1998), pp. 471–488.

[2] M. C. Calderer and B. Mukherjee, *Chevron patterns in liquid crystal flows*, Phys. D, 98 (1996), pp. 201–224.

[3] M. C. Calderer and B. Mukherjee, *On Poiseuille flow of polymeric liquid crystals*, Liq. Cryst., 22 (1997), pp. 121–136.

[4] T. Carlsson, *Theoretical investigation of the shear flow of nematic liquid crystals with the Leslie viscosity $\alpha > 0$: Hydrodynamic analogue of first order phase transitions*, Mol. Cryst. Liq. Cryst., 104 (1984), pp. 307–334.

[5] T. Carlsson, *Unit-sphere description of nematic flows*, Phys. Rev. A, 34 (1986), pp. 3393–3404.

[6] S. Chandrasekhar, *Liquid Crystals*, 2nd ed., Cambridge University Press, Cambridge, 1992.

[7] P. E. Cladis, S. Torza, in *Colloid and Interface Science*, vol. 4, M. Kerker, ed., Academic Press, New York, 1976.

[8] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, Oxford University Press, London, 1993.

[9] A. M. Donald and A. H. Windle, *Liquid Crystalline Polymers*, Cambridge Solid State Sci. Ser., Cambridge University Press, Cambridge, 1992.

[10] M. G. Forest and Q. Wang, *Monodomain response of finite-aspect-ratio macromolecules in shear and related linear flows*, Rheol. Acta, 42 (2003), pp. 20–46.

[11] M. G. Forest, Q. Wang, and H. Zhou, *Exact banded patterns from a Doi-Marrucci-Greco model of nematic liquid crystal polymers*, Phys. Rev. E, 61 (2000), pp. 6655–6662.

[12] M. G. Forest, Q. Wang, and H. Zhou, *Methods for the exact construction of mesoscale spatial structures in liquid crystal polymers*, Phys. D, 152–153 (2001), pp. 288–309.

[13] M. G. Forest, R. Zhou, and Q. Wang, *Full-tensor alignment criteria for sheared nematic polymers*, J. Rheol., 47 (2003), pp. 105–127.

[14] M. G. Forest, Q. Wang, H. Zhou, and R. Zhou, *Structure scaling properties of confined nematic polymers in plane Couette cells: The weak flow limit*, J. Rheol., 48 (2004), pp. 175–192.

[15] M. G. Forest, Q. Wang, and R. Zhou, *The weak shear phase diagram for nematic polymers*, Rheol. Acta, 43 (2004), pp. 17–37.

[16] M. G. Forest, Q. Wang, and R. Zhou, *The flow-phase diagram of Doi-Hess theory for sheared nematic polymers II: Finite shear rates*, Rheol. Acta, 44 (2004), pp. 80–93.

[17] M. G. Forest, Q. Wang, and H. Zhou, *Structure formation in sheared tumbling nematic liquid crystal polymers*, University of North Carolina, Chapel Hill, NC, preprint, 2004.

[18] M. G. Forest, R. Zhou, Q. Wang, X. Zheng, and R. Lipton, *Anisotropy and heterogeneity of nematic polymer nano-composite film properties*, in Modeling of Soft Matter, IMA Math. Appl., Vol. 141, M.-C. T. Calderer and E. M. Terentjev, eds., Springer-Verlag, New York, 2005, pp. 85–98.

[19] M. G. Forest, X. Zheng, R. Zhou, Q. Wang, and R. Lipton, *Anisotropy and dynamic ranges in effective properties of of nematic polymer nano-composites*. Adv. Func. Mat., 15 (2005), pp. 2029–2035.

[20] D. D. Joseph, *Fluid Dynamics of Viscoelastic Liquids*, Applied Math (N.Y.) 84, Springer-Verlag, New York, 1990.

[21] R. Kupferman, M. Kawaguchi, and M. M. Denn, *Emergence of structure in a model of liquid crystalline polymers with elastic coupling*, J. Non-Newtonian Fluid Mech., 91 (2000), pp. 255–271.

[22] R. G. Larson, *The Structure and Rheology of Complex Fluids*, Oxford University Press, London, 1999.

[23] R. G. Larson, *Roll-cell instabilities in shearing flows of nematic polymers*, J. Rheol., 37 (1993), page 175.

[24] R. G. Larson and D. W. Mead, *Development of orientation and texture during shearing of liquid-crystalline polymers*, Liq. Cryst., 12 (1993), pp. 751–768.

[25] R. G. Larson and D. W. Mead, *The Ericksen number and Deborah number cascade in sheared polymeric nematics*, Liq. Cryst., 15 (1993), pp. 151–169.

[26] P. Manneville, *The transition to turbulence in nematic liquid crystals: Part 1, general review. Part 2, on the transition via tumbling*, Mol. Cryst. Liq. Cryst., 70 (1981), pp. 223–250.

[27] G. Marrucci, *Tumbling regime of liquid-crystalline polymers*, Macromol., 24 (1991), pp. 4176–4182.
[28] G. Marrucci and F. Greco, *Flow behavior of liquid crystalline polymers*, Adv. Chem. Phys., 86 (1993), pp. 331–404.
[29] A. D. Rey and M. M. Denn, *Dynamical phenomena in liquid-crystalline materials*, Annual Rev. Fluid Mech., 34 (2002), pp. 233–266.
[30] G. Sgalari, G. L. Leal, and J. Feng, *The shear flow behavior of LCPs based on a generalized Doi model with distortional elasticity*, J. Non-Newtonian Fluid Mech., 102 (2002), pp. 361–382.
[31] Z. Tan and G. C. Berry, *Studies on the texture of nematic solutions of rodlike polymers, 3. Rheo-optical and rheological behavior in shear*, J. Rheol., 47 (2003), pp. 73–104.
[32] T. Tsuji and A. D. Rey, *Effect of long range order on sheared liquid crystalline polymers, Part 1: Compatibility between tumbling behavior and fixed anchoring*, J. Non-Newtonian Fluid Mech., 73 (1997), pp. 127–152.
[33] Q. Wang, *A hydrodynamic theory of nematic liquid crystalline polymers of different configurations*, J. Chem. Phys., 116 (2002), pp. 9120–9136.
[34] X. Zheng, M. G. Forest, R. Lipton, R. Zhou, and Q. Wang, *Exact scaling laws for electrical conductivity properties of nematic polymer nano-composite monodomains*, Adv. Func. Mat., 15 (2005), pp. 627–638.

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1261–1269

DRAW RESONANCE REVISITED*

MICHAEL RENARDY†

Abstract. We consider the problem of isothermal fiber spinning in a Newtonian fluid with no inertia. In particular, we focus on the effect of the downstream boundary condition. For prescribed velocity, it is well known that an instability known as draw resonance occurs at draw ratios in excess of about 20.2. We shall revisit this problem. Using the closed form solution of the differential equation, we shall show that an infinite family of eigenvalues exists and discuss its asymptotics. We also discuss other boundary conditions. If the force in the filament is prescribed, no eigenvalues exist, and the problem is stable at all draw ratios. If the area of the cross section is prescribed downstream, on the other hand, the problem is unstable at any draw ratio. Finally, we discuss the stability when the drawing speed is controlled in response to changes in cross section or force.

Key words. extensional flow, fiber spinning, draw resonance

AMS subject classifications. 76E99

DOI. 10.1137/050634268

1. Formulation of the problem. Fiber spinning is a manufacturing process used in making textile or glass fibers. A highly viscous fluid is extruded vertically from a nozzle. It is then cooled by the ambient air and solidifies. The solidified fiber is then wound on a spool at the end of the spinline.

Many physical effects are potentially significant in the study of this problem: viscosity, inertia, gravity, surface tension, cooling, elasticity, and air drag may all be relevant. In this paper, we focus on the simplest model and study the influence of varying boundary conditions. We assume that the force in the fiber is purely due to viscous effects, and we ignore temperature dependence. We use a one-dimensional model based on slender geometry and cross-sectional averaging. Let $u(x,t)$ denote the axial speed and $A(x,t)$ the area of the cross section. The spinneret is located at $x=0$ and the spool is at $x=L$. The conservation of mass implies that

$$A_t + (uA)_x = 0. \tag{1}$$

If only viscous forces contribute, the tension in the fiber is given by $3\eta A u_x$, where η is the viscosity. The requirement of constant tension in the fiber leads to

$$(Au_x)_x = 0. \tag{2}$$

Boundary conditions in industrial processes are notoriously ill defined. It is customary to assume that A and u are given at the spinneret: $A(0,t)=A_0$, $u(0,t)=u_0$. This of course, is an idealization; in reality there is a transition to an upstream flow, which cannot be described by the one-dimensional model. At the spool, it is sensible to prescribe either the speed or the force with which the fiber is wound. One might also consider control strategies where the flow is monitored and the speed of the spool adjusted to achieve a given objective. Since the goal of the manufacturing process is

*Received by the editors June 22, 2005; accepted for publication (in revised form) October 10, 2005; published electronically March 31, 2006. This research was supported by the National Science Foundation under grant DMS-0405810.

http://www.siam.org/journals/siap/66-4/63426.html

†Department of Mathematics, Virginia Tech, Blacksburg, VA 24061-0123 (renardym@math.vt.edu).

a fiber of uniform cross section, a control strategy might aim to keep the cross section constant. We shall consider what happens in the case of perfect success of such a control, i.e., when constant area is imposed as a boundary condition. We shall thus focus on the following three boundary conditions:

1. Prescribed speed: $u(L,t) = u_1$.
2. Prescribed force: $A(L,t)u_x(L,t) = F$, where F denotes the force divided by the elongational viscosity 3η.
3. Prescribed cross section: $A(L,t) = A_1$.

It is easy to see that the problem admits the steady solution

$$u_s(x) = u_0 e^{kx}, \quad A_s(x) = A_0 e^{-kx}, \tag{3}$$

where, respectively,

$$e^{kL} = u_1/u_0, \quad k = F/(A_0 u_0), \quad e^{kL} = A_0/A_1 \tag{4}$$

for the three choices of boundary conditions. The dimensionless quantity $q = e^{kL}$ is called the draw ratio.

2. Linear stability. The stability of the steady solution was first analyzed by Kase, Matsuo, and Yoshimoto [5] and Pearson and Matovich [6]. For subsequent reviews and textbook chapters, see also [2, 7, 9, 10]. We note that much of the literature on draw resonance is concerned with the effect of additional physical mechanism, which are not included in our analysis. Inertia, elasticity, and cooling generally have a stabilizing effect, while surface tension and shear thinning are destabilizing. This paper, on the other hand, will focus purely on the case of Stokes flow and investigate the effect of varying the downstream boundary condition. In the case of prescribed speed, an instability known as draw resonance is found for draw ratios in excess of about 20.2, while no such instability is found for prescribed force. In [8], the case of a linear combination of speed and force is also investigated; as expected, the stability threshold increases from 20.2 to infinity as the relevant coefficient is varied. The case of prescribed cross section does not seem to have been analyzed in the literature. We shall see that this boundary condition leads to instability at all draw ratios.

We linearize at the steady solution and consider exponentially varying perturbations:

$$u(x,t) = u_s(x) + \tilde{u}(x)e^{\lambda t}, \quad A(x,t) = A_s(x) + \tilde{a}(x)e^{\lambda t}. \tag{5}$$

The linearized equations are

$$\lambda\tilde{a} + (u_s\tilde{a} + A_s\tilde{u})_x = 0, \quad (A_s\tilde{u}_x + \tilde{a}(u_s)_x)_x = 0. \tag{6}$$

It is advantageous to make the transformation $z = e^{kx}$. The steady solution then takes the form $u_s(z) = u_0 z$, $A_s(z) = A_0/z$. The linearized equations (6) are transformed to

$$\lambda\tilde{a} + kz(zu_0\tilde{a} + A_0\tilde{u}/z)_z = 0, \quad (A_0\tilde{u}_z + zu_0\tilde{a})_z = 0. \tag{7}$$

We can rewrite these equations in the form

$$\begin{aligned} \frac{\lambda}{ku_0}\frac{\tilde{a}}{A_0} + z\frac{\tilde{a}}{A_0} + z^2\frac{\tilde{a}_z}{A_0} - \frac{\tilde{u}}{u_0 z} + \frac{\tilde{u}_z}{u_0} &= 0, \\ \frac{\tilde{u}_z}{u_0} + z\frac{\tilde{a}}{A_0} &= C_1. \end{aligned} \tag{8}$$

We simplify by setting $\lambda/(ku_0) = \mu$, $\tilde{a}/A_0 = a$, $\tilde{u}/u_0 = u$. This is equivalent to nondimensionalizing the equations by scaling the velocity and area with their steady state values at the spinneret, length along the filament with $1/k$, and time with $1/(ku_0)$. The resulting dimensionless equations are

$$\begin{aligned} \mu a + za + z^2 a_z - \frac{u}{z} + u_z &= 0, \\ u_z + za &= C_1. \end{aligned} \tag{9}$$

We can now solve the second equation for a:

$$a = \frac{C_1 - u_z}{z}. \tag{10}$$

After inserting this into the first equation, we obtain

$$-u + (-\mu + z)u_z + \mu C_1 - z^2 u_{zz} = 0. \tag{11}$$

Clearly, $u = \mu C_1$ is a particular solution, and $u = z - \mu$ is a particular solution of the homogeneous equation. We can then obtain the full solution using the reduction of order method:

$$u(z) = \mu C_1 + (z - \mu)C_2 + C_3(-ze^{\mu/z} + (\mu - z)\mathrm{Ei}(\mu/z)); \tag{12}$$

see also [8]. Here Ei is the exponential integral defined for $z > 0$ by

$$\mathrm{Ei}(x) = \int_{-\infty}^{z} \frac{e^t}{t}\, dt, \tag{13}$$

where the integral is understood in the principal value sense (see Ch. 5 of [1]).

We have the boundary conditions $u(1) = a(1) = 0$ at the spinneret and one of the following three at the take-up point:

1. Fixed speed: $u(q) = 0$.
2. Fixed force: $u_z(q) + qa(q) = 0$.
3. Fixed cross section: $a(q) = 0$.

The requirement of a nontrivial solution leads to the following characteristic equations. For fixed speed,

$$(e^{\mu} - e^{\mu/q})q + (q - \mu)(\mathrm{Ei}(\mu) - \mathrm{Ei}(\mu/q)) = 0. \tag{14}$$

For fixed force,

$$e^{\mu} = 0. \tag{15}$$

For fixed cross section,

$$\mathrm{Ei}(\mu) - \mathrm{Ei}(\mu/q) = 0. \tag{16}$$

3. Remarks on well-posedness. For the case of fixed force, no eigenvalues exist, and indeed it can be shown for this case that any initial disturbance will decay to zero in finite time. To see this, we note that the linearized equations (9) (without the assumption of exponential time dependence) are

$$\begin{aligned} a_t + z^2 a_z - \frac{u}{z} &= 0, \\ u_z + za &= 0. \end{aligned} \tag{17}$$

We can integrate the second of these equations to find

$$u = -\int_1^z ya\,dy. \tag{18}$$

For any given initial data $a(z,0) = a_0(z)$, we can now solve the equation using the iterative procedure

$$\begin{aligned} u^1 &= 0, \\ a_t^n + z^2 a_z^n - \frac{u^n}{z} &= 0, \quad a^n(1,t) = 0, \quad a_n(z,0) = a_0(z), \\ u^{n+1} &= -\int_0^x ya^n\,dy. \end{aligned} \tag{19}$$

It follows by induction that

$$a(z,t) = 0 \tag{20}$$

for $1 < z < Z(t)$, where

$$Z'(t) = Z(t)^2, \quad Z(0) = 1. \tag{21}$$

The solution will thus become identically zero as soon as $Z(t)$ reaches the value q.

For the case of prescribed speed, a rigorous proof of well-posedness and spectrally determined growth can be given along similar lines as [4]. The case of prescribed cross section is somewhat different and will be discussed now. In this case, the linear problem, again without the assumption of exponential time dependence, is

$$\begin{aligned} a_t + za + z^2 a_z - \frac{u}{z} + u_z &= 0, \\ u_z + za &= \phi(t), \end{aligned} \tag{22}$$

where $\phi(t)$ is a function to be determined after the boundary conditions are imposed. We can integrate the second equation to find

$$u(z,t) = -\int_1^z ya(y,t)\,dy + (z-1)\phi(t) \tag{23}$$

and insert this result into the first equation. This yields

$$a_t + z^2 a_z + \frac{1}{z}\int_1^z ya\,dy + \frac{\phi(t)}{z} = 0. \tag{24}$$

We next set $a = b + \gamma/z$, where γ is independent of z, and b satisfies

$$\int_1^q b(z,t)\chi(z)\,dz, \tag{25}$$

with χ to be determined. We shall denote the projections of a onto b and γ/z by P and Q.

We want χ to be such that we also have

$$\int_1^q z^2 b_z \chi(z)\,dz = 0. \tag{26}$$

We note that the boundary condition $a(1,t) = a(q,t) = 0$ leads to $b(1,t) = qb(q,t)$. We now integrate by parts to find

$$\int_1^q z^2 b_z \chi \, dz = -\int_1^q b(z^2\chi)' \, dz + q^2 b(q,t)\chi(q) - b(1,t)\chi(1). \tag{27}$$

We achieve our objective if

$$\frac{d}{dz}(z^2\chi) = K\chi, \quad \chi(1) = q\chi(q) \tag{28}$$

for some constant K. This leads to

$$\chi(z) = \frac{1}{z^2} e^{-K/z}, \quad e^{-K} = \frac{1}{q} e^{-K/q}. \tag{29}$$

With χ thus determined, (24) can be decomposed as follows:

$$\begin{aligned} \gamma_t - \gamma + zQ\left(\frac{1}{z}\int_1^z yb \, dy\right) + \phi(t) &= 0, \\ b_t + z^2 b_z + P\left(\frac{1}{z}\int_1^z yb \, dy\right) &= 0. \end{aligned} \tag{30}$$

The solution procedure is now obvious. We solve the second equation for b with the boundary condition $b(1,t) = qb(q,t)$, and after b is determined, the first equation, and the condition $\gamma(t) = -b(1,t)$, determine $\phi(t)$. Well-posedness and spectral growth are obvious from this reformulation of the equations. We note that the boundary condition for b is a two-point condition rather than an upstream condition. This is reflected in the nature of the eigenspectrum below; the limit of the real part of large eigenvalues will be a finite number rather than $-\infty$.

We note that a general discussion of boundary conditions for the hyperbolic systems arising in viscoelastic flows is given in [3]; in the context of that discussion the Newtonian case is degenerate, even if inertia is included.

4. Asymptotics of large eigenvalues. In this section, we focus on the asymptotic behavior of large eigenvalues. It is instructive to look at this case for a number of reasons. As we shall see, some instabilities can be predicted from the analysis of this limit. The asymptotic formula also gives insights into the qualitative nature of the eigenspectrum; it shows that there are infinitely many eigenvalues and that they line up along a curve and shows what the approximate spacing is.

We begin with the simpler case of fixed cross section.

We use the asymptotic expansion of the exponential integral for large argument [1]:

$$\mathrm{Ei}(\mu) = \pi i \, \mathrm{sgn}(\mathrm{Im}\, \mu) + \frac{e^\mu}{\mu}\left(1 + O\left(\frac{1}{\mu}\right)\right). \tag{31}$$

Using this, we can approximate the characteristic equation by

$$e^\mu = qe^{\mu/q}, \tag{32}$$

which leads to

$$\mu = \frac{q}{q-1}(2n\pi i + \ln q). \tag{33}$$

Since $\ln q/(q-1)$ is positive, we find an infinite family of unstable eigenvalues for any value of q. For $q = 2$ the following table compares the eigenvalues found from the asymptotic formula (33) with exact roots of the characteristic equation found by Newton's method:

n	Result from (33)	Exact eigenvalue
1	$1.38629 + 12.5664i$	1.35405 + 12.42i
2	$1.38629 + 25.1327i$	1.37705 + 25.0552i
3	$1.38629 + 37.6991i$	1.38204 + 37.6467i
4	$1.38629 + 50.2655i$	1.38387 + 50.226i
5	$1.38629 + 62.8319i$	1.38473 + 62.8002i

For the case of fixed speed, we need to carry the approximation of the exponential integral a little further:

$$\mathrm{Ei}(\mu) = \pi i \,\mathrm{sgn}(\mathrm{Im}\,\mu) + \frac{e^\mu}{\mu}\left(1 + \frac{1}{\mu} + \frac{2}{\mu^2} + O\left(\frac{1}{\mu^3}\right)\right). \tag{34}$$

Using this, we obtain the approximate characteristic equation

$$e^{\mu - \mu/q} = -\frac{q^3}{(q-1)\mu^2}. \tag{35}$$

For large $|\mu|$, we obtain the following asymptotic formula for the eigenvalues

$$\mu_n = \frac{q}{q-1}\left(2n\pi i + \ln\left(\frac{q^3}{q-1}\right) - 2\ln\left(2n\pi\frac{q}{q-1}\right)\right). \tag{36}$$

Here n is any integer. For $n \to \infty$, the real parts of these eigenvalues tend to $-\infty$ logarithmically, i.e., they are stable.

Since the asymptotic approximation depends on $|\mu/q|$ being large in addition to $|\mu|$, the first few eigenvalues are predicted poorly if q is large. The following table illustrates this behavior for $q = 20.218$, the value at which onset of draw resonance occurs:

n	μ_n given by (36)	Exact eigenvalue
1	$2.40565 + 6.61013i$	4.66015i
2	$0.947223 + 13.2203i$	$-0.738622 + 11.4532$i
3	$0.094096 + 19.8304i$	$-1.20379 + 18.2453$i
4	$-0.511207 + 26.4405i$	$-1.55854 + 25.0014$i
5	$-0.980716 + 33.0506i$	$-1.85118 + 31.7307$i
10	$-2.43915 + 66.1013i$	$-2.8734 + 65.1607$i
20	$-3.89758 + 132.203i$	$-4.0729 + 131.605$i
50	$-5.82551 + 330.506i$	$-5.86414 + 330.225$i

Another limit which can be approached by asymptotics is that of large draw ratio. If we consider the case $\mu \to \infty$, $q \to \infty$ in (14) with the expectation that $\mu/q \to 0$, the balance of leading order terms yields

$$e^\mu + \ln q = 0, \tag{37}$$

i.e.,

$$\mu_n = (2n-1)i\pi + \ln\ln q. \tag{38}$$

Since q must be really large for $\ln\ln q$ to be considered "large," this approximation is not useful in practice. For $q = 5 * 10^8$, a totally unrealistic value of course, we have

$$i\pi + \ln\ln q = 2.99724 + 3.14159i, \quad 3i\pi + \ln\ln q = 2.99724 + 9.42478i, \tag{39}$$

compared to exact eigenvalues of $2.72203 + 3.471i$ and $2.76567 + 9.63623i$, respectively.

5. Control strategies. In this section, we consider how the onset of draw resonance is affected if we add a control which adjusts the drawing speed in response to observed fluctuations. Since the goal of the manufacturing process is a uniform thread, it seems natural to change the speed in response to fluctuations in the cross-sectional area. This leads to a downstream boundary condition

$$u(q) - \epsilon a(q) \tag{40}$$

to be imposed on (9). Intuitively, we would be tempted to increase the drawing speed when the cross section becomes larger, i.e., $\epsilon > 0$.

The resulting characteristic equation is

$$(e^{\mu} - e^{\mu/q})q + \left(q + \frac{\epsilon}{q} - \mu\right)(\mathrm{Ei}(\mu) - \mathrm{Ei}(\mu/q)) = 0. \tag{41}$$

The asymptotic behavior of the eigenvalues can be discussed by the same methods as above. We obtain

$$\mu_n \sim \frac{q}{q-1}\left(2\pi n i + \ln\left(\frac{\epsilon}{2n\pi q}\right) - \mathrm{sgn}(\epsilon)\frac{i\pi}{2}\right). \tag{42}$$

For large n, these eigenvalues become stable.

Next, we consider the onset value for draw resonance as a function of ϵ. The results are summarized in the following table:

ϵ	Critical draw ratio
0	20.218
10	18.872
20	17.224
30	14.904

Contrary to intuition, the effect of the control is destabililizing, and the critical draw ratio decreases. For negative ϵ, if we just track the eigenvalue that is responsible for draw resonance at $\epsilon = 0$, the critical draw ratio seems to increase:

ϵ	Critical draw ratio
-20000	200.00
-10000	147.10
-5000	109.13
-1000	57.324
-500	44.843
-100	28.536
-50	25.056
-20	22.420
-10	21.3817

It would be wrong to think, however, that we can achieve stability at any draw ratio by choosing ϵ large and negative. In fact, there are new instabilities at low draw ratios when $|\epsilon|$ is large. We can see this by looking at the asymptotic behavior of eigenvalues assuming that both $|\mu|$ and $|\epsilon|$ are large. The result is

$$\mu_n \sim \frac{q}{q-1}\left(2\pi i n + \ln\left(\frac{\epsilon q}{\epsilon + 2\pi i n q(q-1)}\right)\right). \tag{43}$$

For $q = 4$, $\epsilon = -100$, for instance, this formula yields $\mu_1 = 1.54832 + 9.23897i$, while the actual eigenvalue is $1.21466 + 9.25047i$. The instability resulting from this

eigenvalue persists for $q < 8.01516$. We thus have two separate instabilities, one for low draw ratio and another for high draw ratio. The next table shows the first eigenvalue for $\epsilon = -100$ as a function of the draw ratio q (the higher eigenvalues are more stable):

q	First eigenvalue
2	$1.33788 + 12.9402$i
3	$1.40897 + 10.04$i
4	$1.21466 + 9.25047$i
6	$0.579954 + 8.55471$i
8	$0.004134 + 8.08557$i
10	$-0.533251 + 7.66538$i
15	$-1.59346 + 5.79611$i
20	$-0.601555 + 4.77314$i
25	$-0.173004 + 4.58841$i
30	$0.0558642 + 4.49297$i
40	$0.31093 + 4.37887$i

Another control strategy is to monitor the force in the thread and change the drawing speed in response. This leads to the boundary condition

$$u(q) + \epsilon(u'(q) + qa(q)) = 0. \tag{44}$$

This boundary condition was also considered in [8]. The resulting characteristic equation is

$$(e^{\mu} - e^{\mu/q})q + (q - \mu)(\mathrm{Ei}(\mu) - \mathrm{Ei}(\mu/q)) + \epsilon e^{\mu} = 0. \tag{45}$$

The behavior of large eigenvalues becomes

$$\mu_n \sim \frac{q}{q-1}\left(2n\pi i + \ln\left(\frac{q^3}{q+\epsilon-1}\right) - 2\ln\left(2n\pi\frac{q}{q-1}\right)\right). \tag{46}$$

We see from this asymptotic formula that a positive ϵ is stabilizing, as would heuristically be expected. The effect on draw resonance follows the same trend, and the results show no surprises.

ϵ	Critical draw ratio
-5	5.387
-2	16.786
0	20.218
5	26.561
10	31.632
20	40.137
50	60.37
100	87.468

6. Conclusions. We have investigated the simplest model of fiber spinning in a viscous fluid, which includes only viscous forces, neglecting all other effects. In this simple case, the linear stability problem has a closed form solution in terms of an exponential integral, which can be exploited to gain substantial qualitative insight into the behavior of the eigenvalues. The stability of the flow depends crucially on the choice of downstream boundary conditions. If the speed is prescribed, then, as is well known, the flow becomes unstable beyond a critical draw ratio. On the other hand, prescribed force leads to no instabilities, while prescribed cross section leads to instability at all draw ratios. In terms of strategies to control instability, adjustment

of the speed in reaction to changes in cross section has an effect opposite of what is intuitively expected. In addition to changing the threshold for high draw ratio instabilities, such a control also produces new instabilities at low draw ratios.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1964.

[2] M. M. Denn, *Continuous drawing of liquids to form fibers*, Ann. Rev. Fluid Mech., 12 (1980), pp. 365–387.

[3] M. G. Forest and Q. Wang, *Dynamics of slender viscoelastic jets*, SIAM J. Appl. Math., 54 (1994), pp. 996–1032.

[4] T. Hagen and M. Renardy, *Studies on the linear equations of melt-spinning of viscous fluids*, Differential Integral Equ., 14 (2001), pp. 19–36.

[5] S. Kase, T. Matsuo, and Y. Yoshimoto, *Theoretical analysis of melt spinning. Part* 2: *Surging phenomena in extrusion casting of plastic films*, Seni Kikai Gakkaishi, 19 (1966), pp. T63–72.

[6] J. R. A. Pearson and M. A. Matovich, *On spinning a molten threadline—stability*, Ind. Engrg. Chem. Fund., 8 (1969), pp. 605–609.

[7] C. J. S. Petrie, *Elongational Flows*, in Res. Notes Math. 29, Pitman, Boston, 1979.

[8] W. W. Schultz and S. H. Davis, *Effects of boundary condition on the stability of slender viscous fibers*, ASME J. Appl. Mech., 51 (1984), pp. 1–5.

[9] R. I. Tanner, *Engineering Rheology*, 2nd ed., Oxford University Press, London, 2000.

[10] A. L. Yarin, *Free Liquid Jets and Films: Hydrodynamics and Rheology*, Longman, London, 1993.

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1270–1300

HILBERT FORMULAS FOR r-ANALYTIC FUNCTIONS IN THE DOMAIN EXTERIOR TO SPINDLE*

MICHAEL ZABARANKIN† AND ANDREI F. ULITKO‡

Abstract. Hilbert formulas for an r-analytic function, defined by a generalized Cauchy–Riemann system in the domain exterior to the contour of a spindle in the meridional cross-section plane, have been derived. The derivation is based on the theory of Riemann boundary-value problems for analytic functions. For numerical calculations, Fourier integrals with Hilbert formulas representing the real and imaginary parts of the r-analytic function have been reduced to the form of regular integrals. The problem of the axially symmetric steady motion of a rigid spindle-shaped body in a Stokes fluid has been solved, and the pressure in the fluid has been expressed analytically based on a Hilbert formula. As an illustration, streamlines about the body, vortex and pressure functions at the contour of the body, and the drag force, exerted on the body by the fluid, have been calculated.

Key words. Hilbert formula, r-analytic function, Riemann boundary-value problem, analytic function, spindle, bipolar coordinates, Fourier integral transform, Lame equation, Stokes equations, pressure, vorticity, drag force

AMS subject classifications. 30E20, 35Q15, 35Q30, 76D07

DOI. 10.1137/050632403

1. Introduction.

This paper derives Hilbert formulas for an r-analytic function in the domain exterior to the contour of a spindle in the meridional cross-section plane and applies these formulas in a hydrodynamic problem of axially symmetric Stokes flow about a rigid spindle-shaped body. So-called r-analytic functions are a special case of pseudoanalytic functions [2, 22] or so-called p-analytic and (p,q)-analytic functions introduced by Polozhii [17]. For a given positive continuously differentiable function $p = p(x,y)$, a p-analytic function $F(x,y) = u(x,y) + i\,v(x,y)$ is defined by the generalized Cauchy–Riemann system

$$\frac{\partial u}{\partial x} = \frac{1}{p}\,\frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{1}{p}\,\frac{\partial v}{\partial x}, \tag{1}$$

where $i = \sqrt{-1}$. This system is encountered in different areas of mathematical physics, in particular, the theory of elasticity [6, 17, 21], hydrodynamics [21, 28, 30], and quantum mechanics [12].

We consider a special case of system (1) from the linear theory of elasticity and hydrodynamics of Stokes flows. Let $\mathbf{u}$ be the displacement vector of an isotropic elastic medium, characterized by Poisson number $m \in [2, +\infty)$. In the framework of the linear theory of elasticity, equilibrium of the medium is governed by Lame equation

$$2\,\tfrac{m-1}{m-2}\ \operatorname{grad}\operatorname{div}\mathbf{u} - \operatorname{curl}\operatorname{curl}\mathbf{u} = \mathbf{0}. \tag{2}$$

*Received by the editors May 25, 2005; accepted for publication (in revised form) October 18, 2005; published electronically March 31, 2006.
http://www.siam.org/journals/siap/66-4/63240.html
†Stevens Institute of Technology, Department of Mathematical Sciences, Hoboken, NJ 07030 (mzabaran@stevens.edu).
‡National Taras Shevchenko University of Kiev, Department of Mechanics and Mathematics, Kiev, Ukraine (ulitko@univ.kiev.ua).

For the vector field, $\mathbf{u}$, divergence, θ, and vorticity, $\boldsymbol{\omega}$, are two fundamental characteristics that we introduce by

$$\theta = -2\,\tfrac{m-1}{m-2}\,\operatorname{div}\mathbf{u}, \tag{3}$$

$$\boldsymbol{\omega} = \operatorname{curl}\mathbf{u}. \tag{4}$$

In the case of $m = 2$ and $\operatorname{div}\mathbf{u} = 0$, the function θ is defined as a finite limit of expression $-2\frac{m-1}{m-2}\operatorname{div}\mathbf{u}$ under $m \to 2$ and $\operatorname{div}\mathbf{u} \to 0$, and Lame equation (2) becomes mathematically identical to the Stokes model for a viscous incompressible fluid under low Reynolds numbers (so-called Stokes fluid). In this case, the displacement vector, $\mathbf{u}$, corresponds to the vector of velocity of fluid particles, and θ corresponds to the pressure $\mathcal{P}$, i.e., $\theta = \mathcal{P}/\rho$, where ρ is the shear viscosity. The Stokes model is given by

$$\begin{cases} \operatorname{curl}(\operatorname{curl}\mathbf{u}) = -\operatorname{grad}\theta, \\ \operatorname{div}\mathbf{u} = 0. \end{cases} \tag{5}$$

This model can also be obtained by linearizing Navier equations with the zero Reynolds number [8].

In terms of θ and $\boldsymbol{\omega}$, Lame equation (2) and the first equation in Stokes model (5) take the form

$$\operatorname{grad}\theta = -\operatorname{curl}\boldsymbol{\omega}. \tag{6}$$

Relations (3), (4), and (6) imply that θ and $\boldsymbol{\omega}$ are harmonic functions. Indeed, $\Delta\,\theta = \operatorname{div}\operatorname{grad}\theta = -\operatorname{div}\operatorname{curl}\boldsymbol{\omega} \equiv 0$, and $\operatorname{curl}\operatorname{curl}\boldsymbol{\omega} = -\operatorname{curl}\operatorname{grad}\theta \equiv \mathbf{0}$, where $\Delta = \nabla^2$ is the harmonic operator. Since $\operatorname{div}\boldsymbol{\omega} \equiv 0$, we have $\Delta\,\boldsymbol{\omega} = \operatorname{grad}\operatorname{div}\boldsymbol{\omega} - \operatorname{curl}\operatorname{curl}\boldsymbol{\omega} = \mathbf{0}$. These relations are used for constructing exact solutions to Lame equation (2) and Stokes model (5); see [21].

In planar problems, the vorticity vector has only one nonzero component. Suppose a problem is considered in the plane xy in the system of cartesian coordinates (x, y, z); then $\boldsymbol{\omega} = (0, 0, \omega_z)$. In this case, (6) reduces to the Cauchy–Riemann system for an analytic function $F = \theta + i\,\omega_z$, and the vector $\mathbf{u}$ is expressed by the Kolosov–Muskhelishvili formulas; see, for example, [21]. In axially symmetric three-dimensional (3-D) problems, the vorticity vector can be represented by a scalar vortex function ω, and (6) reduces to the generalized Cauchy–Riemann system. Let (r, φ, z) be a system of cylindric coordinates with basis $(\mathbf{e}_r, \mathbf{e}_\varphi, \mathbf{k})$, and let axis z determine the axis of symmetry. The axially symmetric case means that the vector $\mathbf{u}$ is independent of angular coordinate φ, and the vorticity vector can be represented by $\boldsymbol{\omega} = \omega\,\mathbf{e}_\varphi$, where ω is the vortex function. Here, θ and ω depend only upon r and z. Equation (6) reduces to a special case of the generalized Cauchy–Riemann system with $p(r, z) = r$, i.e.,

$$\frac{\partial\theta}{\partial r} = \frac{1}{r}\frac{\partial}{\partial z}(r\omega), \qquad \frac{\partial\theta}{\partial z} = -\frac{1}{r}\frac{\partial}{\partial r}(r\omega). \tag{7}$$

Function $F(r, z) = \theta(r, z) + i\,r\,\omega(r, z)$ is called r-analytic if it satisfies system (7). Functions θ and $r\,\omega$ are considered to be real and imaginary parts of the r-analytic function, respectively. Let Δ_k define so-called k-harmonic operator

$$\Delta_k = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{\partial^2}{\partial z^2} - \frac{k^2}{r^2}, \tag{8}$$

where $\Delta \equiv \Delta_0$. System (7) implies that $\theta(r,z)$ and $\omega(r,z)$ are harmonic and 1-harmonic functions, respectively, i.e.,

$$\Delta\,\theta = 0, \qquad \Delta_1\omega = 0. \tag{9}$$

Suppose $\mathcal{D}$ is a bounded domain in the meridional cross-section plane rz with smooth boundary $\partial\mathcal{D}$, and suppose boundary values of functions θ and ω at $\partial\mathcal{D}$ are given. Establishing existence and uniqueness of solutions to (9) is a Dirichlet problem, which is discussed in 3-D potential theory [23]. For the domain exterior to $\mathcal{D}$, a harmonic function vanishing at infinity in 3-D space is uniquely determined by its boundary value at $\partial\mathcal{D}$. For domains determined by the surface of bodies of revolution in the meridional cross-section plane, Polozhii [17] obtained integral representations for p-analytic functions via analytic functions of complex variable and generalized Kolosov–Muskhelishvili formulas for axially symmetric problems of the linear theory of elasticity.

Of special interest is the problem of finding the boundary value of the imaginary part of an r-analytic function via the boundary value of its real part, and vice versa. If harmonic functions θ and ω are represented by integrals with densities that are analytic functions, then the problem reduces to finding corresponding relations between those analytic functions. These relations are called Hilbert formulas. Integral representations for r-analytic functions in domains exterior to the contour of lens, spindle, torus, and two spheres in the meridional cross-section plane are discussed in [21]. In our previous work [28], we obtained Hilbert formulas for the domain exterior to the contour of a spindle by integrating the generalized Cauchy–Riemann system (7) in bipolar coordinates. In this paper we derive Hilbert formulas for the same domain based on the theory of Riemann boundary-value problems for analytic functions. For details of this theory, see [4]. We represent functions θ and ω by Fourier integrals in bipolar coordinates and reduce system (7) to a problem for a meromorphic function on three parallel contours in the infinite strip $|\operatorname{Re}\mu| \le 1$. Further, using conformal mapping, we reformulate this problem as the Riemann boundary-value problem for finding a meromorphic function in the plane with the branch cut along the segment $[-1,1]$. A solution to this problem is represented by a Cauchy integral, and boundary values of this solution at the upper and lower banks of the branch cut are expressed by the Sokhotski formulas [4]. For numerical calculations, representations of r-analytic functions in the form of Fourier integrals with Hilbert formulas reduce to the form of regular integrals. In our previous work [29], we found a nonhomogeneous solution to the problem on three parallel contours using complex Fourier transform. In this paper, we apply the approach of Riemann boundary-value problems to obtaining the nonhomogeneous solution from the class of meromorphic functions with only two simple poles at $\mu = \pm\frac{1}{2}$ in the strip $|\operatorname{Re}\mu| \le 1$ and to analyzing the existence of corresponding nontrivial homogeneous solutions.

Hilbert formulas are applied for finding the pressure function in axially symmetric problems of Stokes flows. As discussed, Lame equation (2) with $m = 2$ corresponds to the model for a Stokes fluid, and θ corresponds to the pressure in the fluid. However, there is a crucial difference in solving Lame equation (2) with $m \in (2,+\infty)$ and Stokes model (5). Namely, if $m \in (2,+\infty)$, and $\mathbf{u}$ is already known, then θ can readily be determined by (3). But relation (3) cannot be used for determining θ in the case of $m = 2$, since $\operatorname{div}\mathbf{u} = 0$. Thus, in axially symmetric problems of Stokes flows, we use Hilbert formulas for r-analytic functions to express θ via ω. As an illustration, we consider the problem of axially symmetric steady motion of a rigid spindle-shaped

body in a Stokes fluid. Stokes [19] was the first to study steady motion of a rigid sphere in a viscous incompressible fluid under low Reynolds numbers. Constructing analytical solutions to similar problems with rigid bodies of axially symmetric shape is mainly based on a stream function approach [8, 13, 14]. There are extensive studies of axially symmetric Stokes flows about spherical cap [3, 21], two spheres [18, 27], torus [5, 7, 11, 16, 20, 21, 26], lens-shaped body [3, 21, 24], and spindle-shaped body [15, 28, 30]. However, analytic formulas for the pressure function in these studies were obtained only for torus [21] and spindle-shaped body [28]. For detailed discussion of these issues, see [21]. In this paper we solve the problem of steady axially symmetric motion of a rigid spindle-shaped body in a Stokes fluid using the stream function that we proposed in [30]. Based on Hilbert formulas, we obtain an analytic expression for the pressure in the fluid, which coincides with that derived in our work [28]. To illustrate obtained results, we calculate streamlines about the body, vortex and pressure functions at the contour of the body, and the drag force exerted on the body by the fluid.

The paper is organized as follows. Section 2 represents an r-analytic function in the domain exterior to the contour of a spindle in the meridional cross-section plane. Section 3 derives Hilbert formulas for r-analytic functions in the framework of the theory of Riemann boundary-value problems for analytic functions. Section 4 reduces Fourier integral representations of r-analytic functions with Hilbert formulas to the form of regular integrals. Section 5 solves the problem of steady axially symmetric motion of a rigid spindle-shaped body in a Stokes fluid. Section 6 obtains analytic expressions for the pressure and drag force exerted on the body. Section 7 concludes the paper. The appendix derives some auxiliary formulas.

2. An r-analytic function in the domain exterior to spindle. Let (r, φ, z) be a system of cylindric coordinates with basis $(\mathbf{e}_r, \mathbf{e}_\varphi, \mathbf{k})$, and let the z-axis be the axis of symmetry. In the meridional cross-section plane rz, bipolar coordinates (ξ, η) are defined by

$$r = c\,\frac{\sin\eta}{\cosh\xi - \cos\eta}\,, \qquad z = c\,\frac{\sinh\xi}{\cosh\xi - \cos\eta}\,, \tag{10}$$

where $-\infty < \xi < +\infty$, $0 \le \eta \le \pi$, and c is a metric parameter of bipolar coordinates. Spindle is an axially symmetric body, whose contour in the plane rz is determined by fixing coordinate η, i.e., $\eta = \eta_0$ (see Figure 1). For example, the surface of the spindle for $\eta_0 = \frac{\pi}{2}$ is sphere.

In the system of bipolar coordinates, derivatives $\frac{\partial}{\partial r}$, $\frac{\partial}{\partial z}$ and the k-harmonic operator Δ_k, defined by (8), take the form

$$\begin{aligned}
\frac{\partial}{\partial r} &= -\frac{1}{c}\left(\sinh\xi\sin\eta\frac{\partial}{\partial\xi} - (\cosh\xi\cos\eta - 1)\frac{\partial}{\partial\eta}\right),\\
\frac{\partial}{\partial z} &= -\frac{1}{c}\left((\cosh\xi\cos\eta - 1)\frac{\partial}{\partial\xi} + \sinh\xi\sin\eta\frac{\partial}{\partial\eta}\right),\\
\Delta_k =& \frac{(\cosh\xi - \cos\eta)^2}{c^2}\\
&\times\left(\frac{\partial^2}{\partial\xi^2} + \frac{\partial^2}{\partial\eta^2} - \frac{\sinh\xi}{\cosh\xi - \cos\eta}\frac{\partial}{\partial\xi} + \left(\cot\eta - \frac{\sin\eta}{\cosh\xi - \cos\eta}\right)\frac{\partial}{\partial\eta} - \frac{k^2}{\sin^2\eta}\right).
\end{aligned} \tag{11}$$

Let $F(r, z) = \theta(r, z) + i\,r\,\omega(r, z)$ be an r-analytic function satisfying system (7). In this case, θ and ω are harmonic and 1-harmonic functions defined by (9). In the

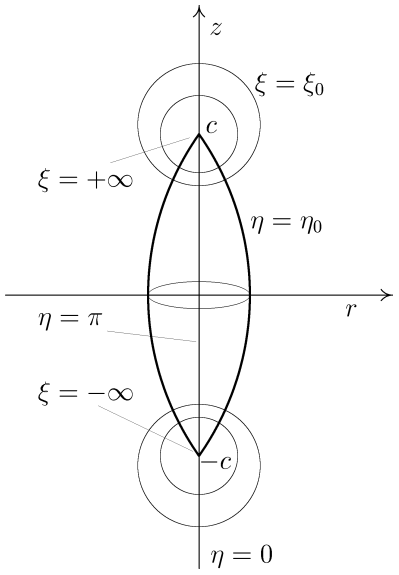


FIG. 1. *Bipolar coordinates and spindle-shaped body.*

domain exterior to the contour of spindle in the plane rz, an arbitrary k-harmonic function is represented by a Fourier integral with respect to variable ξ. Thus, in bipolar coordinates, functions $\theta(\xi,\eta)$ and $\omega(\xi,\eta)$ take the form (see [10])

$$\theta(\xi,\eta)=\frac{1}{2\pi i}\sqrt{\cosh\xi-\cos\eta}\int\limits_{-i\infty}^{+i\infty}X(\mu)\,\mathrm{P}_{-\frac{1}{2}+\mu}(\cos\eta)\,\mathrm{e}^{-\xi\mu}d\mu,\quad 0\le\eta\le\eta_0, \tag{12}$$

$$\omega(\xi,\eta)=\frac{1}{2\pi i}\sqrt{\cosh\xi-\cos\eta}\int\limits_{-i\infty}^{+i\infty}Y(\mu)\,\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta)\,\mathrm{e}^{-\xi\mu}d\mu,\quad 0\le\eta\le\eta_0, \tag{13}$$

where $\mathrm{P}^{(k)}_{-\frac{1}{2}+\mu}(\cos\eta)$ is the associated Legendre function of the first kind of complex index μ; see [1]. For $k=0$, the upper index (k) is omitted. Since $\mathrm{P}_{-\frac{1}{2}+i\tau}(\cos\eta)\sim\frac{1}{\sqrt{2\pi\sin\eta}}\,\mathrm{e}^{\eta|\tau|}$ and $\mathrm{P}^{(1)}_{-\frac{1}{2}+i\tau}(\cos\eta)\sim\frac{|\tau|}{\sqrt{2\pi\sin\eta}}\,\mathrm{e}^{\eta|\tau|}$ at $\tau\to\pm\infty$, we require functions $X(i\tau)$ and $Y(i\tau)$ in Fourier integrals (12) and (13) to have exponentially fast convergence $C\mathrm{e}^{-\gamma|\tau|}$ at $\tau\to\pm\infty$, where C is a constant, and $\gamma>\eta_0$.

Note that the harmonic functions θ and ω represented by (12) and (13), respectively, vanish at infinity $\sqrt{r^2+z^2}\to\infty$, that is, at $\xi\to 0$ and $\eta\to 0$. This guarantees uniqueness of solutions to a Dirichlet problem for (9) in the domain of consideration.

3. Problem for an analytic function on three parallel contours. Let $\mathcal{M}_{[a,b]}$ be the space of functions that are meromorphic in the strip $a\le\operatorname{Re}\mu\le b$ and have exponentially fast convergence at $|\mu|\to\infty$, i.e., vanish as $C\mathrm{e}^{-\gamma|\tau|}$, where C is a constant, and $\gamma>\eta_0$. Functions from the space $\mathcal{M}_{[a,b]}$ may have simple poles at $\mu=\pm\frac{1}{2}$ only.

Suppose $X(\mu),\,Y(\mu)\in\mathcal{M}_{[-1,1]}$, and $\eta\in[0,\eta_0]$. Under these assumptions, the following relations hold [29]:

(14)
$$\frac{\partial\theta}{\partial r} = \frac{1}{4\pi ic}\sqrt{\cosh\xi-\cos\eta}\int\limits_{-i\infty}^{+i\infty}\left(X(\mu+1)-2X(\mu)+X(\mu-1)\right)\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta)\mathrm{e}^{-\xi\mu}d\mu,$$

(15)
$$\frac{\partial\omega}{\partial z} = \frac{1}{4\pi ic}\sqrt{\cosh\xi-\cos\eta}\int\limits_{-i\infty}^{+i\infty}\left(\left(\mu+\tfrac{3}{2}\right)Y(\mu+1)-2\mu\,Y(\mu)+\left(\mu-\tfrac{3}{2}\right)Y(\mu-1)\right)\times\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta)\mathrm{e}^{-\xi\mu}d\mu.$$

The derivation of formulas (14) and (15) is presented in the appendix.

Note that because function $\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta)$ has simple zeroes at $\mu=\pm\frac{1}{2}$, functions $X(\mu)$ and $Y(\mu)$ are allowed to have simple poles at $\mu=\pm\frac{1}{2}$. Substituting (14) and (15) into the first equation of system (7), we obtain an equation for $X(\mu)$ and $Y(\mu)$:

(16)
$$X(\mu+1)-2X(\mu)+X(\mu-1)=\left(\mu+\tfrac{3}{2}\right)Y(\mu+1)-2\mu\,Y(\mu)+\left(\mu-\tfrac{3}{2}\right)Y(\mu-1),$$

where $\mu=i\tau$, $\tau\in\mathbb{R}$.

Equation (16) is the problem on three parallel contours for finding either $X(\mu)$ given $Y(\mu)$ at the contour $\operatorname{Re}\mu=0$ or $Y(\mu)$ given $X(\mu)$ at $\operatorname{Re}\mu=0$. The uniqueness of a solution to (16) with respect to either $X(\mu)$ or $Y(\mu)$ reduces to proving that a corresponding homogeneous equation has only trivial solution. We will also show that if $X(\mu)$, $Y(\mu)\in\mathcal{M}_{[-1,1]}$ solve (16), then the function $X(\mu)$ should necessarily satisfy an additional condition $\int_{-i\infty}^{+i\infty}X(\mu)\,d\mu=0$.

3.1. Hilbert formula for the real part of r-analytic function. In this section, we solve (16) with respect to $X(\mu)\in\mathcal{M}_{[-1,1]}$ assuming that $Y(\mu)\in\mathcal{M}_{[-1,1]}$ is given. If $X(\mu)\in\mathcal{M}_{[-1,1]}$ solves (16), then $X(\mu)$ is unique. Indeed, suppose that $X_1(\mu)\in\mathcal{M}_{[-1,1]}$ and $X_2(\mu)\in\mathcal{M}_{[-1,1]}$ both solve (16), and $X_1(\mu)\neq X_2(\mu)$. This means that $X_0(\mu)=X_1(\mu)-X_2(\mu)\in\mathcal{M}_{[-1,1]}$ is a solution to homogeneous equation (16) such that $X_0(\mu)\not\equiv 0$. We prove the following proposition.

PROPOSITION 1 (homogeneous solution $X_0(\mu)$). *The only $X_0(\mu)\in\mathcal{M}_{[-1,1]}$ that solves homogeneous equation* (16)

$$X_0(\mu+1)-2X_0(\mu)+X_0(\mu-1)=0,\quad \operatorname{Re}\mu=0, \tag{17}$$

is zero function, i.e., $X_0(\mu)\equiv 0$.

Proof. Let a function $Z(\mu)$ be defined by

$$Z(\mu)=X_0(\mu)-X_0(\mu-1)\quad\text{and}\quad Z(\mu+1)=X_0(\mu+1)-X_0(\mu);$$

then $Z(\mu)\in\mathcal{M}_{[0,1]}$, and (17) reduces to

$$Z(\mu+1)-Z(\mu)=0,\quad \operatorname{Re}\mu=0. \tag{18}$$

The function

$$z=i\tan(\pi\mu) \tag{19}$$

maps the complex plane μ with the strip $0\le\operatorname{Re}\mu\le 1$ to the complex plane z with the branch cut along the segment $[-1,1]$ (see Figure 2). The line $\mu=i\tau$, $\tau\in\mathbb{R}$,

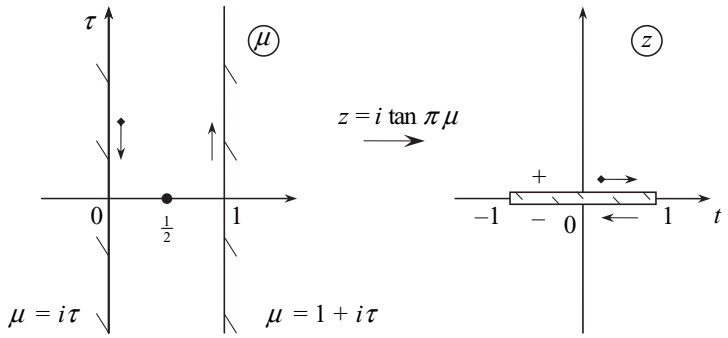


FIG. 2. *Function $z = i\tan(\pi\mu)$ maps the complex plane μ with the strip $0 \le \operatorname{Re}\mu \le 1$ to the complex plane z with the branch cut along the segment $[-1,1]$.*

corresponds to the upper bank of the branch cut, and the line $\mu = 1 + i\tau$, $\tau \in \mathbb{R}$, corresponds to the lower bank of the branch cut with the counterclockwise orientation as shown in Figure 2. The pole at $\mu = \frac{1}{2}$ in the complex plane μ corresponds to the pole at infinity in the complex plane z.

Using conformal mapping (19), we introduce a function $\tilde{Z}(z)$ in the complex plane z such that

$$\tilde{Z}^+(t) = Z(i\tau), \qquad \tilde{Z}^-(t) = Z(i\tau + 1), \qquad \tau \in \mathbb{R},$$

where $\tilde{Z}^+(t)$ and $\tilde{Z}^-(t)$ are boundary values of $\tilde{Z}(z)$ at the upper and lower banks of the branch cut. Problem (18) becomes a Riemann boundary-value problem for finding the function $\tilde{Z}(z)$, analytic in the plane z with the branch cut along the segment $[-1,1]$, such that

$$\tilde{Z}^+(t) = \tilde{Z}^-(t), \quad -1 \le t \le 1.$$

This means that the function $\tilde{Z}(z)$ is analytic in the whole plane z and has the simple pole at infinity. A solution to this problem is given by

$$\tilde{Z}(z) = \tilde{c}_1 z + \tilde{c}_0,$$

where $\tilde{c}_0$ and $\tilde{c}_1$ are constants. However, since $Z(\mu) \in \mathcal{M}_{[0,1]}$, i.e., it vanishes at $|\mu| \to \infty$, the function $\tilde{Z}(z)$ should satisfy $\tilde{Z}(\pm 1) = 0$. Consequently, $\tilde{c}_0 = 0$ and $\tilde{c}_1 = 0$, and we obtain

$$X_0(\mu + 1) - X_0(\mu) = 0, \qquad \operatorname{Re}\mu = 0,$$

for $X_0(\mu) \in \mathcal{M}_{[0,1]}$. This is the same problem as (18). Its only solution is $X_0(\mu) \equiv 0$. Thus, since $\mathcal{M}_{[-1,1]} \subset \mathcal{M}_{[0,1]}$, the only solution to (17) in the space of $\mathcal{M}_{[-1,1]}$ is zero function. □

Consequently, if $X(\mu) \in \mathcal{M}_{[-1,1]}$ solves (16), then it is unique.

THEOREM 1 (Hilbert formula for the real part). *Let the real and imaginary parts of an r-analytic function be represented in bipolar coordinates by Fourier integrals* (12) *and* (13)*, respectively. If $X(\mu) \in \mathcal{M}_{[-1,1]}$ and $Y(\mu) \in \mathcal{M}_{[-1,1]}$, then at the contour $\operatorname{Re}\mu = 0$, the function $X(\mu)$ is represented by the Hilbert formula for the real part of the r-analytic function*

$$X(\mu) = \mu\, Y(\mu) - \frac{i}{2\cos(\pi\mu)} \fint_{-i\infty}^{+i\infty} Y(\nu)\, \frac{\cos(\pi\nu)}{\sin[\pi(\nu - \mu)]}\, d\nu, \qquad \operatorname{Re}\mu = 0, \tag{20}$$

where $⨍$ *means the Cauchy principal value or v.p. (i.e.,* valeur principale*) of a singular integral.*

Proof. For $\operatorname{Re}\mu = 0$, equation (16) may be rewritten as

$$[X(\mu+1) - X(\mu)] - [X(\mu) - X(\mu-1)]$$
$$= \left[\left(\mu+\tfrac{3}{2}\right)Y(\mu+1) - \left(\mu-\tfrac{1}{2}\right)Y(\mu)\right] - \left[\left(\mu+\tfrac{1}{2}\right)Y(\mu) - \left(\mu-\tfrac{3}{2}\right)Y(\mu-1)\right].$$

Introducing a new function $Z(\mu)$ by

$$Z(\mu+1) = [X(\mu+1) - X(\mu)] - \left[\left(\mu+\tfrac{3}{2}\right)Y(\mu+1) - \left(\mu-\tfrac{1}{2}\right)Y(\mu)\right],$$
$$Z(\mu) = [X(\mu) - X(\mu-1)] - \left[\left(\mu+\tfrac{1}{2}\right)Y(\mu) - \left(\mu-\tfrac{3}{2}\right)Y(\mu-1)\right],$$

we reduce (16) to

$$Z(\mu+1) - Z(\mu) = 0, \qquad \operatorname{Re}\mu = 0,$$

for $Z(\mu) \in \mathcal{M}_{[0,1]}$. This is the same problem as (18). Its only solution is $Z(\mu) \equiv 0$. Consequently,

$$X(\mu+1) - X(\mu) = \left(\mu+\tfrac{3}{2}\right)Y(\mu+1) - \left(\mu-\tfrac{1}{2}\right)Y(\mu), \quad \operatorname{Re}\mu = 0, \tag{21}$$
$$X(\mu) - X(\mu-1) = \left(\mu+\tfrac{1}{2}\right)Y(\mu) - \left(\mu-\tfrac{3}{2}\right)Y(\mu-1), \quad \operatorname{Re}\mu = 0. \tag{22}$$

It is sufficient to solve only (21) for $X(\mu) \in \mathcal{M}_{[0,1]}$ given $Y(\mu) \in \mathcal{M}_{[0,1]}$. It can be shown that solutions to (21) and (22) provide the same $X(\mu)$ at $\operatorname{Re}\mu = 0$.

Representing $X(\mu)$ by

$$X(\mu) = \left(\mu+\tfrac{1}{2}\right)Y(\mu) + \hat{X}(\mu), \tag{23}$$

where $\hat{X}(\mu) \in \mathcal{M}_{[0,1]}$ is a new function, we reformulate (21) for $\hat{X}(\mu)$:

$$\hat{X}(\mu+1) - \hat{X}(\mu) = Y(\mu), \quad \operatorname{Re}\mu = 0. \tag{24}$$

Since $\hat{X}(\mu)$ and $Y(\mu)$ have exponentially fast convergence at $|\mu| \to \infty$, we integrate equation (24) at the contour $\operatorname{Re}\mu = 0$ and obtain

$$\operatorname*{Res}_{\mu=\frac{1}{2}} \hat{X}(\mu) = \frac{1}{2\pi i} ⨍_{-i\infty}^{+i\infty} Y(\mu)\, d\mu. \tag{25}$$

Using conformal mapping (19), we introduce functions $\tilde{X}(z)$ and $\tilde{Y}(z)$, meromorphic in the complex plane z, such that

$$\begin{aligned} \tilde{X}^+(t) &= \hat{X}(i\tau), & \tilde{Y}^+(t) &= Y(i\tau), \\ \tilde{X}^-(t) &= \hat{X}(i\tau+1), & \tilde{Y}^-(t) &= Y(i\tau+1), \end{aligned} \qquad \tau \in \mathbb{R},$$

where pairs $\tilde{X}^+(t)$, $\tilde{X}^-(t)$ and $\tilde{Y}^+(t)$, $\tilde{Y}^-(t)$ are boundary values of $\tilde{X}(z)$ and $\tilde{Y}(z)$ at the upper and lower banks of the branch cut, respectively. Note that the pole $\mu = \frac{1}{2}$ in the complex plane μ became the pole at infinity in the complex plane z. Problem (24) reduces to a Riemann boundary-value problem for finding the function

$\tilde{X}(z)$ analytic in the plane z with the branch cut along the segment $[-1,1]$ and having the simple pole at infinity:

$$\tilde{X}^+(t)-\tilde{X}^-(t)=-\tilde{Y}^+(t),\quad -1\le t\le 1. \tag{26}$$

The function $\tilde{X}(z)$ is represented by a Cauchy integral

$$\tilde{X}(z)=-\frac{1}{2\pi i}\oint_{-1}^{1}\frac{\tilde{Y}^+(s)}{s-z}\,ds+\tilde{c}_1 z+\tilde{c}_0,$$

where $\tilde{c}_0$ and $\tilde{c}_1$ are unknown constants. The function $\tilde{Y}^+(s)$ must satisfy Hölder condition, i.e., $|\tilde{Y}^+(t_2)-\tilde{Y}^+(t_1)|\le c'\,|t_2-t_1|^{\lambda}$ for all t_1, $t_2\in[-1,1]$, some $\lambda\in(0,1]$, and nonnegative constant c'. At points $t=\pm1$, it converges to zero not slower than $(1-t^2)^{\frac{\eta_0}{2\pi}}$.

Using the Sokhotski formulas, we obtain

$$\tilde{X}^+(t)=-\frac{1}{2}\tilde{Y}^+(t)-\frac{1}{2\pi i}\oint_{-1}^{1}\frac{\tilde{Y}^+(s)}{s-t}\,ds+\tilde{c}_1 t+\tilde{c}_0. \tag{27}$$

Constants $\tilde{c}_0$ and $\tilde{c}_1$ are found based on the requirement $\tilde{X}^+(\pm1)=0$ and condition $\tilde{Y}^+(t)|_{t\to\pm1}\sim(1-t^2)^{\frac{\eta_0}{2\pi}}$:

$$\tilde{c}_0=-\frac{1}{2\pi i}\oint_{-1}^{1}\frac{s}{1-s^2}\,\tilde{Y}^+(s)\,ds,\qquad \tilde{c}_1=-\frac{1}{2\pi i}\oint_{-1}^{1}\frac{1}{1-s^2}\,\tilde{Y}^+(s)\,ds.$$

Note that $\tilde{c}_1$ satisfies $\tilde{c}_1\operatorname{Res}_{\mu=\frac{1}{2}}[i\tan(\pi\mu)]=\operatorname{Res}_{\mu=\frac{1}{2}}\hat{X}(\mu)$, where $\operatorname{Res}_{\mu=\frac{1}{2}}\hat{X}(\mu)$ is determined by (25). Thus, expression (27) takes the form

$$\tilde{X}^+(t)=-\frac{1}{2}\tilde{Y}^+(t)-\frac{1}{2\pi i}(1-t^2)\oint_{-1}^{1}\frac{\tilde{Y}^+(s)}{1-s^2}\,\frac{ds}{s-t}. \tag{28}$$

Finally, making change of variables $t=i\tan(\pi\mu)$, $\operatorname{Re}\mu=0$, and $s=i\tan(\pi\nu)$, $\operatorname{Re}\nu=0$, in (28), and substituting (28) into (23), we obtain Hilbert formula (20). The change of variables in singular integral (28) is valid because $t'(\mu)=\pi/\cos^2(\pi\mu)$ is a continuous strictly positive function at $\operatorname{Re}\mu=0$ (for details see [4]). □

3.2. Hilbert formula for the imaginary part of *r*-analytic function. Now we solve (16) with respect to $Y(\mu)\in\mathcal{M}_{[-1,1]}$ assuming that $X(\mu)\in\mathcal{M}_{[-1,1]}$ is given. As in the previous case, uniqueness of a solution $Y(\mu)$ to (16) is guaranteed by the fact that the homogeneous equation corresponding to (16) has only trivial solution.

PROPOSITION 2 (homogeneous solution $Y_0(\mu)$). *The only $Y_0(\mu)\in\mathcal{M}_{[-1,1]}$ that solves homogeneous equation*

$$\left(\mu+\tfrac{3}{2}\right)Y_0(\mu+1)-2\mu\,Y_0(\mu)+\left(\mu-\tfrac{3}{2}\right)Y_0(\mu-1)=0,\qquad \operatorname{Re}\mu=0, \tag{29}$$

is zero function, i.e., $Y_0(\mu)\equiv0$.

Proof. Let a function $Z(\mu)$ be defined by

$$Z(\mu+1) = \left(\mu+\tfrac{3}{2}\right) Y_0(\mu+1) - \left(\mu-\tfrac{1}{2}\right) Y_0(\mu),$$
$$Z(\mu) = \left(\mu+\tfrac{1}{2}\right) Y_0(\mu) - \left(\mu-\tfrac{3}{2}\right) Y_0(\mu-1);$$

then $Z(\mu) \in \mathcal{M}_{[0,1]}$, and (29) reduces to $Z(\mu+1) - Z(\mu) = 0$, $\operatorname{Re}\mu = 0$. This problem, being the same as (18), has only zero solution, i.e., $Z(\mu) \equiv 0$. Consequently, for $Y_0(\mu) \in \mathcal{M}_{[0,1]}$ we obtain

$$\left(\mu+\tfrac{3}{2}\right) Y_0(\mu+1) - \left(\mu-\tfrac{1}{2}\right) Y_0(\mu) = 0, \qquad \operatorname{Re}\mu = 0. \tag{30}$$

Multiplying (30) by $\left(\mu+\frac{1}{2}\right)$ and representing $Y_0(\mu) = \frac{1}{\mu^2-\frac{1}{4}}\, Y_1(\mu)$, where $Y_1(\mu)$ is a function analytic in $0 \le \operatorname{Re}\mu \le 1$ and vanishing at $|\mu| \to \infty$, we reduce equation (30) to the problem for function $Y_1(\mu)$: $Y_1(\mu+1) - Y_1(\mu) = 0$, $\operatorname{Re}\mu = 0$. By similar reasoning, we conclude that the only possible solution to this problem is zero function. Thus, since $\mathcal{M}_{[-1,1]} \subset \mathcal{M}_{[0,1]}$, the only solution to (29) in the space of $\mathcal{M}_{[-1,1]}$ is $Y_0(\mu) \equiv 0$. □

This proposition implies that $Y(\mu) \in \mathcal{M}_{[-1,1]}$ solving (16) is unique.

THEOREM 2 (Hilbert formula for the imaginary part). *Let the real and imaginary parts of an r-analytic function be represented in bipolar coordinates by Fourier integrals* (12) *and* (13)*, respectively. If $X(\mu) \in \mathcal{M}_{[-1,1]}$, $Y(\mu) \in \mathcal{M}_{[-1,1]}$, and $\int_{-i\infty}^{+i\infty} X(\mu)\, d\mu = 0$, then at the contour $\operatorname{Re}\mu = 0$, the function $Y(\mu)$ is represented by the Hilbert formula for the imaginary part of the r-analytic function*

$$Y(\mu) = \frac{1}{\mu^2-\frac{1}{4}} \left(\mu\, X(\mu) + \frac{i}{2\cos(\pi\mu)} \fint_{-i\infty}^{+i\infty} X(\nu)\, \frac{\cos(\pi\nu)}{\sin[\pi(\nu-\mu)]}\, d\nu \right), \qquad \operatorname{Re}\mu = 0. \tag{31}$$

Proof. Repeating the same arguments as in the proof of Theorem 1, we obtain (21) and (22), which we now solve with respect to $Y(\mu)$. It can be shown that solutions to (21) and (22) provide the same $Y(\mu)$ at $\operatorname{Re}\mu = 0$. Multiplying (21) by $\left(\mu+\frac{1}{2}\right)$ and representing function $Y(\mu)$ by

$$Y(\mu) = \frac{1}{\mu^2-\frac{1}{4}} \left(\hat{Y}(\mu) + \left(\mu-\tfrac{1}{2}\right) X(\mu) \right), \tag{32}$$

where $\hat{Y}(\mu)$ is analytic in $0 \le \operatorname{Re}\mu \le 1$ and has exponentially fast convergence at $|\mu| \to \infty$, we reduce (21) to

$$\hat{Y}(\mu+1) - \hat{Y}(\mu) = -X(\mu), \quad \operatorname{Re}\mu = 0. \tag{33}$$

This problem is similar to (24). However, while the function $\hat{X}(\mu)$ in (24) has the simple pole at $\mu = \frac{1}{2}$, the function $\hat{Y}(\mu)$ in (33) does not. Consequently, integrating (33) at the contour $\operatorname{Re}\mu = 0$, we obtain $\frac{1}{2\pi i}\int_{-i\infty}^{+i\infty} X(\mu)\, d\mu = 0$. This means that the function $X(\mu)$ should necessarily satisfy this condition.

Using conformal mapping (19), we introduce analytic function $\tilde{X}(z)$ and meromorphic function $\tilde{Y}(z)$ in the complex plane z such that

$$\begin{aligned} &\tilde{Y}^+(t) = \hat{Y}(i\tau), && \tilde{X}^+(t) = X(i\tau), \\ &\tilde{Y}^-(t) = \hat{Y}(i\tau+1), && \tilde{X}^-(t) = X(i\tau+1), \end{aligned} \qquad \tau \in \mathbb{R},$$

where pairs $\tilde{Y}^+(t)$, $\tilde{X}^+(t)$ and $\tilde{Y}^-(t)$, $\tilde{X}^-(t)$ are boundary values of functions $\tilde{Y}(z)$ and $\tilde{X}(z)$ at the upper and lower banks of the branch cut, respectively. Problem (33) reduces to a Riemann boundary-value problem for finding the function $\tilde{Y}(z)$ analytic in the plane z with the branch cut along the segment $[-1,1]$ and bounded at infinity:

$$\tilde{Y}^+(t)-\tilde{Y}^-(t)=\tilde{X}^+(t),\quad -1\le t\le 1. \tag{34}$$

Problem (34) is similar to (26). Thus, a solution to (34) takes the form

$$\tilde{Y}^+(t)=\frac{1}{2}\tilde{X}^+(t)+\frac{1}{2\pi i}\fint_{-1}^{1}\frac{\tilde{X}^+(s)}{s-t}\,ds+\tilde{c}_0, \tag{35}$$

where $\tilde{c}_0$ is a constant. The function $\tilde{X}^+(s)$ must satisfy Hölder condition $|\tilde{X}^+(t_2)-\tilde{X}^+(t_1)|\le c'\,|t_2-t_1|^{\lambda}$ for some $\lambda\in(0,1]$, nonnegative constant c', and all t_1, $t_2\in[-1,1]$. At points $t=\pm1$, it converges to zero not slower than $(1-t^2)^{\frac{\eta_0}{2\pi}}$. Since there is only one constant to satisfy two conditions $\tilde{Y}^+(1)=0$ and $\tilde{Y}^+(-1)=0$, we have

$$\tilde{c}_0=\frac{1}{2\pi i}\fint_{-1}^{1}\frac{s}{1-s^2}\,\tilde{X}^+(s)\,ds,\qquad \frac{1}{2\pi i}\fint_{-1}^{1}\frac{1}{1-s^2}\,\tilde{X}^+(s)\,ds=0,$$

where the second relation is equivalent to $\int_{-i\infty}^{+i\infty}X(\mu)\,d\mu=0$. Thus, expression (35) takes the form

$$\tilde{Y}^+(t)=\frac{1}{2}\tilde{X}^+(t)+\frac{1}{2\pi i}(1-t^2)\fint_{-1}^{1}\frac{\tilde{X}^+(s)}{1-s^2}\,\frac{ds}{s-t}. \tag{36}$$

Finally, making change of variables $t=i\tan(\pi\mu)$, $\operatorname{Re}\mu=0$, and $s=i\tan(\pi\nu)$, $\operatorname{Re}\nu=0$, in (36), and substituting (36) into (32), we obtain Hilbert formula (31). The change of variables in singular integral (36) is valid because $t'(\mu)=\pi/\cos^2(\pi\mu)$ is a continuous strictly positive function at $\operatorname{Re}\mu=0$ (for details see [4]). □

Corollary. Hilbert formula (20) reduces the necessary condition

$$\int_{-i\infty}^{+i\infty}X(\mu)\,d\mu=0$$

to identity, i.e.,

$$\int_{-i\infty}^{+i\infty}\left(\mu\,Y(\mu)-\frac{i}{2\cos(\pi\mu)}\fint_{-i\infty}^{+i\infty}Y(\nu)\frac{\cos(\pi\nu)}{\sin[\pi(\nu-\mu)]}\,d\nu\right)d\mu\equiv 0.$$

Proof. The result follows from the fact that

$$\fint_{-i\infty}^{+i\infty}\frac{d\mu}{\cos(\pi\mu)\,\sin[\pi(\nu-\mu)]}=-2i\,\frac{\nu}{\cos(\pi\nu)},\qquad \operatorname{Re}\nu=0.$$

Conditions for changing the order of integration for singular integrals are discussed in [4]. □

Example. The following two pairs reduce (16) to identity:

$$X(\mu)=\frac{\mu}{\cos(\pi\mu)},\quad Y(\mu)=\frac{1}{\cos(\pi\mu)}\qquad\text{and}\qquad X(\mu)=\frac{\mu^2+\frac{1}{4}}{\cos(\pi\mu)},\quad Y(\mu)=\frac{\mu}{\cos(\pi\mu)}.$$

In each pair, $X(\mu)$ and $Y(\mu)$ are related by Hilbert formulas (20) and (31).

Proof. Verification of Hilbert formulas reduces to the following calculations:

$$\fint_{-i\infty}^{+i\infty}\frac{d\nu}{\sin[\pi(\nu-\mu)]}=0,\qquad \fint_{-i\infty}^{+i\infty}\frac{\nu\,d\nu}{\sin[\pi(\nu-\mu)]}=\frac{i}{2},\qquad \fint_{-i\infty}^{+i\infty}\frac{\nu^2\,d\nu}{\sin[\pi(\nu-\mu)]}=i\mu.$$

Obviously, the condition $\int_{-i\infty}^{+i\infty}X(\mu)\,d\mu=0$ holds: $\int_{-i\infty}^{+i\infty}\frac{\mu}{\cos(\pi\mu)}\,d\mu=0$ and $\int_{-i\infty}^{+i\infty}\frac{(\mu^2+\frac{1}{4})}{\cos(\pi\mu)}\,d\mu=0$, respectively. □

4. Implementation of Fourier integrals with Hilbert formulas. Hilbert formulas (20) and (31) are expressed by singular integrals; consequently, they require special treatment in numerical implementation. In this section, we derive formulas for efficient calculating double integrals in (12) with (20) and (13) with (31).

PROPOSITION 3 (function $\omega(\xi,\eta)$). *If function $Y(\mu)$ is represented at $\operatorname{Re}\mu=0$ by Hilbert formula* (31)*, then the function $\omega(\xi,\eta)$ takes the form*

$$\begin{aligned}\omega(\xi,\eta)=&\frac{1}{2\pi i}\sqrt{\cosh\xi-\cos\eta}\\ &\times\int_{-\infty}^{+\infty}X(i\tau)\left(\frac{\tau}{\tau^2+\frac{1}{4}}\,\mathrm{P}^{(1)}_{-\frac{1}{2}+i\tau}(\cos\eta)\,\mathrm{e}^{-i\xi\tau}+G_1(\xi,\eta,\tau)\right)d\tau,\end{aligned}\tag{37}$$

where

(38)

$$G_1(\xi,\eta,\tau)=\begin{cases}\dfrac{i\sqrt{2}}{\pi\sin\eta}\left(\mathrm{e}^{i\xi\tau}\displaystyle\int_0^{\eta}g(\xi,\tau,t)\sqrt{\cos t-\cos\eta}\,dt-2\,h_1(\xi,\eta)\,\sinh\dfrac{\xi}{2}\right),\ \ \xi\neq0,\\ \dfrac{\sqrt{2}}{\pi\sin\eta}\displaystyle\int_0^{\eta}\cot\tfrac{t}{2}\sinh(\tau t)\sqrt{\cos t-\cos\eta}\,dt,\ \ \xi=0,\end{cases}$$

$$g(\xi,\tau,t)=\frac{\sinh\xi\cosh(\tau t)-i\sin t\,\sinh(\tau t)}{\cosh\xi-\cos t},\tag{39}$$

$$h_1(\xi,\eta)=\frac{\pi}{\sqrt{2}}\left(\sqrt{1+\left(\frac{\sin\frac{\eta}{2}}{\sinh\frac{\xi}{2}}\right)^2}-1\right),\quad \xi\neq0.\tag{40}$$

Both integrals in (38) *are regular and can be efficiently calculated by a Gaussian quadrature.*

Proof. Substituting formula (31) into (13), we obtain

(41)

$$\begin{aligned}\omega(\xi,\eta)=&\frac{1}{2\pi i}\sqrt{\cosh\xi-\cos\eta}\int_{-\infty}^{+\infty}\left(\tau X(i\tau)+\frac{1}{2}\fint_{-\infty}^{+\infty}X(i\tau_1)\frac{\cosh(\pi\tau_1)}{\cosh(\pi\tau)}\frac{d\tau_1}{\sinh[\pi(\tau_1-\tau)]}\right)\\ &\times\frac{\mathrm{P}^{(1)}_{-\frac{1}{2}+i\tau}(\cos\eta)}{\tau^2+\frac{1}{4}}\,\mathrm{e}^{-i\xi\tau}d\tau.\end{aligned}$$

The inner integral in (41) is singular, but the external integral is regular. Consequently, we do not need the Poincaré–Bertrand formula for changing the order of integration in (41) (see [4]):

$$\begin{aligned}\mathcal{I}_Y(\xi,\eta) &= \frac{1}{2}\int\limits_{-\infty}^{+\infty}\left(\mathop{\not\int}\limits_{-\infty}^{+\infty} X(i\tau_1)\,\frac{\cosh(\pi\tau_1)\,d\tau_1}{\sinh[\pi(\tau_1-\tau)]}\right)\frac{\mathrm{P}^{(1)}_{-\frac{1}{2}+i\tau}(\cos\eta)}{\left(\tau^2+\frac{1}{4}\right)\cosh(\pi\tau)}\,\mathrm{e}^{-i\xi\tau}d\tau\\ &= \frac{1}{2}\int\limits_{-\infty}^{+\infty} X(i\tau_1)\cosh(\pi\tau_1)\underbrace{\left(\mathop{\not\int}\limits_{-\infty}^{+\infty}\frac{\mathrm{P}^{(1)}_{-\frac{1}{2}+i\tau}(\cos\eta)}{\left(\tau^2+\frac{1}{4}\right)\cosh(\pi\tau)}\,\frac{\mathrm{e}^{-i\xi\tau}\,d\tau}{\sinh[\pi(\tau_1-\tau)]}\right)}_{I_1(\xi,\eta,\tau_1)}d\tau_1.\end{aligned}$$

The inner integral $I_1(\xi,\eta,\tau_1)$ in $\mathcal{I}_Y(\xi,\eta)$ is calculated based on the following representation [1]:

$$\mathrm{P}^{(1)}_{-\frac{1}{2}+i\tau}(\cos\eta) = \frac{2\sqrt{2}}{\pi}\frac{\left(\tau^2+\frac{1}{4}\right)}{\sin\eta}\int\limits_0^\eta \cosh(\tau t)\sqrt{\cos t-\cos\eta}\,dt.$$

We obtain

$$\begin{aligned}I_1(\xi,\eta,\tau_1) &= \frac{2\sqrt{2}}{\pi\sin\eta}\int\limits_0^\eta J_1(\xi,\tau_1,t)\sqrt{\cos t-\cos\eta}\,dt\\ &= \frac{2\sqrt{2}}{\pi\sin\eta}\,\frac{i}{\cosh(\pi\tau_1)}\\ &\quad\times\left(\mathrm{e}^{-i\xi\tau_1}\int\limits_0^\eta g(\xi,\tau_1,t)\,\sqrt{\cos t-\cos\eta}\,dt-2h_1(\xi,\eta)\sinh\tfrac{\xi}{2}\right),\quad \xi\neq 0,\end{aligned}$$

where

$$\begin{aligned}J_1(\xi,\tau_1,t) &= \mathop{\not\int}\limits_{-\infty}^{+\infty}\frac{\cosh(\tau t)}{\cosh(\pi\tau)}\frac{\mathrm{e}^{-i\xi\tau}\,d\tau}{\sinh[\pi(\tau_1-\tau)]}\\ &= \frac{i}{\cosh(\pi\tau_1)}\left(g(\xi,\tau_1,t)\,\mathrm{e}^{-i\xi\tau_1}-\frac{2\sinh\frac{\xi}{2}\cos\frac{t}{2}}{\cosh\xi-\cos t}\right),\end{aligned}\tag{42}$$

$$h_1(\xi,\eta) = \int\limits_0^\eta\frac{\sqrt{\cos t-\cos\eta}}{\cosh\xi-\cos t}\,\cos\tfrac{t}{2}\,dt = \frac{\pi}{\sqrt{2}}\left[\sqrt{1+\left(\frac{\sin\frac{\eta}{2}}{\sinh\frac{\xi}{2}}\right)^2}-1\right],\quad \xi\neq 0,$$

and function $g(\xi,\tau_1,t)$ is determined by (39). For $\xi=0$, expression (42) takes on finite values for all $t\in[0,\eta]$:

$$J_1(0,\tau_1,t) = \frac{\sinh(\tau_1 t)}{\cosh(\pi\tau_1)}\,\cot\tfrac{t}{2}.\tag{43}$$

Thus,

$$I_1(0,\eta,\tau_1) = \frac{2\sqrt{2}}{\pi\sin\eta}\int\limits_0^\eta J_1(0,\tau_1,t)\sqrt{\cos t-\cos\eta}\,dt,$$

and $G_1(\xi,\eta,\tau_1)=\frac{1}{2}\,I_1(\xi,\eta,\tau_1)\cosh(\pi\tau_1)$. □

PROPOSITION 4 (function $\theta(\xi,\eta)$). *If function $X(\mu)$ is represented at $\operatorname{Re}\mu = 0$ by Hilbert formula* (20)*, then the function $\theta(\xi,\eta)$ takes the form*

(44)
$$\theta(\xi,\eta) = -\frac{1}{2\pi i}\sqrt{\cosh\xi - \cos\eta}\int\limits_{-\infty}^{+\infty} Y(i\tau)\left(\tau\,\mathrm{P}_{-\frac{1}{2}+i\tau}(\cos\eta)\,\mathrm{e}^{-i\xi\tau} + G_2(\xi,\eta,\tau)\right)d\tau,$$

where

(45)
$$G_2(\xi,\eta,\tau) = \begin{cases} -\dfrac{2\sqrt{2}\,i}{\pi\sin^2\eta}\displaystyle\int\limits_0^{\eta}\Bigg[g(\xi,\tau,t)\left(\frac{3}{4}\cos t + \frac{1}{4}\cos\eta\right)\sqrt{\cos t - \cos\eta}\,\mathrm{e}^{-i\xi\tau} \\ \qquad\qquad -\dfrac{1}{3}\dfrac{\partial^2}{\partial\xi^2}\left(g(\xi,\tau,t)\,\mathrm{e}^{-i\xi\tau}\right)(\cos t - \cos\eta)^{\frac{3}{2}}\Bigg]dt \\ \qquad\qquad +\dfrac{i}{\sqrt{2}}\dfrac{\operatorname{sign}\xi}{\sqrt{\cosh\xi - \cos\eta}}, \quad \xi \neq 0, \\ -\dfrac{2\sqrt{2}}{\pi\sin^2\eta}\displaystyle\int\limits_0^{\eta}\Bigg[\cot\frac{t}{2}\,\sinh(\tau t)\left(\frac{3}{4}\cos t + \frac{1}{4}\cos\eta - \frac{1}{2}\right) \\ \qquad\qquad +\tau\,\cos^2\left(\dfrac{t}{2}\right)\cosh(\tau t)\Bigg]\sqrt{\cos t - \cos\eta}\,dt, \quad \xi = 0, \end{cases}$$

and $g(\xi,\tau,t)$ is defined by (39)*. Both integrals in* (45) *are regular and can be efficiently calculated by a Gaussian quadrature.*

Proof. Substituting formula (20) into (12), we obtain

(46)
$$\theta(\xi,\eta) = -\frac{1}{2\pi i}\sqrt{\cosh\xi - \cos\eta}\int\limits_{-\infty}^{+\infty}\left(\tau Y(i\tau) - \frac{1}{2}\,\xint-\limits_{-\infty}^{+\infty} Y(i\tau_1)\frac{\cosh(\pi\tau_1)}{\cosh(\pi\tau)}\frac{d\tau_1}{\sinh[\pi(\tau_1-\tau)]}\right) \times \mathrm{P}_{-\frac{1}{2}+i\tau}(\cos\eta)\,\mathrm{e}^{-i\xi\tau}d\tau.$$

Although the inner integral in (46) is singular, the external integral is regular. Consequently, we do not need the Poincaré–Bertrand formula for changing the order of integration in (46) (see [4]):

$$\begin{aligned}\mathcal{I}_X(\xi,\eta) &= \int\limits_{-\infty}^{+\infty}\left(\xint-\limits_{-\infty}^{+\infty} Y(i\tau_1)\,\frac{\cosh(\pi\tau_1)\,d\tau_1}{\sinh[\pi(\tau_1-\tau)]}\right)\frac{\mathrm{P}_{-\frac{1}{2}+i\tau}(\cos\eta)}{\cosh(\pi\tau)}\,\mathrm{e}^{-i\xi\tau}d\tau \\ &= \int\limits_{-\infty}^{+\infty} Y(i\tau_1)\cosh(\pi\tau_1)\underbrace{\left(\xint-\limits_{-\infty}^{+\infty}\frac{\mathrm{P}_{-\frac{1}{2}+i\tau}(\cos\eta)}{\cosh(\pi\tau)}\frac{\mathrm{e}^{-i\xi\tau}\,d\tau}{\sinh[\pi(\tau_1-\tau)]}\right)}_{I_2(\xi,\eta,\tau_1)}d\tau_1.\end{aligned}$$

The inner integral $I_2(\xi,\eta,\tau_1)$ in $\mathcal{I}_X(\xi,\eta)$ is calculated based on the following representation [1]:

$$\begin{aligned}\mathrm{P}_{-\frac{1}{2}+i\tau}(\cos\eta) =& \frac{4\sqrt{2}}{\pi\sin^2\eta}\int\limits_0^{\eta}\cosh(\tau t)\left(\cos\eta + \tfrac{1}{3}\left(\tau^2 + \tfrac{9}{4}\right)(\cos t - \cos\eta)\right) \\ &\times\sqrt{\cos t - \cos\eta}\,dt.\end{aligned}$$

We have

$$I_2(\xi,\eta,\tau_1) = \frac{4\sqrt{2}}{\pi\sin^2\eta} \times \int\limits_0^\eta \left(J_1(\xi,\tau_1,t)\ \cos\eta\sqrt{\cos t-\cos\eta} + \tfrac{1}{3}J_2(\xi,\tau_1,t)(\cos t-\cos\eta)^{\frac{3}{2}} \right) dt,$$

where the function $J_1(\xi,\tau_1,t)$ is defined by (42), and

$$J_2(\xi,\tau_1,t) = \fint\limits_{-\infty}^{+\infty} \frac{\cosh(\tau t)}{\cosh(\pi\tau)} \frac{\left(\tau^2+\frac{9}{4}\right)\mathrm{e}^{-i\xi\tau}\,d\tau}{\sinh[\pi(\tau_1-\tau)]} = \frac{9}{4}J_1(\xi,\tau_1,t) - \frac{\partial^2}{\partial\xi^2}J_1(\xi,\tau_1,t).$$

Using intermediate calculations

$$\begin{aligned} h_2(\xi,\eta) &= \int\limits_0^\eta \frac{(\cos t-\cos\eta)^{\frac{3}{2}}}{\cosh\xi-\cos t}\ \cos\tfrac{t}{2}\ dt \\ &= (\cosh\xi-\cos\eta)\,h_1(\xi,\eta) - \frac{\pi}{\sqrt{2}}\sin^2(\tfrac{\eta}{2}), \quad \xi\neq 0, \\ \left(2h_1(\xi,\eta)\cos\eta+\tfrac{3}{2}h_2(\xi,\eta)\right)\sinh\tfrac{\xi}{2} &- \frac{2}{3}\frac{\partial^2}{\partial\xi^2}\left(h_2(\xi,\eta)\sinh\tfrac{\xi}{2}\right) \\ &= \frac{\pi}{4}\frac{\sinh\xi}{|\sinh\xi|}\frac{\sin^2\eta}{\sqrt{\cosh\xi-\cos\eta}}, \quad \xi\neq 0, \end{aligned}$$

where the function $h_1(\xi,\eta)$ is defined by (40), we reduce the integral $I_2(\xi,\eta,\tau_1)$ to the form

$$I_2(\xi,\eta,\tau_1) = \frac{1}{\cosh(\pi\tau_1)}\left(R(\xi,\eta,\tau_1,t)\,\mathrm{e}^{-i\xi\tau_1} - \frac{\sqrt{2}\,i\ \operatorname{sign}\xi}{\sqrt{\cosh\xi-\cos\eta}} \right), \quad \xi\neq 0,$$

where

$$\begin{aligned} R(\xi,\eta,\tau_1,t)\,\mathrm{e}^{-i\xi\tau_1} =& \frac{4\sqrt{2}\,i}{\pi\sin^2\eta}\int\limits_0^\eta \Big[g(\xi,\tau_1,t)\left(\tfrac{3}{4}\cos t+\tfrac{1}{4}\cos\eta\right)\sqrt{\cos t-\cos\eta}\,\mathrm{e}^{-i\xi\tau_1} \\ &- \frac{1}{3}\frac{\partial^2}{\partial\xi^2}\left(g(\xi,\tau_1,t)\,\mathrm{e}^{-i\xi\tau_1}\right)(\cos t-\cos\eta)^{\frac{3}{2}} \Big]\, dt, \end{aligned}$$

and function $g(\xi,\tau_1,t)$ is defined by (39). In the case of $\xi=0$, we use (43) to derive expression for $I_2(0,\eta,\tau_1)$:

$$\begin{aligned} I_2(0,\eta,\tau_1) =& \frac{1}{\cosh(\pi\tau_1)}\ \frac{4\sqrt{2}}{\pi\sin^2\eta}\int\limits_0^\eta \Big[\cot\tfrac{t}{2}\ \sinh(\tau_1 t)\left(\tfrac{3}{4}\cos t+\tfrac{1}{4}\cos\eta-\tfrac{1}{2}\right) \\ &+ \tau_1\ \cos^2(\tfrac{t}{2})\ \cosh(\tau_1 t)\Big]\sqrt{\cos t-\cos\eta}\,dt. \end{aligned}$$

Note that the integrand in $I_2(0,\eta,\tau_1)$ takes on finite values for all $t\in[0,\eta]$. Finally, defining $G_2(\xi,\eta,\tau_1) = -\frac{1}{2}\ I_2(\xi,\eta,\tau_1)\cosh(\pi\tau_1)$, we finish the proof of the proposition. □

Remark (function $\theta(\xi,\eta)$). If we represent $\mathrm{P}_{-\frac{1}{2}+i\tau}(\cos\eta)$ by

$$\mathrm{P}_{-\frac{1}{2}+i\tau}(\cos\eta)=\frac{1}{\pi\sqrt{2}}\cosh(\pi\tau)\int\limits_{-\infty}^{+\infty}\frac{\mathrm{e}^{i\tau t}}{\sqrt{\cosh t+\cos\eta}}\,dt$$

(see [1]), then function (45) takes the form

(47)

$$\begin{aligned}G_2(\xi,\eta,\tau_1)&=-\frac{1}{2\pi\sqrt{2}}\,\cosh(\pi\tau_1)\int\limits_{-\infty}^{+\infty}\frac{1}{\sqrt{\cosh t+\cos\eta}}\left(\int\limits_{-\infty}^{+\infty}\frac{\mathrm{e}^{i\tau(t-\xi)}\,d\tau}{\sinh[\pi(\tau_1-\tau)]}\right)dt\\&=\frac{i}{2\pi\sqrt{2}}\,\cosh(\pi\tau_1)\int\limits_{-\infty}^{+\infty}\frac{\tanh\left(\frac{t-\xi}{2}\right)}{\sqrt{\cosh t+\cos\eta}}\,\mathrm{e}^{i\tau_1(t-\xi)}dt.\end{aligned}$$

Expression (47) is simpler than (45). However, although (47) is a regular integral, it is a Fourier integral on an infinite interval. Consequently, from a computational point of view, representation (45) is preferable.

5. Axially symmetric Stokes flow about a spindle-shaped body. Let us consider axially symmetric steady motion of a rigid spindle-shaped body in a Stokes fluid. In this case, the vector of velocity of fluid particles, $\mathbf{u}$, satisfies Stokes model (5). Suppose that the body moves in the fluid with constant velocity V_0 along its axis of symmetry. Let (r,φ,z) be a system of cylindrical coordinates with basis $(\mathbf{e}_r,\mathbf{e}_\varphi,\mathbf{k})$ such that axis z determines a body's axis of symmetry. Then boundary conditions for $\mathbf{u}$ are determined on the surface S of the body by

$$\mathbf{u}|_S=V_0\,\mathbf{k}. \tag{48}$$

We assume that the velocity $\mathbf{u}$ and the pressure function θ vanish at infinity,

$$\mathbf{u}|_\infty=0,\qquad \theta|_\infty=0. \tag{49}$$

The boundary-value problem (5), (48), and (49) is a classical problem in hydrodynamics of Stokes flows [8]. Since we consider only axially symmetric motion, boundary conditions (48) are reformulated for components of the vector $\mathbf{u}$ in cylindrical coordinates as

$$u_r(r,z)|_{\eta=\eta_0}=0,\qquad u_\varphi(r,z)\equiv 0,\qquad u_z(r,z)|_{\eta=\eta_0}=V_0, \tag{50}$$

where $\eta=\eta_0$ determines the contour of the spindle-shaped body in bipolar coordinates (ξ,η) in the meridional cross-section plane rz (see Figure 1).

The problem of the steady motion of a rigid body in a Stokes fluid is closely related to the problem of the Stokes flow about the body immersed in the viscous fluid [8]. The only difference is that in the latest problem, the body is immersed in the uniform flow, and the velocity of the flow is assumed to be constant at infinity. In this case, the boundary conditions take the form $\tilde{\mathbf{u}}|_S=\mathbf{0}$ and $\tilde{\mathbf{u}}|_\infty=-V_0\,\mathbf{k}$, where $\tilde{\mathbf{u}}$ is the velocity of the Stokes flow in this problem. Obviously, the velocities $\mathbf{u}$ and $\tilde{\mathbf{u}}$ are related by $\tilde{\mathbf{u}}=\mathbf{u}-V_0\,\mathbf{k}$.

5.1. Stream function. A standard approach to solving axially symmetric problems of Stokes flows is to represent the vector $\mathbf{u}$ by a stream function $\Psi(r,z)$ in cylindrical coordinates:

$$\mathbf{u} = -\operatorname{curl}(\Psi \mathbf{e}_\varphi). \tag{51}$$

In component form, (51) is rewritten as

$$u_r(r,z) = \frac{1}{r}\frac{\partial}{\partial z}(r\Psi), \qquad u_\varphi(r,z) \equiv 0, \qquad u_z(r,z) = -\frac{1}{r}\frac{\partial}{\partial r}(r\Psi). \tag{52}$$

The stream function Ψ is different from the stream function, Ψ_P, introduced by Payne and Pell [14] as $u_r = -\frac{1}{r}\frac{\partial \Psi_P}{\partial z}$, $u_z = \frac{1}{r}\frac{\partial \Psi_P}{\partial r}$ in the problem of the Stokes flow about a body immersed in a viscous fluid. If the velocity of the Stokes flow at infinity in Payne and Pell's problem is $-V_0\,\mathbf{k}$, then the stream functions Ψ and Ψ_P are related by $\Psi_P = -\left(r\Psi + \frac{1}{2}V_0 r^2\right)$.

Stokes model (5) and equation (51) imply that the stream function Ψ satisfies

$$(\operatorname{curl})^4(\Psi \mathbf{e}_\varphi) = 0.$$

This relation reduces to a so-called bi-1-harmonic equation

$$\Delta_1^2 \Psi(r,z) = 0, \tag{53}$$

where 1-harmonic operator Δ_1 is defined by (8) for $k=1$. Based on (52), boundary conditions (50) are reformulated as

$$\left(\frac{\partial}{\partial z}(r\Psi)\right)\Big|_{\eta=\eta_0} = 0, \qquad \left(\frac{\partial}{\partial r}(r\Psi)\right)\Big|_{\eta=\eta_0} = -V_0\ r|_{\eta=\eta_0}. \tag{54}$$

Substituting derivatives (11) into (54) and solving system (54) with respect to $(\frac{\partial}{\partial \xi}(r\Psi))|_{\eta=\eta_0}$ and $(\frac{\partial}{\partial \eta}(r\Psi))|_{\eta=\eta_0}$, we obtain

$$\begin{aligned}
\left(\frac{\partial}{\partial \xi}(r\Psi)\right)\Big|_{\eta=\eta_0} &= V_0 c^2 \frac{\sinh\xi \sin^2\eta_0}{(\cosh\xi - \cos\eta_0)^3},\\
\left(\frac{\partial}{\partial \eta}(r\Psi)\right)\Big|_{\eta=\eta_0} &= -V_0 c^2 \sin\eta_0 \frac{(\cosh\xi\cos\eta_0 - 1)}{(\cosh\xi - \cos\eta_0)^3},
\end{aligned} \tag{55}$$

where c is a metric parameter of bipolar coordinates. Integrating the first equation of (55) with respect to ξ, we have

$$\Psi|_{\eta=\eta_0} = -\frac{V_0 c}{2}\frac{\sin\eta_0}{\cosh\xi - \cos\eta_0} + \tilde{c}\,\frac{(\cosh\xi - \cos\eta_0)}{\sin\eta_0},$$

where $\tilde{c}$ is a constant of integration. Assuming that Ψ is bounded at $\xi \to \infty$, we put $\tilde{c} = 0$, and consequently,

$$\Psi|_{\eta=\eta_0} = -\frac{V_0 c}{2}\frac{\sin\eta_0}{\cosh\xi - \cos\eta_0}. \tag{56}$$

Note that for spatial doubly connected bodies, e.g., torus [21], $\tilde{c} \neq 0$. Substituting (56) into the second equation of (55), we obtain

$$\left(\frac{\partial \Psi}{\partial \eta}\right)\Big|_{\eta=\eta_0} = -\frac{V_0 c}{2}\frac{(\cosh\xi\cos\eta_0 - 1)}{(\cosh\xi - \cos\eta_0)^2}. \tag{57}$$

Thus, the boundary-value problem (5), (48), and (49) reduces to finding the stream function Ψ that satisfies bi-1-harmonic equation (53) and boundary conditions (56) and (57).

In our paper [30], we showed that a solution to (53) may be represented by one harmonic function $\Phi_0(r,z)$ and one 1-harmonic function $\Phi_1(r,z)$ in the form

$$\Psi(r,z) = \frac{1}{2}\left(r^2+z^2-c^2\right)\Phi_1(r,z) + r\,\Phi_0(r,z), \tag{58}$$

where

$$\Delta_1\,\Phi_1(r,z) = 0, \qquad \Delta\,\Phi_0(r,z) = 0.$$

In bipolar coordinates (ξ,η), functions Φ_0 and Φ_1 are represented by Fourier integrals:

$$\Phi_1(\xi,\eta) = \frac{1}{2\pi i c^2}\sqrt{\cosh\xi-\cos\eta}\int\limits_{-i\infty}^{+i\infty} A(\mu)\,\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta)\,\mathrm{e}^{-\xi\mu}d\mu, \quad \eta\le\eta_0, \tag{59}$$

$$\Phi_0(\xi,\eta) = \frac{1}{2\pi i c}\sqrt{\cosh\xi-\cos\eta}\int\limits_{-i\infty}^{+i\infty} B(\mu)\,\mathrm{P}_{-\frac{1}{2}+\mu}(\cos\eta)\,\mathrm{e}^{-\xi\mu}d\mu, \quad \eta\le\eta_0, \tag{60}$$

where $A(\mu)$ and $B(\mu)$ are meromorphic functions in the strip $-1\le\operatorname{Re}\mu\le 1$. Representations (58), (60), and (59) reduce the function Ψ to the form

$$\Psi(\xi,\eta) = \frac{1}{2\pi i\sqrt{\cosh\xi-\cos\eta}}\left(\cos\eta\int\limits_{-i\infty}^{+i\infty} A(\mu)\,\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta)\,\mathrm{e}^{-\xi\mu}d\mu + \sin\eta\int\limits_{-i\infty}^{+i\infty} B(\mu)\,\mathrm{P}_{-\frac{1}{2}+\mu}(\cos\eta)\,\mathrm{e}^{-\xi\mu}d\mu\right), \quad \eta\le\eta_0. \tag{61}$$

To express right-hand sides of (56) and (57) by Fourier integrals, we use the relation

$$\frac{(2k-1)!!}{2^k}\frac{\sin^k\eta}{(\cosh\xi+\cos\eta)^{k+\frac{1}{2}}} = \frac{1}{i\sqrt{2}}\int\limits_{b-i\infty}^{b+i\infty}\frac{1}{\cos(\pi\mu)}\,\mathrm{P}^{(k)}_{-\frac{1}{2}+\mu}(\cos\eta)\,\mathrm{e}^{-\xi\mu}d\mu,$$

where $k\ge 0$, $(2k-1)!! = \prod_{j=1}^{k}(2j-1)$, and $b\in\left(-k-\frac{1}{2},k+\frac{1}{2}\right)$; see [1]. For $k=0$, we define $(-1)!!=1$.

Thus, boundary conditions (56) and (57) reduce to a system of linear equations with respect to functions $A(\mu)$ and $B(\mu)$:

$$\begin{pmatrix} \cos\eta_0\,\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta_0) & \sin\eta_0\,\mathrm{P}_{-\frac{1}{2}+\mu}(\cos\eta_0) \\ \cos\eta_0\,\mathrm{P}^{(2)}_{-\frac{1}{2}+\mu}(\cos\eta_0)-\sin\eta_0\,\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta_0) & \sin\eta_0\,\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta_0) \end{pmatrix}\begin{pmatrix} A(\mu) \\ B(\mu)\end{pmatrix} = \frac{\pi c V_0}{\sqrt{2}}\frac{\sin\eta_0}{\cos(\pi\mu)}\begin{pmatrix} -\mathrm{P}_{-\frac{1}{2}+\mu}(-\cos\eta_0) \\ \mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(-\cos\eta_0)\end{pmatrix}. \tag{62}$$

Let $D(\mu)$ denote the determinant of system (62):

$$
\begin{aligned}
D(\mu) &= (1+\cos^2\eta_0)\mathrm{P}_{-\frac{1}{2}+\mu}(\cos\eta_0)\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta_0) \\
&\quad + \sin\eta_0\cos\eta_0\left(\left(\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta_0)\right)^2 + \left(\mu^2-\tfrac{1}{4}\right)\left(\mathrm{P}_{-\frac{1}{2}+\mu}(\cos\eta_0)\right)^2\right);
\end{aligned}
\tag{63}
$$

then using a relation for the associated Legendre functions

$$
\mathrm{P}_{-\frac{1}{2}+\mu}(-\cos\eta_0)\,\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta_0)+\mathrm{P}_{-\frac{1}{2}+\mu}(\cos\eta_0)\,\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(-\cos\eta_0)=\frac{2}{\pi}\frac{\cos(\pi\mu)}{\sin\eta_0}
\tag{64}
$$

(see [1]), we obtain a solution to system (62):

$$
\begin{aligned}
A(\mu) &= -V_0 c\sqrt{2}\,\frac{\sin\eta_0}{D(\mu)}, \\
B(\mu) &= \frac{V_0 c}{\sqrt{2}}\left(\frac{2\cos\eta_0}{D(\mu)}\,\frac{\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta_0)}{\mathrm{P}_{-\frac{1}{2}+\mu}(\cos\eta_0)}-\frac{\pi}{\cos(\pi\mu)}\,\frac{\mathrm{P}_{-\frac{1}{2}+\mu}(-\cos\eta_0)}{\mathrm{P}_{-\frac{1}{2}+\mu}(\cos\eta_0)}\right).
\end{aligned}
\tag{65}
$$

Consequently, the velocity vector, $\mathbf{u}$, that solves problem (5), (48), and (49) is expressed analytically by (52), (61), and (65). As an illustration, we calculated streamlines about a rigid spindle-shaped body by solving the equation

$$
r\,\Psi(r,z)+\frac{1}{2}V_0\,r^2=C
\tag{66}
$$

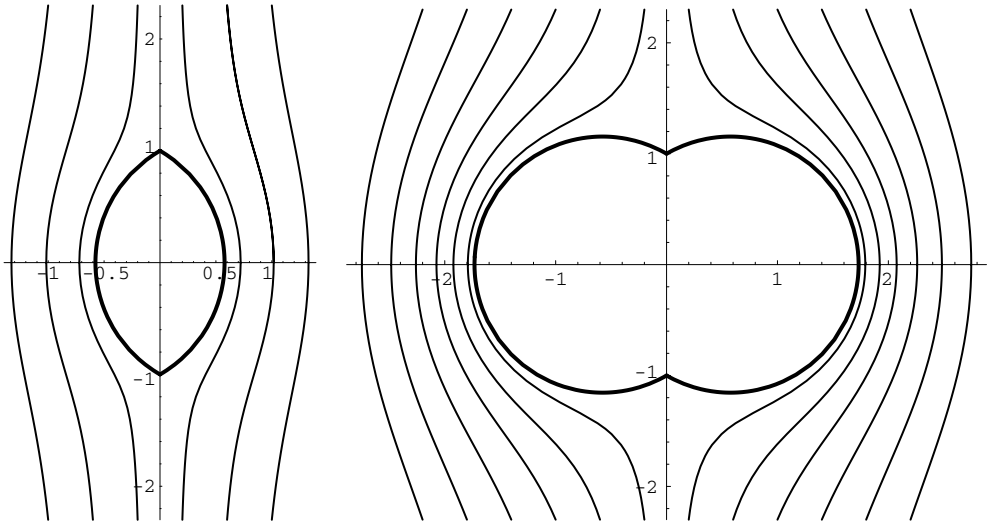
with respect to pairs (r,z) for different values of constant C. It should be noted that (66), in fact, determines streamlines about the body immersed in the uniform Stokes flow with the constant velocity, $-V_0\,\mathbf{k}$, at infinity, while the stream function Ψ corresponds to the motion of the body with the constant velocity $V_0\,\mathbf{k}$. We obtain (66) based on the fact that in terms of Payne and Pell's stream function, Ψ_P, streamlines are defined by Ψ_P = constant, and that Ψ and Ψ_P are related by $\Psi_P = -\left(r\,\Psi+\frac{1}{2}V_0\,r^2\right)$. We used Mathematica 5 to solve (66). Figure 3 shows streamlines about a rigid spindle-shaped body for $\eta_0=\frac{2\pi}{3}$ and $\eta_0=\frac{\pi}{3}$, respectively. Streamlines may also be calculated based on the relation $\frac{dr}{dz}=u_r/\left(u_z-V_0\right)$; see [8].

For the case of sphere, $\eta_0=\frac{\pi}{2}$, the stream function Ψ in cylindrical coordinates takes the form

$$
\Psi=\frac{cV_0}{4}\frac{r}{\sqrt{r^2+z^2}}\left(\frac{c^2}{r^2+z^2}-3\right).
$$

5.2. Analysis of the determinant $D(\mu)$ and functions $A(\mu)$ and $B(\mu)$. Asymptotic behavior of functions (59) and (60) at $\xi\to\infty$ is determined by zeros of determinant (63). The function $D(\mu)$ is even, i.e., $D(-\mu)=D(\mu)$, and equals zero at $\mu=\pm\frac{1}{2}$ and $\mu=\pm\frac{3}{2}$ for all $\eta_0\in(0,\pi)$:

$$
\begin{aligned}
D(\mu)\big|_{\mu\to\pm\frac{1}{2}} &\to -\tfrac{(1-\cos\eta_0)^2}{\sin\eta_0}\left(|\mu|-\tfrac{1}{2}\right), \\
D(\mu)\big|_{\mu\to\pm\frac{3}{2}} &\to -\tfrac{(1-\cos\eta_0)^2}{\sin\eta_0}(1+2\cos\eta_0)\left(|\mu|-\tfrac{3}{2}\right).
\end{aligned}
$$



$\eta_0 = \frac{2\pi}{3}$ $\eta_0 = \frac{\pi}{3}$

FIG. 3. *Streamlines about a rigid spindle-shaped body for* $\eta_0 = \frac{2\pi}{3}$ *and* $\eta_0 = \frac{\pi}{3}$*, respectively.*

TABLE 1
First individual root for $D(\mu)$*.*

η_0	μ_0	η_0	μ_0
$2\pi/12$	$8.564 + i\ 2.614$	$7\pi/12$	1.897
$3\pi/12$	$5.740 + i\ 1.569$	$8\pi/12$	$1.5^{\ddagger}$
$4\pi/12$	$4.341 + i\ 0.960$	$9\pi/12$	1.211
$5\pi/12$	$3.517 + i\ 0.437$	$2.6037^{\S}$	1.000
$9\pi/20^{\dagger}$	$3.345 + i\ 0.000$	$10\pi/12$	0.989
$6\pi/12$	2.5	$11\pi/12$	0.807

† For $\eta_0 \geq 9\pi/20$, the first individual root is real.
‡ Since the root $\frac{3}{2}$ is generic, it becomes the root of multiplicity 2 for $8\pi/12$.
§ For $\eta_0 \geq 2.6037$, the first individual root lies within the strip $-1 \leq \operatorname{Re}\mu \leq 1$.

We call values $\mu = \pm\frac{1}{2}$ and $\mu = \pm\frac{3}{2}$ generic roots of $D(\mu)$. Consequently, functions $A(\mu)$ and $B(\mu)$ have simple poles at $\mu = \pm\frac{1}{2}$ and $\mu = \pm\frac{3}{2}$ with corresponding residues determined by

$$\operatorname*{Res}_{\mu=\pm\frac{1}{2}} A(\mu) = \pm V_0 c\sqrt{2}\left(\tfrac{1+\cos\eta_0}{1-\cos\eta_0}\right), \qquad \operatorname*{Res}_{\mu=\pm\frac{3}{2}} A(\mu) = \mp V_0 c\sqrt{2}\left(\tfrac{1+\cos\eta_0}{1-\cos\eta_0}\right)\tfrac{1}{1+2\cos\eta_0},$$

$$\operatorname*{Res}_{\mu=\pm\frac{1}{2}} B(\mu) = \pm\tfrac{V_0 c}{\sqrt{2}}, \qquad \operatorname*{Res}_{\mu=\pm\frac{3}{2}} B(\mu) = \mp\tfrac{V_0 c}{\sqrt{2}}\tfrac{(1+\cos\eta_0+2\cos^2\eta_0)}{(1-\cos\eta_0)(1+2\cos\eta_0)}.$$

Except for generic roots, the determinant $D(\mu)$ has individual roots for any $\eta_0 \in (0, \pi)$. Table 1 presents the first individual root, μ_0, for different η_0.

Using asymptotic formulas for functions $\mathrm{P}_{-\frac{1}{2}+i\tau}(\cos\eta_0)$ and $\mathrm{P}^{(1)}_{-\frac{1}{2}+i\tau}(\cos\eta_0)$ at $\tau\to\infty$

$$\mathrm{P}_{-\frac{1}{2}+i\tau}(\cos\eta_0)\Big|_{\tau\to\infty}=\frac{\mathrm{e}^{\eta_0|\tau|}}{\sqrt{2\pi\sin\eta_0}}\left(1+\frac{\cot\eta_0}{8\tau}+O\left(\frac{1}{\tau^2}\right)\right),$$
$$\mathrm{P}^{(1)}_{-\frac{1}{2}+i\tau}(\cos\eta_0)\Big|_{\tau\to\infty}=\frac{|\tau|\,\mathrm{e}^{\eta_0|\tau|}}{\sqrt{2\pi\sin\eta_0}}\left(1-\frac{3\cot\eta_0}{8\tau}+O\left(\frac{1}{\tau^2}\right)\right)$$

(see [1]), we determine asymptotic behavior of $D(i\tau)$, $A(i\tau)$, and $B(i\tau)$:

$$D(i\tau)|_{\tau\to\infty}=\frac{\mathrm{e}^{2\eta_0|\tau|}}{2\pi\sin\eta_0}\left(1+O\left(\frac{1}{\tau}\right)\right),$$
$$A(i\tau)|_{\tau\to\infty}=V_0c\,2\pi\sqrt{2}\,(\sin^2\eta_0)\,\mathrm{e}^{-2\eta_0|\tau|}\left(1+O\left(\frac{1}{\tau}\right)\right),$$
$$B(i\tau)|_{\tau\to\infty}=V_0c\,\pi\sqrt{2}\,\sin(2\eta_0)\,|\tau|\,\mathrm{e}^{-2\eta_0|\tau|}\left(1+O\left(\frac{1}{\tau}\right)\right).$$

For the case of sphere, $\eta_0=\frac{\pi}{2}$, functions $D(\mu)$, $A(\mu)$, and $B(\mu)$ take the form

$$D(\mu)=\frac{1}{\pi}\cos(\pi\mu),\qquad A(\mu)=-\frac{V_0c\,\pi\sqrt{2}}{\cos(\pi\mu)},\qquad B(\mu)=-\frac{V_0c\,\pi}{\sqrt{2}}\frac{1}{\cos(\pi\mu)}.$$

6. Hilbert formulas in hydrodynamics of Stokes flows. In this section, we analyze basic hydrodynamic characteristics: vorticity, pressure, and drag force. We use Hilbert formula (20) for analytic representation of the pressure function θ via a vortex function.

6.1. Vorticity and scalar vortex function. The vorticity, $\boldsymbol{\omega}$, is defined by (4). In the case of axially symmetric boundary-value conditions, it may be represented as

$$\boldsymbol{\omega}=-\operatorname{curl}(\operatorname{curl}(\Psi\mathbf{e}_\varphi))=\omega(r,z)\,\mathbf{e}_\varphi,$$

where $\omega(r,z)$ is a scalar vortex function given by

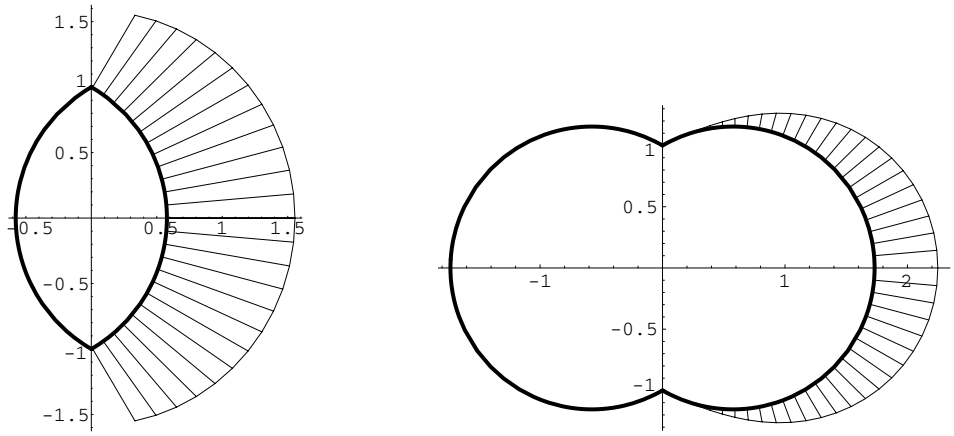
$$\omega(r,z)=\Delta_1\Psi(r,z).$$

Since the stream function Ψ is bi-1-harmonic, the vortex function $\omega(r,z)$ satisfies 1-harmonic equation (9), and in terms of functions Φ_0 and Φ_1, it takes the form

$$\omega(r,z)=2\left(r\frac{\partial}{\partial r}+z\frac{\partial}{\partial z}+\frac{3}{2}\right)\Phi_1+2\frac{\partial\Phi_0}{\partial r}. \tag{67}$$

Consequently, representation of ω by $A(\mu)$ and $B(\mu)$ is straightforward. At the contour $\eta=\eta_0$, the function ω is determined by

$$\omega(\xi,\eta)|_{\eta=\eta_0}=\frac{V_0\sqrt{2}}{\pi ic}\,(\cosh\xi-\cos\eta_0)^{\frac{3}{2}}\int\limits_{-i\infty}^{+i\infty}\frac{1}{D(\mu)}\,\mathrm{P}^{(2)}_{-\frac{1}{2}+\mu}(\cos\eta_0)\,\mathrm{e}^{-\xi\mu}d\mu.$$



$\eta_0 = \frac{2\pi}{3}$ $\eta_0 = \frac{\pi}{3}$

FIG. 4. *Epures for the vortex function,* $\frac{c}{2V_0}\,\omega(\xi,\eta)|_{\eta=\eta_0}$*, at the surface of a rigid spindle-shaped body for* $\eta_0 = \frac{2\pi}{3}$ *and* $\eta_0 = \frac{\pi}{3}$*, respectively. At a particular point on the contour, the value of the function is depicted by the length of the outward normal line if the value is positive and by the length of the inward normal line if the value is negative.*

Figure 4 illustrates behavior of $\frac{c}{2V_0}\,\omega(\xi,\eta)|_{\eta=\eta_0}$ for $\eta_0 = \frac{2\pi}{3}$ and $\eta_0 = \frac{\pi}{3}$. Since in the case of $\eta_0 > \frac{2\pi}{3}$, the determinant $D(\mu)$ has first individual root $\mu_0 < \frac{3}{2}$, the vortex function $\omega(\xi,\eta)|_{\eta=\eta_0}$ diverges at $\xi \to \infty$. For $\eta_0 = \frac{2\pi}{3}$, it takes on a nonzero finite value at $\xi \to \infty$:

$$\lim_{\xi\to\infty} \omega(\xi,\eta)|_{\eta=\eta_0=\frac{2\pi}{3}} = 1.275\,\frac{V_0}{c}.$$

For the case of sphere, $\eta_0 = \frac{\pi}{2}$, the vortex function takes the form

$$\omega(r,z) = \frac{3}{2}\,c\,V_0\frac{r}{(r^2+z^2)^{\frac{3}{2}}}.$$

In this case, if the function ω is represented by Fourier integral (13), then $Y(\mu) = \frac{3\pi}{\sqrt{2}}\frac{V_0}{c}\frac{1}{\cos(\pi\mu)}$.

6.2. Pressure. We associate the function θ in Stokes model (5) with the pressure in a Stokes fluid. In an axially symmetric case, the pressure θ and the vortex function ω are independent of angular coordinate φ and may be considered as real and imaginary parts of an r-analytic function $F(r,z) = \theta(r,z) + i\,r\,\omega(r,z)$ that satisfies the generalized Cauchy–Riemann system (7). Consequently, we may use Hilbert formula (20) to express θ via ω.

PROPOSITION 5 (pressure). *Let the vortex function* ω *be determined by* (67)*. Then for* $\eta_0 < 2.6037$*, the pressure* θ *is a real-valued function represented by*

(68)
$$\theta(\xi,\eta)=\frac{1}{\pi i c^2}\sqrt{\cosh\xi-\cos\eta}\left(\cosh\xi\sin\eta\int\limits_{-\infty}^{+\infty}\tau\,A(i\tau)\,\mathrm{P}^{(1)}_{-\frac{1}{2}+i\tau}(\cos\eta)\,\mathrm{e}^{-i\xi\tau}d\tau\right.$$
$$-(\cosh\xi\cos\eta-1)\int\limits_{-\infty}^{+\infty}\tau\,B(i\tau)\,\mathrm{P}_{-\frac{1}{2}+i\tau}(\cos\eta)\,\mathrm{e}^{-i\xi\tau}d\tau$$
$$-i\,\sinh\xi\sin\eta\int\limits_{-\infty}^{+\infty}\left(\tfrac{1}{2}A(i\tau)+B(i\tau)\right)\mathrm{P}^{(1)}_{-\frac{1}{2}+i\tau}(\cos\eta)\,\mathrm{e}^{-i\xi\tau}d\tau$$
$$-i\,\sinh\xi\cos\eta\int\limits_{-\infty}^{+\infty}\left(\left(\tau^2+\tfrac{1}{4}\right)A(i\tau)+\tfrac{1}{2}B(i\tau)\right)\mathrm{P}_{-\frac{1}{2}+i\tau}(\cos\eta)\,\mathrm{e}^{-i\xi\tau}d\tau$$
$$+\frac{3}{2}\int\limits_{-\infty}^{+\infty}\left[-\tau\,A(i\tau)+\frac{1}{2\cosh(\pi\tau)}\oint\limits_{-\infty}^{+\infty}A(i\tau_1)\frac{\cosh(\pi\tau_1)\,d\tau_1}{\sinh[\pi(\tau_1-\tau)]}\right]$$
$$\left.\times\,\mathrm{P}_{-\frac{1}{2}+i\tau}(\cos\eta)\,\mathrm{e}^{-i\xi\tau}d\mu\right).$$

The double integral in (68) *can be efficiently calculated by* (45).

Proof. Using representation (67) and the fact that $\Delta\,\Phi_0=0$ and $\Delta_1\,\Phi_1=0$, we obtain identities

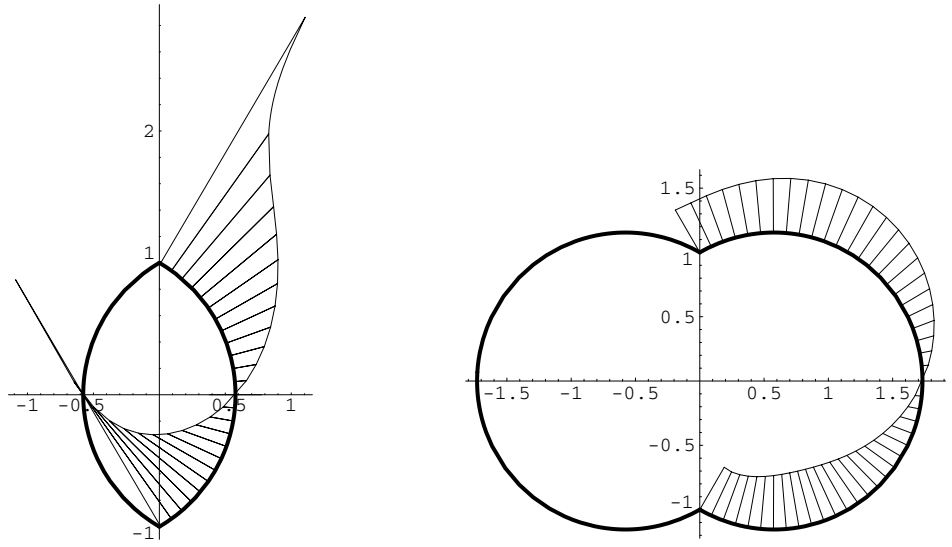
$$\frac{\partial\omega}{\partial z}\equiv 2\,\frac{\partial}{\partial z}\left[\left(r\frac{\partial}{\partial r}+z\frac{\partial}{\partial z}+\frac{3}{2}\right)\Phi_1+\frac{\partial\Phi_0}{\partial r}\right]$$
$$=2\,\frac{\partial}{\partial r}\left[-\frac{z}{r}\frac{\partial}{\partial r}\,(r\,\Phi_1)+r\frac{\partial\Phi_1}{\partial z}+\frac{\partial\Phi_0}{\partial z}\right]+3\,\frac{\partial\Phi_1}{\partial z},$$
$$-\frac{1}{r}\frac{\partial}{\partial r}\,(r\omega)\equiv-\frac{2}{r}\frac{\partial}{\partial r}\left(r\left[\left(r\frac{\partial}{\partial r}+z\frac{\partial}{\partial z}+\frac{3}{2}\right)\Phi_1+\frac{\partial\Phi_0}{\partial r}\right]\right)$$
$$=2\,\frac{\partial}{\partial z}\left[-\frac{z}{r}\frac{\partial}{\partial r}\,(r\,\Phi_1)+r\frac{\partial\Phi_1}{\partial z}+\frac{\partial\Phi_0}{\partial z}\right]-3\,\frac{1}{r}\frac{\partial}{\partial r}\,(r\,\Phi_1).$$

If we represent the pressure function θ by

$$\theta=-2\,\frac{z}{r}\frac{\partial}{\partial r}\,(r\,\Phi_1)+2\,r\frac{\partial\Phi_1}{\partial z}+2\,\frac{\partial\Phi_0}{\partial z}+3\,\tilde{\theta}+\tilde{c},\tag{69}$$

where $\tilde{\theta}=\tilde{\theta}(r,z)$ is a new function, and $\tilde{c}$ is a constant, then system (7) for functions θ and ω reduces to (7) for functions $\tilde{\theta}$ and Φ_1, i.e., $\tilde{F}(r,z)=\tilde{\theta}(r,z)+i\,r\,\Phi_1(r,z)$ is an r-analytic function. In an axially symmetric case, $\theta(-\xi,\eta)=-\theta(\xi,\eta)$, $\Phi_0(-\xi,\eta)=\Phi_0(\xi,\eta)$, $\Phi_1(-\xi,\eta)=\Phi_1(\xi,\eta)$, and $\tilde{\theta}(-\xi,\eta)=-\tilde{\theta}(\xi,\eta)$. Consequently, the left-hand side in (69) is an odd function with respect to ξ, only if $\tilde{c}=0$.

Recall that the function Φ_1 is represented by Fourier integral (59) with the density $A(\mu)$ determined by (65). For $\eta_0<2.6037$, $A(\mu)$ is meromorphic within the strip $-1\le\mathrm{Re}\,\mu\le 1$ with only two simple poles at $\mu=\pm\frac{1}{2}$, i.e., it belongs to the space $\mathcal{M}_{[-1,1]}$. Let $\tilde{\theta}$ be represented by Fourier integral (12) with density $X(\mu)\in\mathcal{M}_{[-1,1]}$. Consequently, functions $A(\mu)$ and $X(\mu)$ satisfy conditions of Theorem 1. We use Hilbert formula (20) to represent $X(i\tau)$ by $A(i\tau)$ and then express (69) in terms of $A(i\tau)$ and $B(i\tau)$, where $\tau\in\mathbb{R}$. ∎



$\eta_0 = \frac{2\pi}{3}$ $\eta_0 = \frac{\pi}{3}$

FIG. 5. *Epures for the pressure function,* $\frac{c}{2V_0}\,\theta(\xi,\eta_0)$*, at the surface of a rigid spindle-shaped body for* $\eta_0 = \frac{2\pi}{3}$ *and* $\eta_0 = \frac{\pi}{3}$*, respectively. At a particular point on the contour, the value of the function is depicted by the length of the outward normal line if the value is positive and by the length of the inward normal line if the value is negative.*

TABLE 2
Function $\frac{c}{V_0}\,\theta(\xi,\eta)|_{\eta=\eta_0}$ *at* $\xi \to +\infty$ *for different* η_0*.*

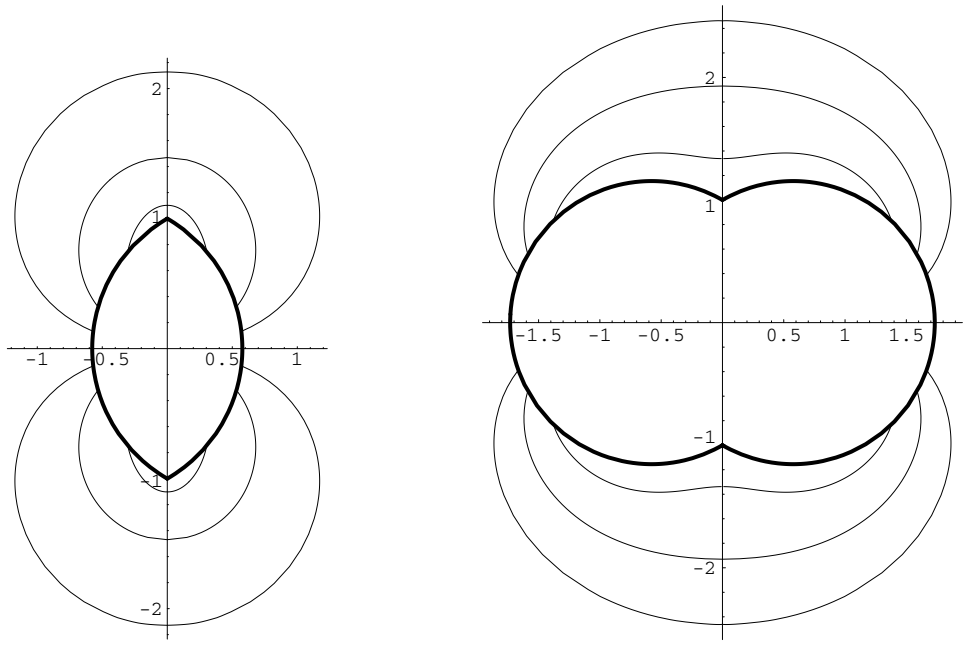
η_0	$\frac{c}{V_0}\,\theta(+\infty,\eta_0)$	η_0	$\frac{c}{V_0}\,\theta(+\infty,\eta_0)$
$\pi/12$	0.164	$5\pi/12$	1.034
$2\pi/12$	0.338	$6\pi/12$	1.5
$3\pi/12$	0.526	$7\pi/12$	2.692
$4\pi/12$	0.746	$8\pi/12$	$+\infty$

As an illustration to (68), epures of the pressure, $\frac{c}{V_0}\,\theta(\xi,\eta)|_{\eta=\eta_0}$, are calculated at the contour of a spindle-shaped body for $\eta_0 = \frac{2\pi}{3}$ and $\eta_0 = \frac{2\pi}{3}$; see Figure 5. For $0 < \eta_0 < \frac{2\pi}{3}$, the function $\theta(\xi,\eta)|_{\eta=\eta_0}$ takes on finite nonzero values at $\xi \to +\infty$:

$$\lim_{\xi\to+\infty}\theta(\xi,\eta)|_{\eta=\eta_0} = \frac{6V_0}{c}\left(-\frac{\cot^2\left(\frac{\eta_0}{2}\right)\cos\eta_0}{1+2\cos\eta_0} + \frac{1}{2\pi}\sin\eta_0\int\limits_0^{+\infty}\frac{d\tau}{D(i\tau)}\right).$$

Table 2 presents values of $\frac{c}{V_0}\,\theta(\xi,\eta)|_{\eta=\eta_0}$ at the tip point, i.e., $\xi \to +\infty$, for different η_0.

Figure 6 shows isobars about a spindle-shaped body for $\eta_0 = \frac{2\pi}{3}$ and $\eta_0 = \frac{2\pi}{3}$. Isobars are determined by equation $\theta(\xi,\eta) = C$ for different values of constant C. To solve this equation numerically, we represented $\theta(\xi,\eta)$ by (68) and used Mathematica 5. An alternative approach for computing isobars is based on the fact that at an isobar the relation $d\theta = \frac{\partial\theta}{\partial r}dr + \frac{\partial\theta}{\partial z}dz = 0$ holds. Consequently, using system (7), we obtain the explicit first-order differential equation $\frac{dz}{dr} = -\frac{\partial\theta}{\partial r}\Big/\frac{\partial\theta}{\partial z} = \frac{\partial\omega}{\partial z}\Big/\frac{1}{r}\frac{\partial}{\partial r}(r\omega)$, which can be solved by the Runge–Kutta method. We compared both approaches with respect to running time and accuracy. In comparison to the alternative approach, solving



$\eta_0 = \frac{2\pi}{3}$ $\eta_0 = \frac{\pi}{3}$

FIG. 6. *Isobars about a rigid spindle-shaped body for* $\eta_0 = \frac{2\pi}{3}$ *and* $\eta_0 = \frac{\pi}{3}$*, respectively.*

$\theta(\xi, \eta) = C$ is faster and more accurate. This proves superiority of the analytical solution based on the Hilbert formula.

For the case of sphere, $\eta_0 = \frac{\pi}{2}$, the pressure function takes the form

$$\theta(r, z) = \frac{3}{2}\, c\, V_0 \frac{z}{(r^2 + z^2)^{\frac{3}{2}}}.$$

In this case, if the function θ is represented by Fourier integral (12), then $X(\mu) = \frac{3\pi}{\sqrt{2}} \frac{V_0}{c} \frac{\mu}{\cos(\pi\mu)}$.

6.3. Drag force. Drag force is the characteristics that attracts most of the attention devoted to problems of motion of rigid bodies in a viscous fluid [8]. Approximate calculation of drag force by means of variational principles is discussed in [9]. We derive an analytical formula for the drag force exerted on a rigid spindle-shaped body using expressions for pressure and vortex functions obtained in previous sections.

PROPOSITION 6 (drag force). *The magnitude of the force exerted by a Stokes fluid on the spindle-shaped body is determined by*

$$F_0 = -4\sqrt{2}\,\rho \int_0^{+\infty} \left(\left(\tau^2 + \tfrac{1}{4}\right) A(i\tau) + 2B(i\tau)\right) d\tau, \tag{70}$$

where ρ *is the shear viscosity, and functions* $A(\mu)$ *and* $B(\mu)$ *are given by* (65).

Proof. Let $\mathbf{n} = n_r \mathbf{e}_r + n_z \mathbf{k}$ be the outer normal to the surface of the body, S, where $(\mathbf{e}_r, \mathbf{e}_\varphi, \mathbf{k})$ is the basis of the system of cylindric coordinates. By definition,

$n_r = \frac{\partial r}{\partial n}$ and $n_z = \frac{\partial z}{\partial n}$. The force, exerted by the fluid at a particular point and acting in the direction $\mathbf{n}$, is given by

$$\frac{1}{2\rho}\mathbf{P}_n = (\mathbf{n}\cdot\operatorname{grad})\,\mathbf{u} + \frac{1}{2}\,[\mathbf{n}\times\operatorname{curl}\mathbf{u}] - \frac{1}{2}\,\theta\,\mathbf{n};$$

see [21]. Since the body moves along its axis of symmetry, the resultant force has only the component in the direction $\mathbf{k}$. Thus, the magnitude of the total drag force is the integral of the projection $\mathbf{P}_n$ onto $(-\mathbf{k})$ over the surface S:

$$\frac{1}{2\rho}F_0 = -\frac{1}{2\rho}\iint\limits_S \mathbf{P}_n\cdot\mathbf{k}\,dS = -\iint\limits_S\left(\left(n_r\frac{\partial}{\partial r}+n_z\frac{\partial}{\partial z}\right)u_z + \frac{1}{2}\,\omega\,n_r - \frac{1}{2}\,\theta\,n_z\right)dS.$$

To simplify this expression, we use representation (52), formula $dS = r\,d\varphi\,ds$, and relations

$$n_r = \frac{\partial z}{\partial s}, \qquad n_z = -\frac{\partial r}{\partial s},$$
$$\frac{\partial}{\partial s} = \frac{1}{h}\frac{\partial}{\partial\xi}, \qquad \frac{\partial}{\partial n} = -\frac{1}{h}\frac{\partial}{\partial\eta},$$

where $ds = h\,d\xi$ is the element of the contour of the surface S in the meridional cross-section plane rz, and $h = \frac{c}{\cosh\xi - \cos\eta_0}$ is the Lame coefficient. Directional derivative $\frac{\partial}{\partial s}$ corresponds to the vector $\mathbf{s}$, which is orthogonal to $\mathbf{n}$ and oriented towards an increase of coordinate ξ. We have

$$\left(n_r\frac{\partial}{\partial r}+n_z\frac{\partial}{\partial z}\right)u_z = -\omega\,n_r + \frac{1}{r}\frac{\partial}{\partial s}\left(r\frac{\partial\Psi}{\partial z}\right)$$

and

$$\iint\limits_S \frac{1}{r}\frac{\partial}{\partial s}\left(r\frac{\partial\Psi}{\partial z}\right)dS = 2\pi\int\limits_{-\infty}^{+\infty}\frac{\partial}{\partial\xi}\left(r\frac{\partial\Psi}{\partial z}\right)\Big|_{\eta=\eta_0}d\xi = 2\pi\left(r\frac{\partial\Psi}{\partial z}\right)\Big|_{(\xi,\eta)=(-\infty,\eta_0)}^{(\xi,\eta)=(+\infty,\eta_0)} = 0.$$

Thus, the expression for the total drag force reduces to

$$\frac{1}{2\rho}F_0 = \frac{1}{2}\iint\limits_S(\omega\,n_r + \theta\,n_z)\,dS. \tag{71}$$

Using representations (67) and (69) for functions ω and θ, respectively, we obtain

$$\omega\,n_r + \theta\,n_z = 2\,r\frac{\partial\Phi_1}{\partial n} + 2\,\frac{\partial\Phi_0}{\partial n} + \frac{2}{r}\frac{\partial}{\partial s}(rz\,\Phi_1) + \Phi_1 n_r + 3\,\tilde\theta\,n_z.$$

Note that the integral contribution of the term $\frac{1}{r}\frac{\partial}{\partial s}(rz\,\Phi_1)$ to (71) is zero. Indeed,

$$\iint\limits_S \frac{1}{r}\frac{\partial}{\partial s}(rz\,\Phi_1)\,dS = 2\pi\int\limits_{-\infty}^{+\infty}\frac{\partial}{\partial\xi}(rz\,\Phi_1)\Big|_{\eta=\eta_0}d\xi = 2\pi\,(rz\,\Phi_1)\big|_{(\xi,\eta)=(-\infty,\eta_0)}^{(\xi,\eta)=(+\infty,\eta_0)} = 0.$$

We may avoid the use of Hilbert formulas for expressing $\tilde\theta$. Indeed, representing the generalized Cauchy–Riemann system (7) in terms of $\frac{\partial}{\partial s}$ and $\frac{\partial}{\partial n}$:

$$\frac{\partial\tilde\theta}{\partial n} = \frac{1}{r}\frac{\partial}{\partial s}(r\,\Phi_1), \qquad \frac{\partial\tilde\theta}{\partial s} = -\frac{1}{r}\frac{\partial}{\partial n}(r\,\Phi_1),$$

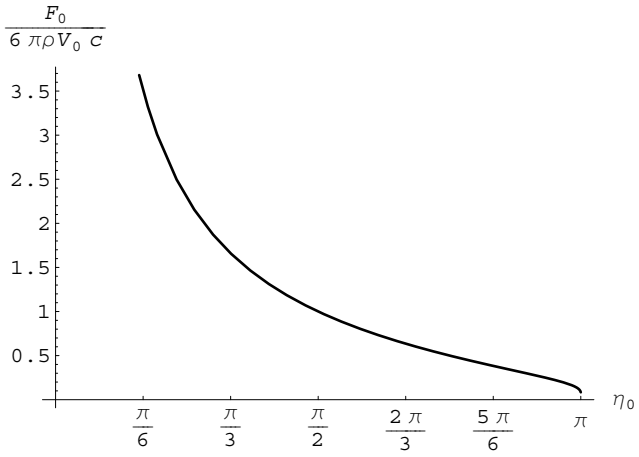


FIG. 7. *Normalized drag force,* $\frac{F_0}{6\pi\rho V_0 c}$*, as a function of* η_0.

we obtain

$$\tilde{\theta}\, n_z = -\frac{1}{2r}\frac{\partial}{\partial s}\left(r^2\tilde{\theta}\right) - \frac{1}{2}\,\frac{\partial}{\partial n}\left(r\,\Phi_1\right),$$

where the integral contribution of the term $\frac{1}{r}\frac{\partial}{\partial s}\left(r^2\tilde{\theta}\right)$ to (71) is zero:

$$\iint\limits_S \frac{1}{r}\frac{\partial}{\partial s}\left(r^2\tilde{\theta}\right) dS = 2\pi \int\limits_{-\infty}^{+\infty} \frac{\partial}{\partial \xi}\left(r^2\tilde{\theta}\right)\Big|_{\eta=\eta_0} d\xi = 2\pi\, \left(r^2\tilde{\theta}\right)\Big|_{(\xi,\eta)=(-\infty,\eta_0)}^{(\xi,\eta)=(+\infty,\eta_0)} = 0.$$

Thus, expression (71) reduces to

$$\frac{1}{2\rho}F_0 = -\frac{\pi}{2}\int\limits_{-\infty}^{+\infty}\left(r^2\,\frac{\partial\Phi_1}{\partial\eta} - \Phi_1\, r\frac{\partial r}{\partial\eta} + 4r\,\frac{\partial\Phi_0}{\partial\eta}\right)\Big|_{\eta=\eta_0} d\xi.$$

Finally, substituting representations (59) and (60) into the last expression and using relation (64), we obtain (70). □

Note that since the derivation of (70) does not use Hilbert formulas, formula (70) holds for all values of $\eta_0 \in (0,\pi)$. Figure 7 illustrates behavior of the normalized drag force $\frac{F_0}{6\pi\rho V_0 c}$ as a function of η_0. Table 3 presents values of $\frac{F_0}{6\pi\rho V_0 c}$ for $\eta_0 = \frac{\pi k}{12}$, $1 \le k \le 12$. In the case of $\eta_0 \to 0$, we have $F_0 \to \infty$.

The drag force may also be calculated as the limit of the stream function at $z = 0$ and $r \to \infty$

$$F_0 = -8\pi\rho \lim_{r\to\infty} \Psi|_{z=0}\,;$$

see [8]. For the stream function given by (61), this expression reduces to (70).

7. Concluding remarks. This paper addresses the problem of obtaining the Hilbert formulas for so-called r-analytic functions in the domain exterior to the contour of a spindle. In the meridional cross-section plane of the spindle, the real and

TABLE 3
Normalized drag force, $\frac{F_0}{6\pi\rho V_0 c}$, as a function of η_0.

η_0	$\frac{F_0}{6\pi\rho V_0 c}$	η_0	$\frac{F_0}{6\pi\rho V_0 c}$
$\pi/12$	7.113	$7\pi/12$	0.797
$2\pi/12$	3.510	$8\pi/12$	0.635
$3\pi/12$	2.287	$9\pi/12$	0.501
$4\pi/12$	1.660	$10\pi/12$	0.383
$5\pi/12$	1.271	$11\pi/12$	0.272
$6\pi/12$	1	$12\pi/12$	0

imaginary parts of the r-analytic function are represented by Fourier integrals with densities $X(\mu)$ and $Y(\mu)$, respectively, and the problem reduces to solving equation (16). To our knowledge, three approaches are available for obtaining the Hilbert formulas:

(1) Integrating system (7) in bipolar coordinates analytically [28],
(2) Solving (16) by complex Fourier transform [28, 29],
(3) Solving (16) in the framework of Riemann boundary-value problems for analytic functions.

In our previous work [28], we obtained Hilbert formulas by integrating the generalized Cauchy–Riemann system (7) under the condition that functions $\theta(\xi,\eta)$ and $\omega(\xi,\eta)$ had the same asymptotic behavior at $\xi \to \infty$. The disadvantage of this approach is that it is relatively complex in the sense of analytical computations and is based on special relations for Legendre functions. In our paper [29], we derived Hilbert formulas using modified Green's functions in the representations of θ and ω. The approach reduced the original problem to the equation similar to (16) and obtained the Hilbert formulas for the modified Green's functions by complex Fourier transform. Although we took into account the fact that the modified Green's functions might have simple poles $\mu = \pm\frac{1}{2}$ in the strip $|\operatorname{Re}\mu| \le 1$, this issue was not linked to specifying the class of functions for $X(\mu)$ and $Y(\mu)$, and as a result, the Hilbert formulas were stated with the accuracy of an additional term associated with homogeneous solutions to (16). In contrast to [29], this paper develops a new approach based on Riemann boundary-value problems to solving (16) for the class of meromorphic functions with two simple poles $\mu = \pm\frac{1}{2}$ in the strip $|\operatorname{Re}\mu| \le 1$ and having exponentially fast convergence at $\tau \to \infty$. The chosen class of meromorphic functions $X(\mu)$ and $Y(\mu)$ is determined by the hydrodynamic problem of the steady motion of a rigid-spindle shaped body in a viscous fluid: the function $A(\mu)$ in (65) has at least two simple poles at $\mu = \pm\frac{1}{2}$ for all values of the parameter η_0. In particular, in the framework of this approach, we show that the homogeneous equations $X(\mu+1) - 2X(\mu) + X(\mu-1) = 0$ and $\left(\mu + \frac{3}{2}\right) Y(\mu+1) - 2\mu\, Y(\mu) + \left(\mu - \frac{3}{2}\right) Y(\mu-1) = 0$ have only trivial solutions from the specified class of meromorphic functions. However, this may not be the case if we consider another class of meromorphic functions. Also, in this work, we derive formulas for efficient calculating Fourier integrals with the Hilbert formulas, i.e., double integrals one of which is singular.

The problem of axially symmetric Stokes flow about a rigid spindle-shaped body was originally considered by Pell and Payne [15]. However, they did not address the issue of determining and analyzing the pressure function. In our paper [30], we solved this hydrodynamic problem using the stream function in the form of (58) and presented the Hilbert formula for θ, obtained by integrating system (7) analytically. However, at the surface of the body, the pressure was calculated by integrating system

(7) numerically. Moreover, the Hilbert formula for the vortex function ω was not presented. This paper closes the gap: it presents both Hilbert formulas for θ and ω, respectively, and calculates isobars and the pressure at the surface using the special representations for the Fourier integrals with Hilbert formulas.

We should note that the class of meromorphic functions with only two simple poles at $\mu = \pm\frac{1}{2}$ in the strip $\mathrm{Re}\,|\mu| \leq 1$ precludes applying the Hilbert formulas for determining the pressure when $\eta_0 \geq 2.6037$, since in this case, the function $A(\mu)$ has other poles in the strip $\mathrm{Re}\,|\mu| \leq 1$; see Table 1. We may show that under some conditions, the Hilbert formulas (20) and (31) will be the same for a class of meromorphic functions $X(\mu)$ and $Y(\mu)$ with more than two simple poles in the strip $\mathrm{Re}\,|\mu| \leq 1$. Consequently, formulas (20) and (31) can be used for all $\eta_0 \in [0, \pi)$. However, addressing this issue calls for a separate publication.

Appendix A. The derivation of formulas (14) and (15). Formula (14) is derived as follows. Using the expression for the derivative $\frac{\partial}{\partial r}$ in (11) and the relation $\frac{\partial}{\partial r}\sqrt{\cosh\xi - \cos\eta} = -\frac{1}{2c}\cosh\xi\sin\eta\sqrt{\cosh\xi - \cos\eta}$, we have

$$\frac{\partial\theta}{\partial r} = \frac{1}{2\pi ic}\sqrt{\cosh\xi - \cos\eta}\left(-\frac{1}{2}\cosh\xi\sin\eta\int\limits_{-i\infty}^{+i\infty} X(\mu)\,\mathrm{P}_{-\frac{1}{2}+\mu}(\cos\eta)\mathrm{e}^{-\xi\mu}d\mu\right.$$
$$+\ \sinh\xi\sin\eta\int\limits_{-i\infty}^{+i\infty}\mu\,X(\mu)\,\mathrm{P}_{-\frac{1}{2}+\mu}(\cos\eta)\mathrm{e}^{-\xi\mu}d\mu$$
$$\left.+\ (\cosh\xi\cos\eta - 1)\int\limits_{-i\infty}^{+i\infty} X(\mu)\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta)\mathrm{e}^{-\xi\mu}d\mu\right).$$

Next, using the formulas

$$\sin\eta\,\mathrm{P}_{-\frac{1}{2}+\mu}(\cos\eta) = \frac{1}{2\mu}\left(\mathrm{P}^{(1)}_{-\frac{3}{2}+\mu}(\cos\eta) - \mathrm{P}^{(1)}_{\frac{1}{2}+\mu}(\cos\eta)\right),$$
$$\cos\eta\,\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta) = \frac{1}{2\mu}\left(\left(\mu+\tfrac{1}{2}\right)\mathrm{P}^{(1)}_{-\frac{3}{2}+\mu}(\cos\eta) + \left(\mu-\tfrac{1}{2}\right)\mathrm{P}^{(1)}_{\frac{1}{2}+\mu}(\cos\eta)\right),$$

we obtain

$$\frac{\partial\theta}{\partial r} = \frac{1}{4\pi ic}\sqrt{\cosh\xi - \cos\eta}\left(\cosh\xi\int\limits_{-i\infty}^{+i\infty} X(\mu)\left(\mathrm{P}^{(1)}_{-\frac{3}{2}+\mu}(\cos\eta) + \mathrm{P}^{(1)}_{\frac{1}{2}+\mu}(\cos\eta)\right)\mathrm{e}^{-\xi\mu}d\mu\right.$$
$$+\ \sinh\xi\int\limits_{-i\infty}^{+i\infty} X(\mu)\left(\mathrm{P}^{(1)}_{-\frac{3}{2}+\mu}(\cos\eta) - \mathrm{P}^{(1)}_{\frac{1}{2}+\mu}(\cos\eta)\right)\mathrm{e}^{-\xi\mu}d\mu$$
$$\left.-\ 2\int\limits_{-i\infty}^{+i\infty} X(\mu)\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta)\mathrm{e}^{-\xi\mu}d\mu\right),$$

which reduces to

$$\frac{\partial\theta}{\partial r} = \frac{1}{4\pi i c}\sqrt{\cosh\xi - \cos\eta}\left(\int_{-i\infty}^{+i\infty} X(\mu)\,\mathrm{P}^{(1)}_{-\frac{3}{2}+\mu}(\cos\eta)\,\mathrm{e}^{-\xi(\mu-1)}d\mu\right.$$
$$+\int_{-i\infty}^{+i\infty} X(\mu)\,\mathrm{P}^{(1)}_{\frac{1}{2}+\mu}(\cos\eta)\,\mathrm{e}^{-\xi(\mu+1)}d\mu \tag{72}$$
$$\left.-\,2\int_{-i\infty}^{+i\infty} X(\mu)\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta)\mathrm{e}^{-\xi\mu}d\mu\right).$$

Assuming $X(\mu)$ to be a meromorphic function with only two simple poles at $\mu = \pm\frac{1}{2}$ in the strip $|\operatorname{Re}\mu| \le 1$, and noticing that functions $\mathrm{P}^{(1)}_{-\frac{3}{2}+\mu}(\cos\eta)$ and $\mathrm{P}^{(1)}_{\frac{1}{2}+\mu}(\cos\eta)$ have nulls at $\mu = \frac{1}{2}$ and $\mu = -\frac{1}{2}$, respectively, we can write

$$\int_{-i\infty}^{+i\infty} X(\mu)\,\mathrm{P}^{(1)}_{-\frac{3}{2}+\mu}(\cos\eta)\,\mathrm{e}^{-\xi(\mu-1)}d\mu = \int_{1-i\infty}^{1+i\infty} X(\mu)\,\mathrm{P}^{(1)}_{-\frac{3}{2}+\mu}(\cos\eta)\,\mathrm{e}^{-\xi(\mu-1)}d\mu$$
$$= \int_{-i\infty}^{+i\infty} X(\mu+1)\,\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta)\,\mathrm{e}^{-\xi\mu}d\mu,$$

$$\int_{-i\infty}^{+i\infty} X(\mu)\,\mathrm{P}^{(1)}_{\frac{1}{2}+\mu}(\cos\eta)\,\mathrm{e}^{-\xi(\mu+1)}d\mu = \int_{-1-i\infty}^{-1+i\infty} X(\mu)\,\mathrm{P}^{(1)}_{\frac{1}{2}+\mu}(\cos\eta)\,\mathrm{e}^{-\xi(\mu+1)}d\mu$$
$$= \int_{-i\infty}^{+i\infty} X(\mu-1)\,\mathrm{P}^{(1)}_{-\frac{1}{2}+\mu}(\cos\eta)\,\mathrm{e}^{-\xi\mu}d\mu.$$

It should be noted that the two equalities above do not hold if the function $X(\mu)$ has other poles in the strip $|\operatorname{Re}\mu| \le 1$. Finally, substituting the last two equalities into expression (72), we obtain formula (14). Formula (15) is derived similarly.

Acknowledgments. We are grateful to the anonymous referees for their valuable comments and suggestions, which greatly helped to improve the quality of the paper.

REFERENCES

[1] H. Bateman and A. Erdelyi, *Higher Transcendental Functions,* vol. 1, McGraw–Hill, New York, 1953.

[2] L. Bers, *Theory of Pseudo-Analytic Functions*, Institute of Mathematics and Mechanics, New York University, New York, 1953.

[3] W. D. Collins, *A note on the axisymmetric Stokes flow of viscous fluid past a spherical cap*, Mathematika, 10 (1963), pp. 72–78.

[4] F. D. Gakhov, *Boundary Value Problems*, Pergamon Press, Oxford, New York, 1966.

[5] S. Ghosh, *On the steady motion of a viscous liquid due to translation of a torus parallel to its axis*, Bull. Calcutta Math. Soc., 18 (1927), pp. 185–194.

[6] O. G. Goman, *Representation in terms of p-analytic functions of the general solution of equations of the theory of elasticity of a transversely isotropic body*, J. Appl. Math. Mech., 48 (1984), pp. 62–67.

[7] S. L. Goren and M. E. O'Neill, *Asymmetric creeping motion of an open torus*, J. Fluid Mech., 101 (1980), pp. 97–110.
[8] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics*, Prentice–Hall, New York, 1965.
[9] R. Hill and G. Power, *Extremum principles for slow viscous flow and the approximate calculation of drag*, Quart. J. Mech. Appl. Math., 9 (1956), pp. 313–319.
[10] E. W. Hobson, *The Theory of Spherical and Ellipsoidal Harmonics*, Chelsea Publishing Co., New York, 1955.
[11] R. E. Johnson and T. Y. Wu, *Hydrodynamics of low-Reynolds-number flow. Part* 5. *Motion of a slender torus*, J. Fluid Mech., 95 (1979), pp. 263–277.
[12] V. V. Kravchenko, *On the relationship between p-analytic functions and the Schrödinger equation*, Z. Anal. Anwendungen, 24 (2005), pp. 487–496.
[13] L. E. Payne, *On axially symmetric flow and the method of generalized electrostatics*, Q. Appl. Math., 10 (1952), pp. 197–204.
[14] L. E. Payne and W. H. Pell, *The Stokes flow problem for a class of axially symmetric bodies*, J. Fluid Mech., 7 (1960), pp. 529–549.
[15] W. H. Pell and L. E. Payne, *The Stokes flow about a spindle*, Q. Appl. Math., 18 (1960), pp. 257–262.
[16] W. H. Pell and L. E. Payne, *On Stokes flow about a torus*, Mathematika, 7 (1960), pp. 78–92.
[17] G. N. Polozhii, *Theory and Application of p-Analytic and* (p, q)*-Analytic Functions*, 2nd ed., Naukova Dumka, Kiev, 1973 (in Russian).
[18] M. Stimson and G. G. Jeffery, *The motion of two-spheres in a viscous fluid*, Proc. Roy. Soc., London, 111 (1926), pp. 110–116.
[19] G. G. Stokes, *Mathematical and Physical Papers*, vol. I, University Press, Cambridge, UK, 1880.
[20] H. Takagi, *Slow viscous flow due to the motion of a closed torus*, J. Phys. Soc. Japan, 35 (1973), pp. 1225–1227.
[21] A. F. Ulitko, *Vectorial Decompositions in the Three-Dimensional Theory of Elasticity*, Akademperiodika, Kiev, 2002 (in Russian).
[22] I. N. Vekua, *Generalized Analytic Functions,* Pergamon Press, Oxford, 1962.
[23] V. S. Vladimirov, *Equations of Mathematical Physics*, Marcel Dekker, New York, 1971.
[24] S. Wakiya, *Axisymmetric Stokes flow about a body made of intersection of two spherical surfaces*, Archi. Mech. Stos., 32 (1980), pp. 809–817.
[25] S. Wakiya, *Axisymmetric flow of a viscous fluid near the vertex of a body*, Fluid Mech., 78 (1976), pp. 737–747.
[26] S. Wakiya, *On the exact solution of the Stokes equations for a torus*, J. Phys. Soc. Japan, 37 (1974), pp. 780–783.
[27] S. Wakiya, *Slow motion of a viscous fluid around two spheres*, J. Phys. Soc. Japan, 22 (1967), pp. 1101–1109.
[28] M. Zabarankin, *Exact Solutions to Displacement Boundary-Value Problems for an Elastic Medium with a Spindle-Shaped Inclusion*, Ph.D. Thesis, National Taras Shevchenko University of Kiev, Kiev, 1999 (in Russian).
[29] M. Zabarankin, *General approach to solving the generalized Cauchy–Riemann system*, Rep. National Acad. Sci. Ukraine, 5 (1999), pp. 30–33 (in Russian).
[30] M. Zabarankin and A. F. Ulitko, *The Stokes flow about a spindle in axisymmetric case*, Bull. National Taras Shevchenko Univ. Kiev Math. Mech., 3 (1999), pp. 58–66 (in Ukrainian).

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1301–1320

MULTISTABILITY IN RECURRENT NEURAL NETWORKS*

CHANG-YUAN CHENG†, KUANG-HUI LIN†, AND CHIH-WEN SHIH‡

Abstract. Stable stationary solutions correspond to memory capacity in the application of associative memory for neural networks. In this presentation, existence of multiple stable stationary solutions for Hopfield-type neural networks with delay and without delay is investigated. Basins of attraction for these stationary solutions are also estimated. Such a scenario of dynamics is established through formulating parameter conditions based on a geometrical setting. The present theory is demonstrated by two numerical simulations on the Hopfield neural networks with delays.

Key words. neural network, multistability, delay equations

AMS subject classifications. 34D20, 34D45, 92B20

DOI. 10.1137/050632440

1. Introduction.

The studies of neural networks have attracted considerable multidisciplinary research interest in recent years. The developments for neural network models and the theory for the models are, on the one hand, driven by application motif or inspired by biological neuronal behaviors. On the other hand, the neural network theory has motivated and elicited further progress in dynamical system theory. For example, theory for existence of many stable patterns or chaotic dynamics for systems in phase space of large dimension is in strong demand for neural network applications. The progress in this direction of research has also enriched dynamical system theory [6, 17, 27].

The applications of neural networks range from classifications, associative memory, image processing, and pattern recognition to parallel computation and its ability to solve optimization problems. The theory on the dynamics of the networks has been developed according to the purposes of the applications. In the application to parallel computation and signal processing involving finding the solution of an optimization problem, the existence of a computable solution for all possible initial states is the best situation. Mathematically, this means that the network needs to have a unique equilibrium which is globally attractive. Such a convergent behavior is referred to as "monostability" of a network. On the other hand, when a neural network is employed as an associative memory storage or for pattern recognition, the existence of many equilibria is a necessary feature [7, 11, 16, 21]. The notion of "multistability" of a neural network is used to describe coexistence of multiple stable patterns such as equilibria or periodic orbits. In general, if the dynamics for a system are bounded, the existence of multiple stable patterns is accompanied with coexistence of stable and unstable equilibria or periodic orbits. The existence of unstable equilibria is essential in certain applications of neural network. For example, unstable equilibria are related to digital constraints on selection in winner-take-all problems [32, 33].

*Received by the editors May 25, 2005; accepted for publication (in revised form) November 17, 2005; published electronically March 31, 2006. This work was partially supported by The National Science Council and The National Center of Theoretical Sciences of Republic of China on Taiwan.
http://www.siam.org/journals/siap/66-4/63244.html
†Department of Applied Mathematics, National Chiao Tung University, Hsinchu, Taiwan, Republic of China (chengcy13@yahoo.com.tw, hs3893@mail.nc.hcc.edu.tw).
‡Corresponding author. Department of Applied Mathematics, National Chiao Tung University, Hsinchu, Taiwan, Republic of China (cwshih@math.nctu.edu.tw).

Classical recurrent neural networks are usually systems of ordinary differential equations. Recently, neural network systems with delays have also been studied extensively, thanks to the need from practical applications and mathematical interests. In this presentation, we propose an approach to investigate existence of multiple stationary solutions and their stability for recurrent neural networks with delay and without delay. We shall illustrate our approach through the Hopfield-type model.

Hopfield-type neural networks and their various generalizations have been widely studied and applied in various scientific areas. A typical form for such a network is given by

$$C_i \frac{dx_i(t)}{dt} = -\frac{x_i(t)}{R_i} + \sum_{j=1}^{n} T_{ij} g_j(x_j(t-\tau_{ij})) + I_i, \quad i = 1, 2, \ldots, n, \tag{1.1}$$

where $C_i > 0$ and $R_i > 0$ are, respectively, the input capacitance and resistance associated with neuron i; I_i is the constant input; T_{ij} are the connection strengths between neurons; $\tau_{ij} > 0$ are the transmission delays; and $g_i, i = 1, 2, \ldots, n$, are neuron activation functions.

The classical Hopfield-type neural network [16] is system (1.1) without delay, that is, $\tau_{ij} = 0$ for all i, j. For the Hopfield-type neural networks, the theory of unique equilibrium and global convergence to the equilibrium has been extensively studied; cf. [9, 10] for the networks without delays and [5, 13, 19, 23, 24, 29, 30, 31, 34, 35] for the delay cases.

In contrast to these studies, we propose a treatment to explore the existence of multiple stationary solutions for (1.1) through a geometrical formulation on the parameter conditions. Stability of these equilibria for (1.1) with and without delay shall also be investigated. In addition, estimations of basins of attraction for these stable stationary solutions are derived. The stationary equations are identical for system (1.1) with delay and without delay. Thus, confirmation for the existence of equilibrium points is valid for both cases. However, stability of the equilibrium points and dynamical behaviors can be very different for the systems with delay and without delay. It is very interesting to explore such a difference as well as a possible coincidence of behaviors.

The theory for existence of multiple stable patterns has been developed for cellular neural networks [8, 17, 26, 27]. The neurons in such a system are locally connected and no time lags were considered therein. Our approach can be adopted to such a network with delays, as remarked in the later section. There are other interesting studies on delayed neural networks in [1, 2, 12, 22, 25].

This presentation is organized as follows. In section 2, we establish conditions for existence of 3^n equilibria for the Hopfield network. 2^n equilibria among them will be shown to be asymptotically stable for the system without delays, through a linearization analysis. In section 3, we shall verify that under the same conditions, there are 2^n regions in $\mathbb{R}^n$, each containing an equilibrium, which are positively invariant under the flow generated by the system with delays and without delays. Subsequently, it is argued that these 2^n equilibria are asymptotically stable, even in the presence of delays. We also formulate more sufficient conditions for stability of these 2^n equilibria. We extend our theory to more general activation functions, including those with saturations, in section 4. Two numerical simulations on the dynamics of two-neuron networks, which illustrate the present theory, are given in section 5. We summarize our results with a discussion (section 6).

2. Existence of multiple equilibria and their stability. In this section, we shall formulate sufficient conditions for the existence of multiple stationary solutions for Hopfield neural networks with and without delays. Our approach is based on a geometrical observation. The derived parameter conditions are concrete and can be examined easily. We also establish stability criteria of these equilibria for the system without delays, through estimations on the eigenvalues of the linearized system. Stability for the system with delays will be discussed in the next section. After rearranging the parameters, we consider system (1.1) in the following forms: for the network without delay,

$$\frac{dx_i(t)}{dt} = -b_i x_i(t) + \sum_{j=1}^{n} \omega_{ij} g_j(x_j(t)) + J_i, \quad i = 1, 2, \ldots, n, \tag{2.1}$$

and for the network with delays,

$$\frac{dx_i(t)}{dt} = -b_i x_i(t) + \sum_{j=1}^{n} \omega_{ij} g_j(x_j(t - \tau_{ij})) + J_i, \quad i = 1, 2, \ldots, n. \tag{2.2}$$

Herein, $b_i > 0$, $0 < \tau_{ij} \le \tau := \max_{1 \le i,j \le n} \tau_{ij}$. While (2.1) is a system of ordinary differential equations, (2.2) is a system of functional differential equations. The initial condition for (2.2) is

$$x_i(\theta) = \phi_i(\theta), \quad -\tau \le \theta \le 0, \quad i = 1, 2, \ldots, n,$$

and it is usually assumed that $\phi_i \in \mathcal{C}([-\tau, 0], \mathbb{R})$. Let $\ell > 0$. For $\mathbf{x} \in \mathcal{C}([-\tau, \ell], \mathbb{R}^n)$ and $t \in [0, \ell]$, we define

$$\mathbf{x}_t(\theta) = \mathbf{x}(t + \theta), \quad \theta \in [-\tau, 0]. \tag{2.3}$$

Let us denote $\tilde{F} = (\tilde{F}_1, \ldots, \tilde{F}_n)$, where $\tilde{F}_i$ is the right-hand side of (2.2),

$$\tilde{F}_i(\mathbf{x}_t) := -b_i x_i(t) + \sum_{j=1}^{n} \omega_{ij} g_j(x_j(t - \tau_{ij})) + J_i,$$

where $\mathbf{x} = (x_1, \ldots, x_n)$. A function $\mathbf{x} = \mathbf{x}(t)$ is called a solution of (2.2) on $[-\tau, \ell)$ if $\mathbf{x} \in \mathcal{C}([-\tau, \ell), \mathbb{R}^n)$ and $\mathbf{x}_t$ defined as (2.3) lies in the domain of $\tilde{F}$ and satisfies (2.2) for $t \in [0, \ell)$. For a given $\phi \in \mathcal{C}([-\tau, 0], \mathbb{R}^n)$, let us denote by $\mathbf{x}(t; \phi)$ the solution of (2.2) with $\mathbf{x}_0(\theta; \phi) := \mathbf{x}(0 + \theta; \phi) = \phi(\theta)$ for $\theta \in [-\tau, 0]$.

The activation functions g_j usually have sigmoidal configuration or are nondecreasing with saturations. Herein, we consider the typical logistic or Fermi function: for all $j = 1, 2, \ldots, n$,

$$g_j(\xi) = g(\xi) := \frac{1}{1 + e^{-\xi/\varepsilon}}, \quad \varepsilon > 0. \tag{2.4}$$

One may also adopt $g_j(\xi) = 1/(1 + e^{-\xi/\varepsilon_j})$, $\varepsilon_j > 0$, or other output functions, as discussed in section 4. Note that the stationary equations for systems (2.1) and (2.2) are identical; namely,

$$F_i(\mathbf{x}) := -b_i x_i + \sum_{j=1}^{n} \omega_{ij} g_j(x_j) + J_i = 0, \quad i = 1, 2, \ldots, n, \tag{2.5}$$

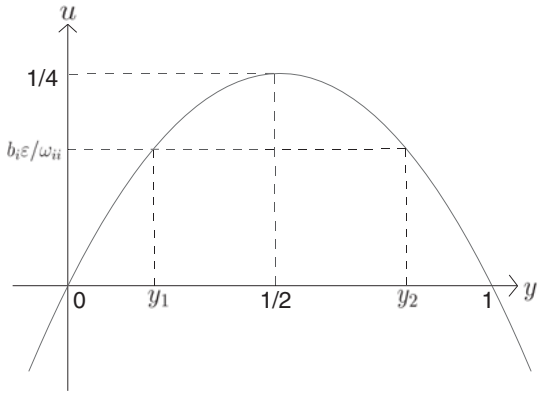


FIG. 1. *The graph for function* $u(y) = y - y^2$ *and* $y_1 = g(p_i)$, $y_2 = g(q_i)$.

where $\mathbf{x} = (x_1, \ldots, x_n)$. For our formulation in the following discussions, we introduce a single neuron analogue (no interaction among neurons),

$$\frac{d\xi}{dt} = f_i(\xi) := -b_i\xi + \omega_{ii}g(\xi) + J_i, \quad \xi \in \mathbb{R}.$$

Let us propose the first parameter condition:

$$(\mathrm{H}_1)\colon 0 < \frac{b_i\varepsilon}{\omega_{ii}} < \frac{1}{4}, \quad i = 1, 2, \ldots, n.$$

LEMMA 2.1. *Under condition* (H_1)*, there exist two points* p_i *and* q_i *with* $p_i < 0 < q_i$ *such that* $f_i'(p_i) = 0$, $f_i'(q_i) = 0$ *for* $i = 1, 2, \ldots, n$.

Proof. We compute that

$$g'(\xi) = \frac{1}{\varepsilon}(1 + e^{-\xi/\varepsilon})^{-2}e^{-\xi/\varepsilon}. \tag{2.6}$$

Note that g is strictly increasing and that the graph of function $g'(\xi)$ is concave down and has its maximal value at $\xi = 0$. We let $y = g(\xi)$, $\xi \in \mathbb{R}$. Then $y \in (0, 1)$ and $g(0) = 1/2$. It follows from (2.6) that

$$g'(\xi) = \frac{1}{\varepsilon}y^2\left(\frac{1}{y} - 1\right) = \frac{1}{\varepsilon}(y - y^2).$$

On the other hand, for each i, since $f_i'(\xi) = -b_i + \omega_{ii}g'(\xi)$, we have $f_i'(\xi) = 0$ if and only if $b_i = \omega_{ii}g'(\xi)$; equivalently,

$$\frac{b_i\varepsilon}{\omega_{ii}} = y - y^2.$$

From the configuration in Figure 1, it follows that, for each i, there exist two points p_i, q_i, $p_i < 0 < q_i$, such that $f_i'(p_i) = f_i'(q_i) = 0$ if the parameter condition $0 < b_i\varepsilon/\omega_{ii} < 1/4$ holds. This completes the proof. ∎

Note that condition (H_1) implies $\omega_{ii} > 0$ for all $i = 1, 2, \ldots, n$, since each b_i is already assumed to be a positive constant. We define, for $i = 1, 2, \ldots, n$,

$$\hat{f}_i(\xi) = -b_i\xi + \omega_{ii}g(\xi) + k_i^+,$$
$$\check{f}_i(\xi) = -b_i\xi + \omega_{ii}g(\xi) + k_i^-,$$

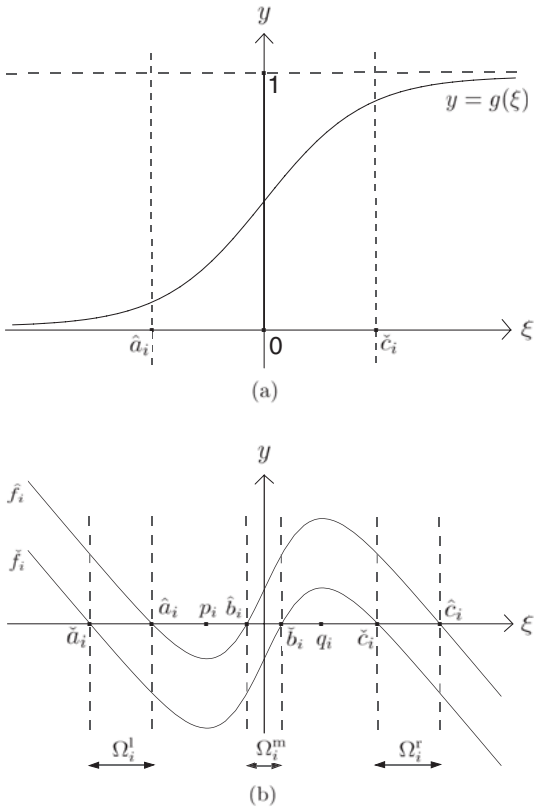


FIG. 2. (a) *The graph of g with* $\varepsilon = 0.5$*;* (b) *Configurations for* $\hat{f}_i$ *and* $\check{f}_i$.

where

$$k_i^+ := \sum_{j=1, j\neq i}^{n} |\omega_{ij}| + J_i, \;\; k_i^- := -\sum_{j=1, j\neq i}^{n} |\omega_{ij}| + J_i.$$

It follows that

$$\check{f}_i(x_i) \leq F_i(\mathbf{x}) \leq \hat{f}_i(x_i) \tag{2.7}$$

for all $\mathbf{x} = (x_1, \ldots, x_n)$ and $i = 1, 2, \ldots, n$, since $0 \leq g_j \leq 1$ for all j.

We consider the second parameter condition which is concerned with the existence of multiple equilibria for (2.1) and (2.2):

$$(\mathrm{H}_2)\text{: } \hat{f}_i(p_i) < 0, \quad \check{f}_i(q_i) > 0, \quad i = 1, 2, \ldots, n.$$

The configuration that motivates (H_2) is depicted in Figure 2. Such a configuration is due to the characteristics of the output function g. Under assumptions (H_1) and (H_2), there exist points $\hat{a}_i, \hat{b}_i, \hat{c}_i$ with $\hat{a}_i < \hat{b}_i < \hat{c}_i$ such that $\hat{f}_i(\hat{a}_i) = \hat{f}_i(\hat{b}_i) = \hat{f}_i(\hat{c}_i) = 0$ as well as points $\check{a}_i, \check{b}_i, \check{c}_i$ with $\check{a}_i < \check{b}_i < \check{c}_i$ such that $\check{f}_i(\check{a}_i) = \check{f}_i(\check{b}_i) = \check{f}_i(\check{c}_i) = 0$.

THEOREM 2.2. *Under* (H_1) *and* (H_2)*, there exist* 3^n *equilibria for systems* (2.1) *and* (2.2).

Proof. The equilibria of systems (2.1) and (2.2) are zeros of (2.5). Under conditions (H_1) and (H_2), the graphs of $\hat{f}_i$ and $\check{f}_i$ defined above are as depicted in Figure

2. According to the configurations, there are 3^n disjoint closed regions in $\mathbb{R}^n$. Set $\Omega^\alpha = \{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \mid x_i \in \Omega_i^{\alpha_i}\}$ with $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$, and $\alpha_i =$ "l," "m," or "r," where

$$\begin{aligned} &\Omega_i^{\mathrm{l}} := \{x \in \mathbb{R} |\ \check{a}_i \le x \le \hat{a}_i\}, \quad \Omega_i^{\mathrm{m}} := \{x \in \mathbb{R} |\ \hat{b}_i \le x \le \check{b}_i\}, \\ &\Omega_i^{\mathrm{r}} := \{x \in \mathbb{R} |\ \check{c}_i \le x \le \hat{c}_i\}. \end{aligned} \tag{2.8}$$

Herein, "l," "m," and "r" mean, respectively, "left," "middle," and "right." Consider any fixed one of these regions Ω^α. For a given $\tilde{\mathbf{x}} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n) \in \Omega^\alpha$, we solve

$$h_i(x_i) := -b_i x_i + \omega_{ii} g(x_i) + \sum_{j=1, j\neq i}^{n} \omega_{ij} g(\tilde{x}_j) + J_i = 0$$

for x_i, $i = 1, 2, \ldots, n$. According to an estimate similar to (2.7), the graph of h_i lies between the graphs of $\hat{f}_i$ and $\check{f}_i$. In fact, the graph of h_i is a vertical shift of the graph of $\hat{f}_i$ or $\check{f}_i$. Thus, one can always find three solutions, and each of them lies in one of the regions in (2.8) for each i. Let us pick the one lying in $\Omega_i^{\alpha_i}$ and set it as $\underline{x}_i$ for each i. We define a mapping $\mathbf{H}_\alpha : \Omega^\alpha \to \Omega^\alpha$ by $\mathbf{H}_\alpha(\tilde{\mathbf{x}}) = \underline{\mathbf{x}} = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)$. Restated, we set

$$\begin{aligned} \underline{x}_i &= (h_i|_{\Omega_i^{\mathrm{l}}})^{-1}(0) \text{ if } \alpha_i = \text{"l,"} \\ \underline{x}_i &= (h_i|_{\Omega_i^{\mathrm{m}}})^{-1}(0) \text{ if } \alpha_i = \text{"m,"} \\ \underline{x}_i &= (h_i|_{\Omega_i^{\mathrm{r}}})^{-1}(0) \text{ if } \alpha_i = \text{"r."} \end{aligned}$$

Since g is continuous and h_i is a vertical shift of function $\xi \mapsto -b_i\xi + \omega_{ii} g(\xi)$ by the quantity $\sum_{j=1, j\neq i}^{n} \omega_{ij} g(\tilde{x}_j) + J_i$, the map $\mathbf{H}_\alpha$ is continuous. It follows from Brouwer's fixed point theorem that there exists one fixed point $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$ of $\mathbf{H}_\alpha$ in Ω^α which is also a zero of the function F, where $F = (F_1, F_2, \ldots, F_n)$. Consequently, there exist 3^n zeros of F, hence 3^n equilibria for systems (2.1) and (2.2), and each of them lies in one of the 3^n regions Ω^α. This completes the proof. □

We consider the following criterion concerning stability of the equilibria:

(2.9)

$$(\mathrm{H}_3)\colon -b_i + \sum_{j=1}^{n} |\omega_{ij}| g'(\eta_j) < 0, \quad g'(\eta_j) := \max\{g'(x_j) \mid x_j = \check{c}_j, \hat{a}_j\}, \quad i = 1, 2, \ldots, n.$$

A simplified yet more restrictive version for condition (H_3) is that for $i = 1, 2, \ldots, n$,

$$b_i > g'(\eta) \sum_{j=1}^{n} |\omega_{ij}| \text{ with } g'(\eta) := \max\{g'(x_j) \mid x_j = \check{c}_j, \hat{a}_j, j = 1, 2, \ldots, n\}. \tag{2.10}$$

THEOREM 2.3. *Under conditions* (H_1), (H_2), *and* (H_3), *there exist* 2^n *asymptotically stable equilibria for the Hopfield neural networks without delay* (2.1).

Proof. Among the 3^n equilibria in Theorem 2.2, we consider those $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_n)$ with $\bar{x}_i \in \Omega_i^{\mathrm{l}}$ or Ω_i^{r} for each i. The linearized system of (2.1) at equilibrium $\bar{\mathbf{x}}$ is

$$\frac{dy_i}{dt} = -b_i y_i + \sum_{j=1}^{n} \omega_{ij} g_j'(\overline{x}_j) y_j, \quad i = 1, 2, \ldots, n.$$

Restated, $\dot{\mathbf{y}} = A\mathbf{y}$, where $DF(\bar{\mathbf{x}}) =: A = [a_{ij}]_{n\times n}$ with

$$[a_{ij}] = \begin{pmatrix} -b_1 + \omega_{11}g'(\bar{x}_1) & \omega_{12}g'(\bar{x}_2) & \cdots & \omega_{1n}g'(\bar{x}_n) \\ \omega_{21}g'(\bar{x}_1) & -b_2 + \omega_{22}g'(\bar{x}_2) & & \omega_{2n}g'(\bar{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{n1}g'(\bar{x}_1) & \omega_{n2}g'(\bar{x}_2) & \cdots & -b_n + \omega_{nn}g'(\bar{x}_n) \end{pmatrix}.$$

Let

$$r_i = \sum_{j=1,j\neq i}^{n} |a_{ij}| = \sum_{j=1,j\neq i}^{n} |\omega_{ij}g'(\bar{x}_j)| = \sum_{j=1,j\neq i}^{n} |\omega_{ij}|g'(\bar{x}_j), \quad i = 1, 2, \ldots, n.$$

According to Gerschgorin's theorem,

$$\lambda_k \in \bigcup_{i=1}^{n} B(a_{ii}, r_i)$$

for all $k = 1, 2, \ldots, n$, where λ_k are the eigenvalues of A and $B(a_{ii}, r_i) := \{\zeta \in \mathbb{C} \mid |\zeta - a_{ii}| < r_i\}$. Hence, for each k, there exists some $i = i(k)$ such that

$$\mathrm{Re}(\lambda_k) < -b_i + \omega_{ii}g'(\bar{x}_i) + \sum_{j=1,j\neq i}^{n} |\omega_{ij}|g'(\bar{x}_j).$$

Notice that for each j, $g'(\xi) \le g'(\check{c}_j)$ (resp., $g'(\xi) \le g'(\hat{a}_j)$) if $\xi \ge \check{c}_j$ (resp., $\xi \le \hat{a}_j$). Since $\bar{\mathbf{x}}$ is such that $\bar{x}_j \in \Omega_j^{\mathrm{l}}$ or Ω_j^{r}, we have $\bar{x}_j \ge \check{c}_j$ or $\bar{x}_j \le \hat{a}_j$ for all $j = 1, 2, \ldots, n$. It follows that $\mathrm{Re}(\lambda_k) < 0$ by (2.9). Thus, under (H_3), all the eigenvalues of A have negative real parts. Therefore, there are 2^n asymptotically stable equilibria for system (2.1). The proof is completed. □

We certainly can replace condition (H_3) by weaker ones, such as an individual condition for each equilibrium. Let $\bar{\mathbf{x}}$ be an equilibrium lying in Ω^α with $\alpha = (\alpha_1, \ldots, \alpha_n)$ and $\alpha_i =$ "r" or $\alpha_i =$ "l," that is, $\bar{x}_i \in \Omega_i^{\mathrm{l}}$ or Ω_i^{r}, for each i. For such an equilibrium we consider, for $i = 1, 2, \ldots, n$,

$$b_i > \omega_{ii}g'(\xi_i) + \sum_{j=1,j\neq i}^{n} |\omega_{ij}|g'(\xi_j), \quad \xi_k = \check{c}_k \text{ if } \alpha_k = \text{"r,"} \quad \xi_k = \hat{a}_k \text{ if } \alpha_k = \text{"l,"} \quad k = 1, \ldots, n.$$

Such conditions are obviously much more tedious than (H_3).

3. Stability of equilibria and the basins of attraction. We plan to investigate the stability of equilibrium for system (2.2), that is, with delays. We shall also explore the basins of attraction for the asymptotically stable equilibria, for both systems (2.1) and (2.2), in this section.

Note that the function $\xi \mapsto [\omega_{ii} + \sum_{j=1,j\neq i}^{n} |\omega_{ij}|]g'(\xi)$ is continuous for all $i = 1, 2, \ldots, n$. From (2.9) and $\omega_{ii} > 0$, it follows that there exists a positive constant ϵ_0 such that

$$b_i > \max\left\{ \left[\omega_{ii} + \sum_{j=1,j\neq i}^{n} |\omega_{ij}| \right] g'(\xi) : \xi = \hat{a}_i + \epsilon_0, \check{c}_i - \epsilon_0 \right\}, \ i = 1, 2, \ldots, n. \tag{3.1}$$

Herein, we choose ϵ_0 such that $\epsilon_0 < \min\{|\hat{a}_i - p_i|, |\check{c}_i - q_i|\}$ for all $i = 1, 2, \ldots, n$. For system (2.1), we consider the following 2^n subsets of $\mathbb{R}^n$. Let $\alpha = (\alpha_1, \ldots, \alpha_n)$ with $\alpha_i =$ "l" or "r," and set

$$\tilde{\Omega}^\alpha = \{(x_1, x_2, \ldots, x_n) \mid x_i \in \tilde{\Omega}_i^{\mathrm{l}} \text{ if } \alpha_i = \text{"l,"}\ x_i \in \tilde{\Omega}_i^{\mathrm{r}} \text{ if } \alpha_i = \text{"r"}\}, \tag{3.2}$$

where $\tilde{\Omega}_i^{\mathrm{l}} := \{\xi \in \mathbb{R} \mid \xi \le \hat{a}_i + \epsilon_0\}$, $\tilde{\Omega}_i^{\mathrm{r}} := \{\xi \in \mathbb{R} \mid \xi \ge \check{c}_i - \epsilon_0\}$. For system (2.2), we consider the following 2^n subsets of $\mathcal{C}([-\tau, 0], \mathbb{R}^n)$. Let $\alpha = (\alpha_1, \ldots, \alpha_n)$ with $\alpha_i =$ "l" or "r," and set

$$\Lambda^\alpha = \{\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_n) \mid \varphi_i \in \Lambda_i^{\mathrm{l}} \text{ if } \alpha_i = \text{"l,"}\ \varphi_i \in \Lambda_i^{\mathrm{r}} \text{ if } \alpha_i = \text{"r"}\}, \tag{3.3}$$

where

$$\Lambda_i^{\mathrm{l}} := \{\varphi_i \in \mathcal{C}([-\tau, 0], \mathbb{R}) \mid \varphi_i(\theta) \le \hat{a}_i + \epsilon_0 \text{ for all } \theta \in [-\tau, 0]\},$$
$$\Lambda_i^{\mathrm{r}} := \{\varphi_i \in \mathcal{C}([-\tau, 0], \mathbb{R}) \mid \varphi_i(\theta) \ge \check{c}_i - \epsilon_0 \text{ for all } \theta \in [-\tau, 0]\}.$$

THEOREM 3.1. *Assume that* (H_1) *and* (H_2) *hold. Then each* $\tilde{\Omega}^\alpha$ *and each* Λ^α *are positively invariant with respect to the solution flow generated by systems* (2.1) *and* (2.2)*, respectively.*

Proof. We prove only the delay case, i.e., system (2.2). Consider any one of the 2^n sets Λ^α. For any initial condition $\phi = (\phi_1, \phi_2, \ldots, \phi_n) \in \Lambda^\alpha$, we claim that the solution $\mathbf{x}(t; \phi)$ remains in Λ^α for all $t \ge 0$. If this is not true, there exists a component $x_i(t)$ of $\mathbf{x}(t; \phi)$ which is the first (or one of the first) escaping from Λ_i^{l} or Λ_i^{r}. Restated, there exist some i and $t_1 > 0$ such that either $x_i(t_1) = \check{c}_i - \epsilon_0$, $\frac{dx_i}{dt}(t_1) \le 0$, and $x_i(t) \ge \check{c}_i - \epsilon_0$ for $-\tau \le t \le t_1$ or $x_i(t_1) = \hat{a}_i + \epsilon_0$, $\frac{dx_i}{dt}(t_1) \ge 0$, and $x_i(t) \le \hat{a}_i + \epsilon_0$ for $-\tau \le t \le t_1$. For the first case $x_i(t_1) = \check{c}_i - \epsilon_0$ and $\frac{dx_i}{dt}(t_1) \le 0$, we derive from (2.2) that

$$\frac{dx_i}{dt}(t_1) = -b_i(\check{c}_i - \epsilon_0) + \omega_{ii} g(x_i(t_1 - \tau_{ii})) + \sum_{j=1, j\ne i}^{n} \omega_{ij} g(x_j(t_1 - \tau_{ij})) + J_i \le 0. \tag{3.4}$$

On the other hand, recalling (H_2) and previous descriptions of $\check{c}_i$ and ϵ_0, we have $\check{f}_i(\check{c}_i - \epsilon_0) > 0$ which gives

$$\begin{aligned} &-b_i(\check{c}_i - \epsilon_0) + \omega_{ii} g(\check{c}_i - \epsilon_0) + k_i^- \\ &= -b_i(\check{c}_i - \epsilon_0) + \omega_{ii} g(\check{c}_i - \epsilon_0) - \sum_{j=1, j\ne i}^{n} |\omega_{ij}| + J_i > 0. \end{aligned} \tag{3.5}$$

Notice that t_1 is the first time for x_i to escape from Λ_i^{r}. We have $g(x_i(t_1 - \tau_{ii})) \ge g(\check{c}_i - \epsilon_0)$, by the monotonicity of function g. In addition, by $\omega_{ii} > 0$ and $|g(\cdot)| \le 1$, we obtain from (3.5) that

$$\begin{aligned} &-b_i(\check{c}_i - \epsilon_0) + \omega_{ii} g(x_i(t_1 - \tau_{ii})) + \sum_{j=1, j\ne i}^{n} \omega_{ij} g(x_j(t_1 - \tau_{ij})) + J_i \\ &\ge -b_i(\check{c}_i - \epsilon_0) + \omega_{ii} g(\check{c}_i - \epsilon_0) - \sum_{j=1, j\ne i}^{n} |\omega_{ij}| + J_i > 0, \end{aligned}$$

which contradicts (3.4). Hence, $x_i(t) \geq \check{c}_i - \epsilon_0$ for all $t > 0$. Similar arguments can be employed to show that $x_i(t) \leq \hat{a}_i + \epsilon_0$ for all $t > 0$ for the situation that $x_i(t_1) = \hat{a}_i + \epsilon_0$ and $\frac{dx_i}{dt}(t_1) \geq 0$. Therefore, Λ^α is positively invariant under the flow generated by system (2.2). The assertion for system (2.1) can be justified similarly. □

THEOREM 3.2. *Under conditions* (H_1), (H_2), *and* (H_3), *there exist* 2^n *exponentially stable equilibria for system* (2.2).

Proof. Consider an equilibrium $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n) \in \Omega^\alpha$ for some $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$, with $\alpha_i =$ "l" or "r," obtained in Theorem 2.2. We consider the single-variable functions $G_i(\cdot)$, defined by

$$G_i(\zeta) = b_i - \zeta - \sum_{j=1}^{n} |\omega_{ij}| g'(\xi_j) e^{\zeta \tau_{ij}},$$

where $\xi_j = \hat{a}_j + \epsilon_0$ (resp., $\check{c}_j - \epsilon_0$) if $\alpha_j =$ "l" (resp., "r"). Then, $G_i(0) > 0$ from (3.1) or (H_3). Moreover, there exists a constant $\mu > 0$ such that $G_i(\mu) > 0$ for $i = 1, 2, \ldots, n$, due to continuity of G_i. Let $\mathbf{x}(t) = \mathbf{x}(t; \phi)$ be the solution to (2.2) with initial condition $\phi \in \Lambda^\alpha$ defined in (3.3). Under the translation $\mathbf{y}(t) = \mathbf{x}(t) - \bar{\mathbf{x}}$, system (2.2) becomes

$$\frac{dy_i(t)}{dt} = -b_i y_i(t) + \sum_{j=1}^{n} \omega_{ij} [g(x_j(t - \tau_{ij})) - g(\bar{x}_j)], \tag{3.6}$$

where $\mathbf{y} = (y_1, \ldots, y_n)$. Now, consider functions $z_i(\cdot)$ defined by

$$z_i(t) = e^{\mu t} |y_i(t)|, \quad i = 1, 2, \ldots, n. \tag{3.7}$$

The domain of definition for $z_i(\cdot)$ is identical to the interval of existence for $y_i(\cdot)$. We shall see in the following computations that the domain can be extended to $[-\tau, \infty)$. Let $\delta > 1$ be an arbitrary real number and let

$$K := \max_{1 \leq i \leq n} \left\{ \sup_{\theta \in [-\tau, 0]} |x_i(\theta) - \bar{x}_i| \right\} > 0. \tag{3.8}$$

It follows from (3.7) and (3.8) that $z_i(t) < K\delta$ for $t \in [-\tau, 0]$ and all $i = 1, 2, \ldots, n$. Next, we claim that

$$z_i(t) < K\delta \quad \text{for all } t > 0,\ i = 1, 2, \ldots, n. \tag{3.9}$$

Suppose this is not the case. Then there are an $i \in \{1, 2, \ldots, n\}$ (say $i = k$) and a $t_1 > 0$ for the first time such that

$$\begin{aligned} z_i(t) &\leq K\delta, \quad t \in [-\tau, t_1], \quad i = 1, 2, \ldots, n,\ i \neq k, \\ z_k(t) &< K\delta, \quad t \in [-\tau, t_1), \\ z_k(t_1) &= K\delta \quad \text{with } \frac{d}{dt} z_k(t_1) \geq 0. \end{aligned}$$

Note that $z_k(t_1) = K\delta > 0$ implies $y_k(t_1) \neq 0$. Hence $|y_k(t)|$ and $z_k(t)$ are differentiable at $t = t_1$. From (3.6), we derive that

$$\frac{d}{dt} |y_k(t_1)| \leq -b_k |y_k(t_1)| + \sum_{j=1}^{n} |\omega_{kj}| g'(\varsigma_j) |y_j(t_1 - \tau_{kj})| \tag{3.10}$$

for some ς_j between $x_j(t_1-\tau_{kj})$ and $\bar{x}_j$. Hence, from (3.7) and (3.10),

$$\begin{aligned}\frac{dz_k(t_1)}{dt} &\le \mu e^{\mu t_1}|y_k(t_1)| + e^{\mu t_1}\left[-b_k|y_k(t_1)| + \sum_{j=1}^{n}|\omega_{kj}|g'(\varsigma_j)|y_j(t_1-\tau_{kj})|\right]\\ &\le \mu z_k(t_1) - b_k z_k(t_1) + \sum_{j=1}^{n}|\omega_{kj}|g'(\varsigma_j)e^{\mu\tau_{kj}}z_j(t_1-\tau_{kj})\\ &\le -(b_k-\mu)z_k(t_1) + \sum_{j=1}^{n}|\omega_{kj}|g'(\xi_j)e^{\mu\tau_{kj}}\left[\sup_{\theta\in[t_1-\tau,t_1]}z_j(\theta)\right], \qquad (3.11)\end{aligned}$$

where $\xi_j = \hat{a}_j + \epsilon_0$ (resp., $\check{c}_j - \epsilon_0$) if $\alpha_j =$ "l" (resp., "r"). Herein, the invariance property of Λ^α in Theorem 3.1 has been applied. Due to $G_i(\mu) > 0$, we obtain

$$\begin{aligned}0 \le \frac{dz_k(t_1)}{dt} &\le -(b_k-\mu)z_k(t_1) + \sum_{j=1}^{n}|\omega_{kj}|g'(\xi_j)e^{\mu\tau_{kj}}\left[\sup_{\theta\in[t_1-\tau,t_1]}z_j(\theta)\right]\\ &< -\left\{b_i - \mu - \sum_{j=1}^{n}|\omega_{ij}|g'(\xi_j)e^{\mu\tau_{kj}}\right\}K\delta\\ &< 0, \qquad (3.12)\end{aligned}$$

which is a contradiction. Hence the claim (3.9) holds. Since $\delta > 1$ is arbitrary, by allowing $\delta \to 1^+$, we have $z_i(t) \le K$ for all $t > 0$, $i = 1, 2, \ldots, n$. We then use (3.7) and (3.8) to obtain

$$|x_i(t) - \bar{x}_i| \le e^{-\mu t}\max_{1\le j\le n}\left(\sup_{\theta\in[-\tau,0]}|x_j(\theta) - \bar{x}_j|\right)$$

for $t > 0$ and all $i = 1, 2, \ldots, n$. Therefore, $\mathbf{x}(t)$ is exponentially convergent to $\bar{\mathbf{x}}$. This completes the proof. □

In the following, we employ the theory of the local Lyapunov functional [15] and the Halanay-type inequality [4, 14] to establish other sufficient conditions for asymptotic stability and exponential stability for the equilibria of system (2.2).

THEOREM 3.3. *There exist 2^n asymptotically stable equilibria for system* (2.2) *under conditions* (H_1) *and* (H_2) *and one of the following conditions:*

(H_4): $2b_i > \sum_{j=1}^{n}|\omega_{ij}| + [g'(\eta_i)]^2\sum_{j=1}^{n}|\omega_{ji}|$ for $\eta_i = \hat{a}_i$ and $\check{c}_i$, $i = 1, 2, \ldots, n$,

(H_5): $\min_{1\le i\le n}\left[2b_i - \sum_{j=1}^{n}|\omega_{ij}|g'(\xi_j)\right] > \max_{1\le i\le n}\left[\sum_{j=1}^{n}|\omega_{ji}|g'(\eta_i)\right]$ for $\xi_j = \hat{a}_j$ and $\check{c}_j$, $\eta_i = \hat{a}_i$ and $\check{c}_i$.

Proof. Similarly to (3.1), there exists $\epsilon_0 > 0$ such that (H_4) holds for $\eta_i = \hat{a}_i + \epsilon_0$, $\check{c}_i - \epsilon_0$, and (H_5) holds for $\xi_j = \hat{a}_j + \epsilon_0$, $\check{c}_j - \epsilon_0$, $\eta_i = \hat{a}_i + \epsilon_0$, $\check{c}_i - \epsilon_0$, $i = 1, 2, \ldots, n$, by continuity of g'. We thus define Λ^α as in (3.3). The following computations are reserved for solutions lying entirely within each of the 2^n positively invariant regions Λ^α.

(i) We employ the following Lyapunov functional:

$$V(\mathbf{y})(t) = \sum_{i=1}^{n} y_i^2(t) + \sum_{i=1}^{n}\sum_{j=1}^{n} |\omega_{ij}| \int_{t-\tau_{ij}}^{t} [g(x_j(s)) - g(\overline{x}_j)]^2 ds,$$

where $\mathbf{y}(t) = \mathbf{x}(t) - \overline{\mathbf{x}}$. By recalling (3.6) and using (H_4), we derive

$$\begin{aligned}
\frac{dV(\mathbf{y})(t)}{dt} &= 2\sum_{i=1}^{n} y_i(t) \left\{ -b_i y_i(t) + \sum_{j=1}^{n} \omega_{ij}[g(x_j(t-\tau_{ij})) - g(\overline{x}_j)] \right\} \\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{n} |\omega_{ij}|[g(x_j(t)) - g(\overline{x}_j)]^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} |\omega_{ij}|[g(x_j(t-\tau_{ij})) - g(\overline{x}_j)]^2 \\
&\le -2\sum_{i=1}^{n} b_i y_i^2(t) + \sum_{i=1}^{n}\sum_{j=1}^{n} |\omega_{ij}| y_i^2(t) + \sum_{i=1}^{n}\sum_{j=1}^{n} |\omega_{ij}|[g'(\eta_j)]^2 y_j^2(t) \\
&= \sum_{i=1}^{n} \left\{ -2b_i + \sum_{j=1}^{n} |\omega_{ij}| + [g'(\eta_i)]^2 \sum_{j=1}^{n} |\omega_{ji}| \right\} y_i^2(t) < 0.
\end{aligned}$$

We thus conclude the asymptotic stability for equilibrium $\bar{\mathbf{x}}$ via applying the theory of the local Lyapunov functional; cf. [15].

(ii) Recall (3.6), and let

$$W(\mathbf{y})(t) = \frac{1}{2}\sum_{i=1}^{n} y_i^2(t). \tag{3.13}$$

Then,

$$\begin{aligned}
\frac{dW(\mathbf{y})(t)}{dt} &= \sum_{i=1}^{n} y_i(t) \left\{ -b_i y_i(t) + \sum_{j=1}^{n} \omega_{ij}[g(x_j(t-\tau_{ij})) - g(\overline{x}_j)] \right\} \\
&\le \sum_{i=1}^{n} \left\{ -b_i y_i^2(t) + \frac{1}{2}\sum_{j=1}^{n} |\omega_{ij}| g'(\varsigma_j)[y_i^2(t) + y_j^2(t-\tau_{ij})] \right\} \\
&\le -\sum_{i=1}^{n} \left[b_i - \frac{1}{2}\sum_{j=1}^{n} |\omega_{ij}| g'(\xi_j) \right] y_i^2(t) \\
&\quad + \frac{1}{2} \left[\max_{1\le i\le n} \sum_{j=1}^{n} |\omega_{ji}| g'(\eta_i) \right] \sum_{i=1}^{n} \sup_{t-\tau\le s\le t} y_i^2(s) \\
&\le -\beta W(\mathbf{y})(t) + \zeta \sup_{t-\tau\le s\le t} W(\mathbf{y})(s),
\end{aligned}$$

where

$$\begin{aligned}
\beta &:= \min_{1\le i\le n} \left\{ 2b_i - \sum_{j=1}^{n} |\omega_{ij}| g'(\xi_j), \xi_j = \hat{a}_j + \epsilon_0,\ \check{c}_j - \epsilon_0 \right\}, \\
\zeta &:= \max_{1\le i\le n} \left\{ \sum_{j=1}^{n} |\omega_{ji}| g'(\eta_i), \eta_i = \hat{a}_i + \epsilon_0,\ \check{c}_i - \epsilon_0 \right\}.
\end{aligned}$$

By (H_5), we have $\beta > \zeta > 0$. By using the Halanay inequality, we obtain that

$$W(\mathbf{y})(t) \le \left(\sup_{-\tau \le s \le 0} W(\mathbf{y})(s) \right) e^{-\gamma t} \tag{3.14}$$

for all $t \ge 0$, where γ is the unique solution of $\gamma = \beta - \zeta e^{\gamma \tau}$. It follows that

$$\frac{1}{2} \sum_{i=1}^{n} y_i^2(t) \le \left[\sup_{-\tau \le s \le 0} \left(\frac{1}{2} \sum_{i=1}^{n} y_i^2(s) \right) \right] e^{-\gamma t}. \tag{3.15}$$

Hence, the equilibrium $\bar{\mathbf{x}}$ is asymptotically stable. □

COROLLARY 3.4. *Under conditions* (H_1), (H_2), *and* (H_5), *there exist* 2^n *exponentially stable equilibria for system* (2.2).

We observe from (2.1) and (2.2) that for every i,

$$F_i(\mathbf{x}),\ \tilde{F}_i(\mathbf{x}_t) < 0 \text{ whenever } x_i > 0 \text{ is sufficiently large,}$$
$$F_i(\mathbf{x}),\ \tilde{F}_i(\mathbf{x}_t) > 0 \text{ whenever } x_i < 0 \text{ and } |x_i| \text{ is sufficiently large,}$$

since $b_i > 0$ and $\sum_{j=1}^{n} \omega_{ij} g_j(x_j(t)) + J_i$ and $\sum_{j=1}^{n} \omega_{ij} g_j(x_j(t - \tau_{ij})) + J_i$ are bounded for any $\mathbf{x}$ and $\mathbf{x}_t$. Therefore, it can be concluded that every solution of (2.1) and (2.2) is bounded in forward time.

4. Further extension. We shall extend our studies in sections 2 and 3 to more general activation functions in this section.

4.1. Activation functions in general form. Let us consider the activation functions $\{g_i(\cdot)\}_1^n$ which are $\mathcal{C}^2$ and satisfy

$$(C): \begin{cases} u_i \le g_i(\xi) \le v_i, & g_i'(\xi) > 0, \\ (\xi - \sigma_i) g_i''(\xi) < 0 & \text{for all } \xi \in \mathbb{R}, \end{cases}$$

$i = 1, 2, \ldots, n$. Herein, u_i, v_i, and σ_i are constants with $u_i < v_i$, $i = 1, 2, \ldots, n$. Under these circumstances, (H_1) can be modified to

$$(H_1'):\ 0 = \inf_{\xi \in \mathbb{R}} g_i'(\xi) < \frac{b_i}{\omega_{ii}} < \max_{\xi \in \mathbb{R}} g_i'(\xi)\ (= g_i'(\sigma_i)),\ i = 1, 2, \ldots, n.$$

As in section 2, we define

$$f_i(\xi) = -b_i \xi + \omega_{ii} g_i(\xi) + J_i.$$

LEMMA 4.1. *For* g_i *in the class* (C), *under condition* (H_1'), *there exist constants* $\{p_i\}_1^n$ *and* $\{q_i\}_1^n$ *with* $p_i < \sigma_i < q_i$ *such that* $f_i'(p_i) = f_i'(q_i) = 0$ *for each* $i = 1, 2, \ldots, n$.

We define

$$\hat{f}_i(\xi) = -b_i \xi + \omega_{ii} g_i(\xi) + k_i^+, \quad \check{f}_i(\xi) = -b_i \xi + \omega_{ii} g_i(\xi) + k_i^-, \tag{4.1}$$

where

$$k_i^+ := \sum_{j=1, j \ne i}^{n} \rho_j |\omega_{ij}| + J_i, \quad k_i^- := -\sum_{j=1, j \ne i}^{n} \rho_j |\omega_{ij}| + J_i \tag{4.2}$$

with $\rho_j = \max\{|u_j|, |v_j|\}$. We locate the points $\hat{a}_i < \hat{b}_i < \hat{c}_i$ and $\check{a}_i < \check{b}_i < \check{c}_i$, where $\hat{f}_i(\hat{a}_i) = \hat{f}_i(\hat{b}_i) = \hat{f}_i(\hat{c}_i) = 0$ and $\check{f}_i(\check{a}_i) = \check{f}_i(\check{b}_i) = \check{f}_i(\check{c}_i) = 0$.

Let $\eta \in \mathbb{R}$ and $k \in \{1, \dots, n\}$ be such that $g_k'(\eta) = \max\{g_i'(\xi) : \xi = \hat{a}_i, \check{c}_i, i = 1, 2, \dots, n\}$. Consider

$$(\mathrm{H}_3'): \; b_i > g_k'(\eta) \left[\omega_{ii} + \sum_{j=1, j\neq i}^{n} |\omega_{ij}| \right], \quad i = 1, 2, \dots, n.$$

THEOREM 4.2. *Let g_i be in the class* (C). *Under conditions* (H_1'), (H_2), *and* (H_3'), *there exist 3^n equilibria for systems* (2.1) *and* (2.2) *with 2^n among them being exponentially stable.*

4.2. Saturated activation functions. In this subsection, we investigate systems (2.1), (2.2) with saturated activation functions. In particular, we consider the following continuous functions:

$$g_i(\xi) = \begin{cases} u_i & \text{if } -\infty < \xi \le p_i, \\ \text{increasing} & \text{if } p_i \le \xi \le q_i, \\ v_i & \text{if } q_i \le \xi < \infty, \end{cases}$$

where p_i, q_i are constants with $p_i < q_i$ for $i = 1, 2, \dots, n$. Such a class of functions includes the piecewise linear function with saturations:

$$g_i(\xi) = \begin{cases} u_i & \text{if } -\infty < \xi \le p_i, \\ u_i + \frac{v_i - u_i}{q_i - p_i}(\xi - p_i) & \text{if } p_i \le \xi \le q_i, \\ v_i & \text{if } q_i \le \xi < \infty \end{cases}$$

for each i. Typical graphs for these functions are depicted in Figures 3(a) and (c). With such activation functions, existence of multiple equilibria for (2.1) and (2.2) can be obtained under condition

$$(\mathrm{H}_\mathrm{s}): b_i > 0, -b_i p_i + \omega_{ii} u_i + k_i^+ < 0, \quad -b_i q_i + \omega_{ii} v_i + k_i^- > 0, \quad i = 1, 2, \dots, n,$$

where k_i^+, k_i^- are defined as in (4.2). We define $\hat{f}_i, \check{f}_i$ as in (4.1). The graphs of $\hat{f}_i$ and $\check{f}_i$ are depicted in Figures 3(b) and (d). Under condition (H_s), we also locate the points $\hat{a}_i < \hat{b}_i < \hat{c}_i$ and $\check{a}_i < \check{b}_i < \check{c}_i$, where $\hat{f}_i(\hat{a}_i) = \hat{f}_i(\hat{b}_i) = \hat{f}_i(\hat{c}_i) = 0$ and $\check{f}_i(\check{a}_i) = \check{f}_i(\check{b}_i) = \check{f}_i(\check{c}_i) = 0$.

Note that we do not need differentiability at corner points p_i, q_i of g_i in our analysis; moreover, $g_i'(\xi) = 0$ for $\xi < p_i$ and $\xi > q_i$. Thus, (H_3) is already satisfied if $b_i > 0$ for $i = 1, 2, \dots, n$. With these formulations, we can derive that there exist 3^n equilibria for systems (2.1) and (2.2), and that 2^n of them are exponentially stable under condition (H_s).

4.3. Unbounded activation functions. Our theory can also be extended to certain unbounded activation functions with controlled slopes, for example, the activation functions g_i with bounded slopes in Figure 4. Herein, we require that the slopes m_i^r of the right- and m_i^l of the left-hand parts of g_i satisfy $b_i > \omega_{ii} m_i^\mathrm{r}$, $b_i > \omega_{ii} m_i^\mathrm{l}$ for $i = 1, \dots, n$.

5. Numerical illustrations. In this section, we present two examples (with delays) to illustrate our theory.

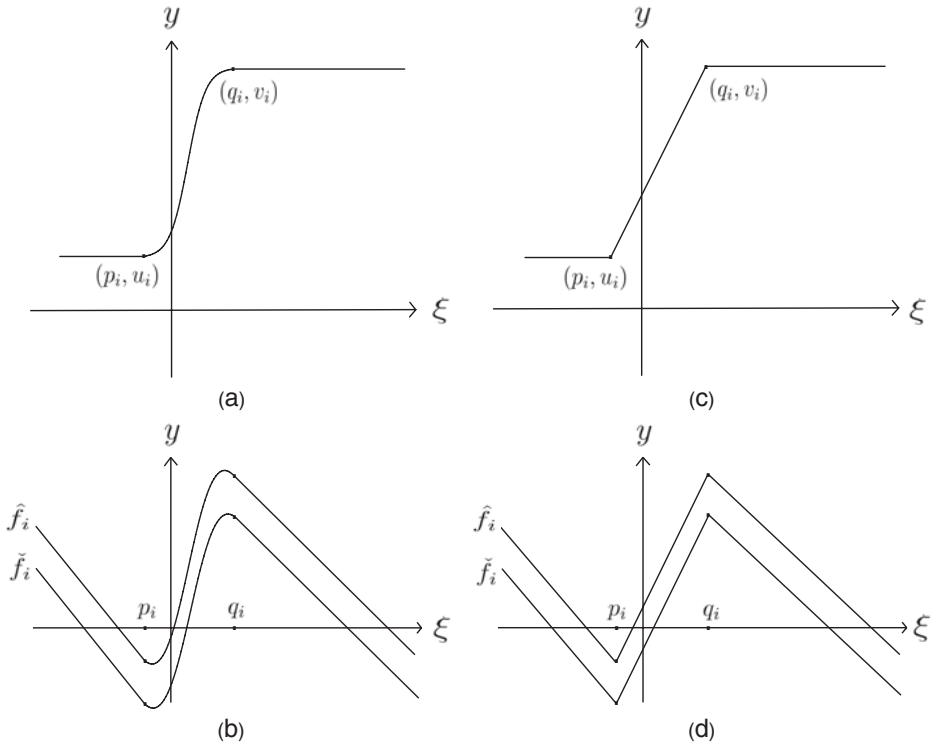


FIG. 3. (a) *The graph for a continuous activation function* g_i *with saturations.* (b) *The graphs for* $\hat{f}_i$ *and* $\check{f}_i$ *induced from the activation function in* (a). (c) *The graph for a piecewise linear activation function* g_i *with saturations.* (d) *The graphs for* $\hat{f}_i$ *and* $\check{f}_i$ *induced from the activation function in* (c).

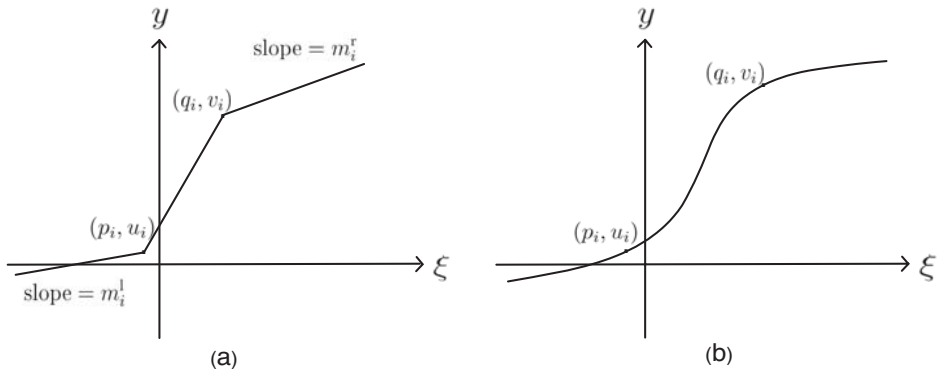


FIG. 4. (a) *The graph for an unbounded piecewise linear activation function.* (b) *The graph for an unbounded activation function with bounded slopes.*

Example 5.1. Consider the two-dimensional neural network

$$\frac{dx_1(t)}{dt} = -x_1(t) + 18g_1(x_1(t-10)) + 5g_2(x_2(t-10)) - 9,$$
$$\frac{dx_2(t)}{dt} = -3x_2(t) + 5g_1(x_1(t-10)) + 30g_2(x_2(t-10)) - 15,$$

TABLE 1
Local extreme points and zeros of $\hat{f}_1, \check{f}_1, \hat{f}_2, \check{f}_2$.

$\hat{a}_1 = -3.993889$	$p_1 = -1.762747$	$\hat{b}_1 = -0.757751$	q_1=1.762747	$\hat{c}_1$=14
$\check{a}_1 = -14$		$\check{b}_1$=0.757751		$\check{c}_1$=3.993889
$\hat{a}_2 = -3.320288$	$p_2 = -1.443635$	$\hat{b}_2 = -0.452309$	q_2=1.443635	$\hat{c}_2$=6.666650
$\check{a}_2 = -6.666650$		$\check{b}_2$=0.452309		$\check{c}_2$=3.320288

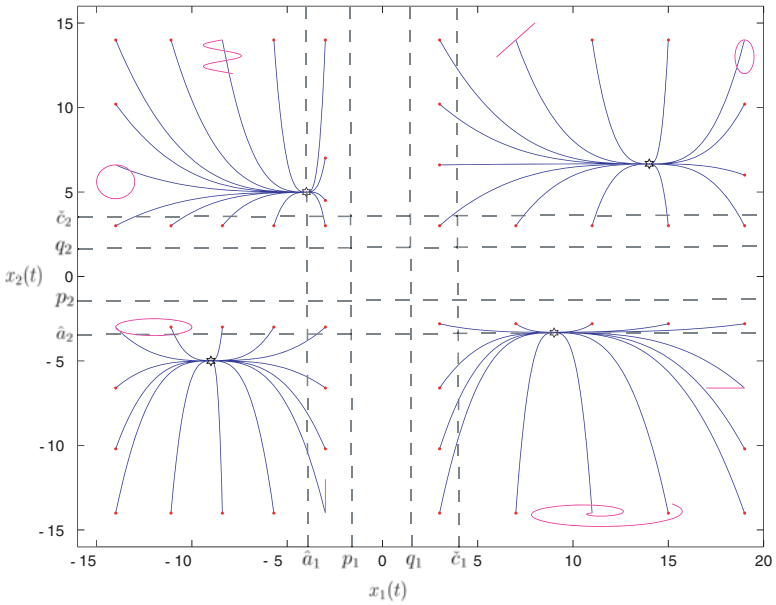


FIG. 5. *Illustrations for the dynamics in Example* 5.1.

where $g_1(x) = g_2(x) = g(x)$ in (2.4) with $\varepsilon = 0.5$. A computation gives

$$\hat{f}_1(x_1) = -x_1 + 18g(x_1) - 4, \quad \check{f}_1(x_1) = -x_1 + 18g(x_1) - 14,$$
$$\hat{f}_2(x_2) = -3x_2 + 30g(x_2) - 10, \quad \check{f}_2(x_2) = -3x_2 + 30g(x_2) - 20.$$

Herein, the parameters satisfy our conditions in Theorem 3.2:

$$\text{Condition } (\mathrm{H}_1)\text{: } 0 < \frac{b_1\varepsilon}{\omega_{11}} = \frac{1}{36} < \frac{1}{4}, \quad 0 < \frac{b_2\varepsilon}{\omega_{22}} = \frac{1}{20} < \frac{1}{4}.$$
$$\text{Condition } (\mathrm{H}_2)\text{: } \hat{f}_1(p_1) = -1.722534 < 0, \quad \check{f}_1(q_1) = 1.722534 > 0,$$
$$\hat{f}_2(p_2) = -4.085501 < 0, \quad \check{f}_2(q_2) = 4.085501 > 0.$$
$$\text{Condition } (\mathrm{H}_3)\text{: } b_1 = 1 > 0.025246 = \omega_{11}g'(\eta_1) + |\omega_{12}|g'(\eta_2),$$
$$b_2 = 3 > 0.081566 = |\omega_{21}|g'(\eta_1) + \omega_{22}g'(\eta_2),$$

where $\eta_1 = \pm 3.993889$, $\eta_2 = \pm 3.320288$ are defined in (2.9). Local extreme points and zeros of $\hat{f}_1, \check{f}_1, \hat{f}_2, \check{f}_2$ are listed in Table 1. The dynamics of this system are illustrated in Figure 5, where evolutions of 56 initial conditions have been tracked. The constant initial conditions are plotted in red dots, and the time-dependent initial conditions are plotted in purple curves. The evolutions of components $x_1(t)$ and $x_2(t)$ are depicted

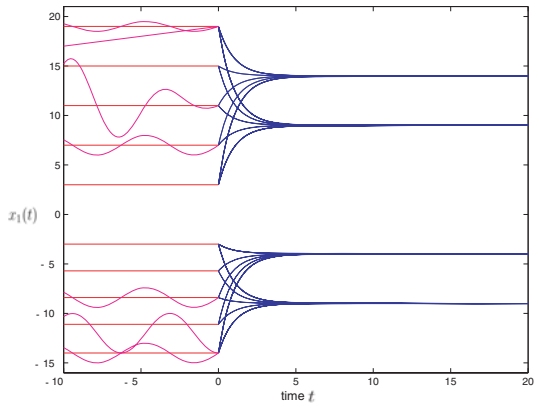


FIG. 6. *Evolution of state variable* $x_1(t)$ *in Example* 5.1.

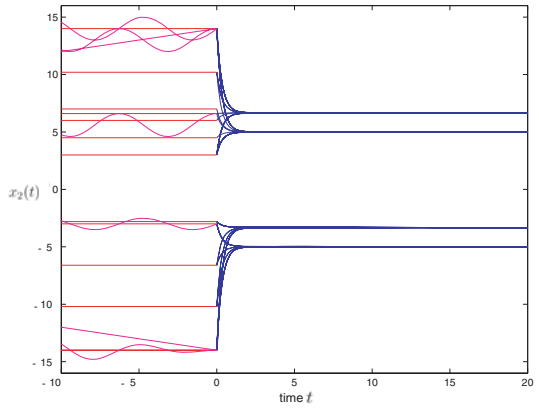


FIG. 7. *Evolution of state variable* $x_2(t)$ *in Example* 5.1.

in Figures 6 and 7, respectively. There are four exponentially stable equilibria in the system, as confirmed by our theory. The simulations demonstrate the convergence to these four equilibria from initial functions ϕ lying in the basin of the respective equilibrium.

Example 5.2. In this example, we simulate the neural network

$$\frac{dx_1(t)}{dt} = -x_1(t) + 18g_1(x_1(t-10)) + 11g_2(x_2(t-10)) + 1,$$
$$\frac{dx_2(t)}{dt} = -3x_2(t) + 11g_1(x_1(t-10)) + 30g_2(x_2(t-10)) + 4$$

with the output function $g_i(\xi) = h(\xi)$, where

$$h(\xi) = \frac{1}{2}(|\xi + 1| - |\xi - 1|), \tag{5.1}$$

for each i. The parameters also satisfy the conditions in our formulations with such an output function. We demonstrate the dynamics as well as evolutions of components $x_1(t)$, $x_2(t)$ for the system in Figures 8, 9, and 10, respectively.

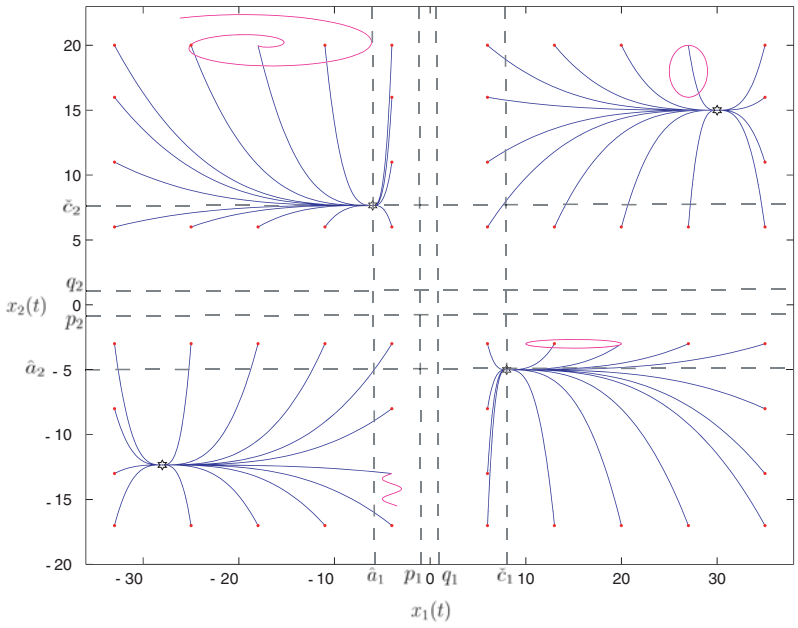


FIG. 8. *Illustrations for the dynamics in Example* 5.2.

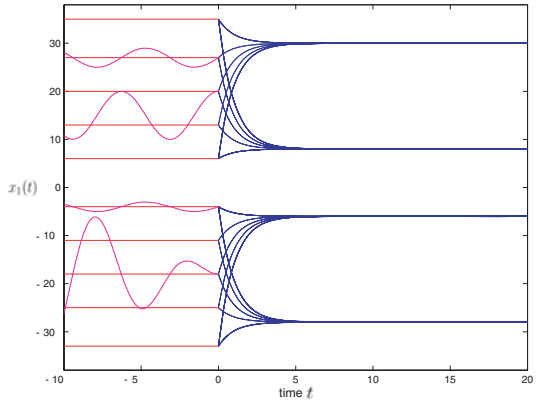


FIG. 9. *Evolution of state variable* $x_1(t)$ *in Example* 5.2.

6. Discussions. Our approach can also be adapted to the cellular neural networks with delays. The cellular neural networks (CNNs) were introduced by Chua and Yang [8] in 1988. A model called delayed cellular neural network [24] is given by

$$\frac{dx_i(t)}{dt} = -x_i(t) + \sum_{j\in N_r(i)} a_{ij}h(x_j(t)) + \sum_{j\in N_r(i)} b_{ij}h(x_j(t-\tau)) + J_i, \tag{6.1}$$

where $N_r(i) = \{i-1, i, i+1\}$ if $r = 1$. The standard activation function for such a network is the piecewise linear h defined in (5.1). Notably, (6.1) is a system of CNNs with cells coupled in the one-dimensional manner, and its local coupling structure is expressed in the equations. Global exponential stability of a single equilibrium for

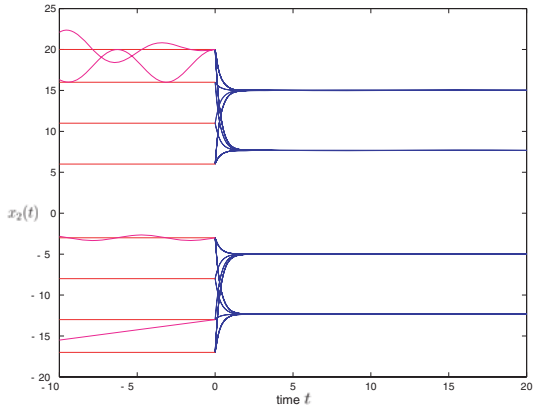


FIG. 10. *Evolution of state variable* $x_2(t)$ *in Example* 5.2.

(6.1) has been studied by many researchers, for instance, the authors of [3, 20]. The CNNs can be built by multidimensional couplings among cells. Since there are finitely many cells at most, the CNNs can always be rewritten in a one-dimensional coupling form by renaming the indices [28]. It can then be written in a form similar to (1.1). Such an arrangement, however, destroys the local connection representation. While previous studies on multistability for the CNNs without delays [17, 26, 27] employed the structure of local connections among cells of CNNs, our approach does not rely on such a structure. Moreover, our theory generalized the multistability to the CNNs with delays (6.1).

In this investigation, we have obtained existence of 2^n stable stationary solutions for recurrent neural networks comprised of n neurons, with delays and without delays. The theory is primarily based upon an observation on the structures of the equations. It is thus rather general and can be applied to at least the Hopfield-type neural networks and the cellular neural networks. The analysis is valid for the networks with various activation functions, including the typical sigmoidal ones and the saturated linear ones, as well as some unbounded activation functions. In fact, our formulation depends on the configuration of the activation functions instead of the precise form of the functions. The theorems thus developed are pertinent in neural network theory.

Stable periodic orbits and limit cycle attractors are also important for memory storage and other neural activities. By similar analysis, we can also establish existence of multiple limit cycles for systems (1.1) and (6.1) with periodic inputs $J_i = J_i(t) = J_i(t + T)$. The result will be reported in another article. The approach in this presentation can be adopted to discrete-time neural networks as well.

The major discussions on neural networks have been centered around monostability, in an abundance of articles in the areas of physics, information sciences, electrical engineering, and mathematics. Multistability in neural networks is, however, essential in numerous applications such as content-addressable memory storage and pattern recognition. Recently, further application potentials of multistability have been found in decision making, digital selection, and analogy amplification [18].

We have exploited further interesting structures of Hopfield-type neural networks in this study. Our investigations have provided computable parameter conditions for multistable dynamics in the recurrent neural networks and are expected to contribute toward practical applications.

Acknowledgment. The authors are grateful to the reviewers for their suggestions on improving the presentation.

REFERENCES

[1] J. Bélair, S. A. Campbell, and P. Van Den Driessche, *Frustration, stability, and delay-induced oscillations in a neural network model*, SIAM J. Appl. Math., 56 (1996), pp. 245–255.

[2] S. A. Campbell, R. Edwards, and P. Van Den Driessche, *Delayed coupling between two neural network loops*, SIAM J. Appl. Math., 65 (2004), pp. 316–335.

[3] J. Cao, *New results concerning exponential stability and periodic solutions of delayed cellular neural networks*, Phys. Lett. A, 307 (2003), pp. 136–147.

[4] J. Cao and J. Wang, *Absolute exponential stability of recurrent neural networks with Lipschitz-continuous activation functions and time delays*, Neural Netw., 17 (2004), pp. 379–390.

[5] T. Chen, *Global exponential stability of delayed Hopfield neural networks*, Neural Netw., 14 (2001), pp. 977–980.

[6] S. S. Chen and C. W. Shih, *Transversal homoclinic orbits in a transiently chaotic neural network*, Chaos, 12 (2002), pp. 654–670.

[7] L. O. Chua, *CNN: A Paradigm for Complexity*, World Scientific, River Edge, NJ, 1998.

[8] L. O. Chua and L. Yang, *Cellular neural networks: Theory*, IEEE Trans. Circuits and Systems, 35 (1988), pp. 1257–1272.

[9] F. Forti, *On global asymptotic stability of a class of nonlinear systems arising in neural network theory*, J. Differential Equations, 113 (1994), pp. 246–264.

[10] M. Forti and A. Tesi, *New conditions for global stability of neural networks with application to linear and quadratic programming problems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 42 (1995), pp. 354–366.

[11] J. Foss, A. Longtin, B. Mensour, and J. Milton, *Multistability and delayed recurrent loops*, Phys. Rev. Lett., 76 (1996), pp. 708–711.

[12] K. Gopalsamy, *Stability and Oscillations in Delay Differential Equations of Population Dynamics*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992.

[13] K. Gopalsamy and X. He, *Stability in asymmetric Hopfield nets with transmission delays*, Phys. D, 76 (1994), pp. 344–358.

[14] A. Halany, *Differential Equations*, Academic Press, New York, 1966.

[15] J. Hale and S. V. Lunel, *Introduction to Functional-Differential Equations*, Springer-Verlag, New York, 1993.

[16] J. Hopfield, *Neurons with graded response have collective computational properties like those of two-state neurons*, Proc. Natl. Acad. Sci. USA, 81 (1984), pp. 3088–3092.

[17] J. Juang and S.-S. Lin, *Cellular neural networks: Mosaic pattern and spatial chaos*, SIAM J. Appl. Math., 60 (2000), pp. 891–915.

[18] R. L. T. Hahnloser, *On the piecewise analysis of networks of linear threshold neurons*, Neural Netw., 11 (1998), pp. 691–697.

[19] X. Liao, G. Chen, and E. N. Sanchez, *Delay-dependent exponential stability analysis of delayed neural networks: An LMI approach*, Neural Netw., 15 (2002), pp. 855–866.

[20] S. Mohamad and K. Gopalsamy, *Exponential stability of continuous-time and discrete-time cellular neural networks with delays*, Appl. Math. Comput., 135 (2003), pp. 17–38.

[21] M. Morita, *Associative memory with non-monotone dynamics*, Neural Netw., 6 (1993), pp. 115–126.

[22] L. Olien and J. Bélair, *Bifurcations, stability, and monotonicity properties of a delayed neural network model*, Phys. D, 102 (1997), pp. 349–363.

[23] J. Peng, H. Qiao, and Z. B. Xu, *A new approach to stability of neural networks with time-varying delays*, Neural Netw., 15 (2002), pp. 95–103.

[24] T. Roska and L. O. Chua, *Cellular neural networks with non-linear and delay-type template elements and non-uniform grids*, Int. J. Circuit Theory Appl., 20 (1992), pp. 469–481.

[25] L. P. Shayer and S. A. Campbell, *Stability, bifurcation, and multistability in a system of two coupled neurons with multiple time delays*, SIAM J. Appl. Math., 61 (2000), pp. 673–700.

[26] C.-W. Shih, *Pattern formation and spatial chaos for cellular neural networks with asymmetric templates*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 8 (1998), pp. 1907–1936.

[27] C.-W. Shih, *Influence of boundary conditions on pattern formation and spatial chaos in lattice systems*, SIAM J. Appl. Math., 61 (2000), pp. 335–368.

[28] C.-W. Shih and C. W. Weng, *On the template corresponding to cycle-symmetric connectivity in cellular neural networks*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 12 (2002), pp. 2957–2966.

[29] P. Van Den Driessche and X. Zou, *Global attractivity in delayed Hopfield neural network models*, SIAM J. Appl. Math., 58 (1998), pp. 1878–1890.
[30] P. Van Den Driessche, J. Wu, and X. Zou, *Stabilization role of inhibitory self-connections in a delayed neural network*, Phys. D, 150 (2001), pp. 84–90.
[31] D. Xu, H. Zhao, and H. Zhu, *Global dynamics of Hopfield neural networks involving variable delays*, Comput. Math. Appl., 42 (2001), pp. 39–45.
[32] J. F. Yang and C. M. Chen, *Winner-take-all neural networks using the highest threshold*, IEEE Trans. Neural Networks, 11 (2000), pp. 194–199.
[33] Z. Yi, P. A. Heng, and P. F. Fung, *Winner-take-all discrete recurrent neural networks*, IEEE Trans. Circuits Syst. II, 47 (2000), pp. 1584–1589.
[34] J. Zhang and X. Jin, *Global stability analysis in delayed Hopfield neural network models*, Neural Netw., 13 (2002), pp. 745–753.
[35] H. Zhao, *Global asymptotic stability of Hopfield neural network involving distributed delays*, Neural Netw., 17 (2004), pp. 47–53.

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1321–1338

MULTIPLE EQUILIBRIA IN COMPLEX CHEMICAL REACTION NETWORKS: II. THE SPECIES-REACTION GRAPH*

GHEORGHE CRACIUN† AND MARTIN FEINBERG‡

Abstract. For mass action kinetics, the capacity for multiple equilibria in an isothermal homogeneous continuous flow stirred tank reactor is determined by the structure of the underlying network of chemical reactions. We suggest a new graph-theoretical method for discriminating between complex reaction networks that can admit multiple equilibria and those that cannot. In particular, we associate with each network a species-reaction graph, which is similar to reaction network representations drawn by biochemists, and we show that, if the graph satisfies certain weak conditions, the differential equations corresponding to the network cannot admit multiple equilibria *no matter what values the rate constants take*. Because these conditions are very mild, they amount to powerful (and quite delicate) necessary conditions that a network must satisfy if it is to have the capacity to engender multiple equilibria. Broad qualitative results of this kind are especially apt, for individual reaction rate constants are rarely known fully for complex reaction networks (if they are known at all). Some concluding remarks address connections to biology.

Key words. equilibrium points, chemical reaction networks, mass action kinetics, SR graph

AMS subject classifications. 92C45, 65H10, 80A30, 37C25

DOI. 10.1137/050634177

1. Introduction.

The purpose of this article is to provide theory for distinguishing between complex chemical reaction networks that have the capacity to admit multiple positive equilibria and those that do not. In particular, we shall be interested in networks governed by mass action kinetics and operating in the context of what chemical engineers call the continuous flow stirred tank reactor (CFSTR [1]). Models in cell biology sometimes invoke pictures and mathematics reminiscent of CFSTRs [9, 12, 17, 14], so it not unreasonable to expect that theory presented here might ultimately provide insight that is useful in biological applications. Indeed, in biology one rarely has detailed knowledge of reaction rate constants; at the outset, then, it is especially appropriate to seek a *qualitative* understanding of the relationship between reaction network structure and the capacity for various kinds of behavior (e.g., bistability). As we indicated in the first article of this series [4], the connection between the two is quite delicate. The theory offered here is intended to render the relationship between reaction network structure and behavior more concrete.

Our principal results will serve to describe very large classes of networks, including highly complex ones, that cannot give rise to multiple steady states regardless of parameter values. These results provide very strong necessary conditions that a network must satisfy if it is to have the capacity to give rise, for example, to bistable behavior.

*Received by the editors June 21, 2005; accepted for publication (in revised form) December 19, 2005; published electronically March 31, 2006.
http://www.siam.org/journals/siap/66-4/63417.html

†Mathematical Biosciences Institute, The Ohio State University, Columbus, OH 43210, and Department of Mathematics and Department of Biomolecular Chemistry, University of Wisconsin, Madison, WI 53706 (craciun@math.wisc.edu). This author was supported by the National Science Foundation under agreement 0112050.

‡Department of Chemical Engineering and Department of Mathematics, The Ohio State University, Columbus, OH 43210 (feinberg.14@osu.edu). This author was supported by the National Science Foundation through grants BES 0222367 and BES-0425459.

Denial of the capacity of a reaction network to admit multiple equilibria follows from inspection of what we call the Species-Reaction Graph (or SR graph [3]) for the network, which is similar to the reaction diagram often drawn by biochemists. Properties of the SR graph and results about it are similar to those in earlier work [8, 15, 16] on properties of a related graph, called the Species-Complex-Linkage Graph (or SCL graph). However, the newer results presented here are substantially more generous in the information they give. Inspection of the SR graph often tells one very quickly that the network under study is, in the sense of a previous article [4], *injective*, which in turn implies that multiple positive equilibria are impossible. That is, inspection of the SR graph for a reaction network will often tell one that the complex nonlinear system of differential equations associated with the network cannot admit multiple positive equilibria, no matter what values the (generally unknown) parameter values might take.

Our aim in this introductory section is to present, in an informal way, the main theorem of this article, largely motivated by a single example. More formal definitions are given in section 2, which will prepare the groundwork for proofs.

A CFSTR consists of a perfectly stirred vessel along with two streams, a feed stream that carries reactants to the vessel and an outflow stream that leaves the vessel, carrying away mixture having the same instantaneous composition as that within the vessel. Hereafter we suppose that the mixtures involved are liquids, all of which have the same time-invariant density, that the mixture within the vessel is maintained at a fixed temperature, and that the feed and outflow streams have the same volumetric flow rate, g (volume/time). For the purposes of an example, we will suppose that (1.1) is a network of chemical reactions among species A, B, M, N, R, X, Y, and Z:

$$A + M \rightleftharpoons X, \quad B + N \rightleftharpoons Y \to 2A + N, \quad B + X \rightleftharpoons Z \to R + M. \tag{1.1}$$

By virtue of the occurrence of chemical reactions, the molar concentrations of the various species within the vessel will generally depend on time. These we denote by $c_A(t), c_B(t), \ldots, c_Z(t)$, which, by supposition, are identical to the species concentrations in the outflow stream. We denote by $c_A^f, c_B^f, \ldots, c_Z^f$ the (fixed) concentrations of the species in the feed stream. We assume hereafter that the rates of the chemical reactions are governed by mass action kinetics [4, 6, 7, 10, 11]. In this case, the system of differential equations associated with network (1.1) is the following:

$$\begin{aligned}
\dot{c}_A &= (g/V)(c_A^f - c_A) - k_{A+M\to X}c_Ac_M + k_{X\to A+M}c_X + 2k_{Y\to 2A+N}c_Y,\\
\dot{c}_B &= (g/V)(c_B^f - c_B) - k_{B+X\to Z}c_Bc_X + k_{Z\to B+X}c_Z + k_{Y\to B+N}c_Y\\
&\quad - k_{B+N\to Y}c_Bc_N,\\
\dot{c}_M &= (g/V)(c_M^f - c_M) - k_{A+M\to X}c_Ac_M + k_{X\to A+M}c_X + k_{Z\to R+M}c_Z,\\
\dot{c}_N &= (g/V)(c_N^f - c_N) - k_{B+N\to Y}c_Bc_N + k_{Y\to B+N}c_Y + k_{Y\to 2A+N}c_Y,\\
\dot{c}_R &= (g/V)(c_R^f - c_R) + k_{Z\to R+M}c_Z,\\
\dot{c}_X &= (g/V)(c_X^f - c_X) - k_{X\to A+M}c_X + k_{A+M\to X}c_Ac_M - k_{B+X\to Z}c_Bc_X\\
&\quad + k_{Z\to B+X}c_Z,\\
\dot{c}_Y &= (g/V)(c_Y^f - c_Y) - k_{Y\to B+N}c_Y + k_{B+N\to Y}c_Bc_N - k_{Y\to 2A+N}c_Y,\\
\dot{c}_Z &= (g/V)(c_Z^f - c_Z) + k_{B+X\to Z}c_Bc_X - k_{Z\to B+X}c_Z - k_{Z\to R+M}c_Z,
\end{aligned} \tag{1.2}$$

where g is the volumetric flow rate (volume/time), V is the reactor volume, $k_{A+M\to X}$ is the rate constant of the reaction $A + M \to X$, $k_{X\to A+M}$ is the rate constant of the

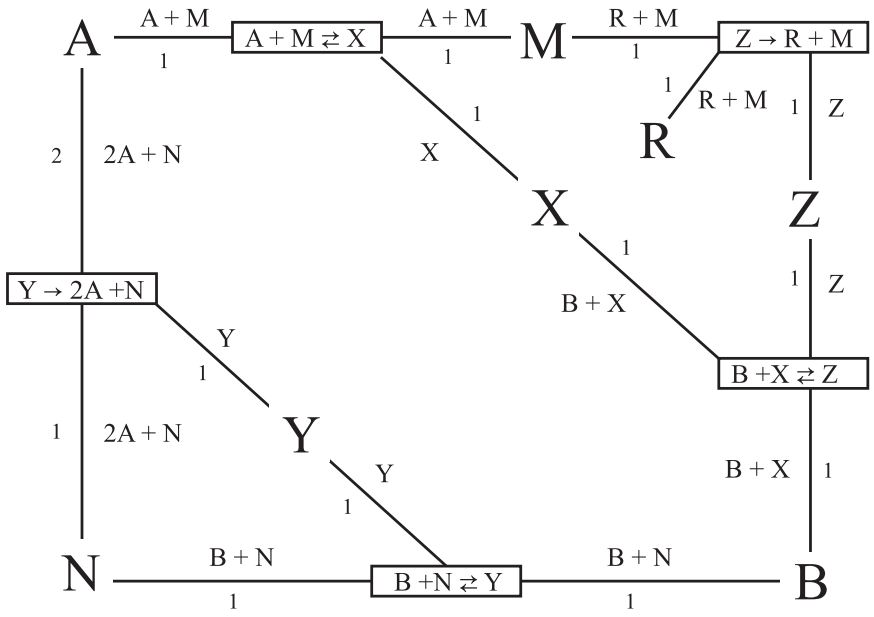


FIG. 1.1. *The SR graph* Γ *of the reaction network* (1.1).

reaction $X \to A + M$, and so on. See [4, 6] for a detailed explanation of how, given a reaction network, we obtain such a system of differential equations.

We say that the reaction network (1.1) *has the capacity for multiple positive equilibria* if there exist some positive flow rate g, positive volume V, nonnegative feed concentrations $c_A^f, \ldots, c_Z^f$, and positive rate constants $k_{A+M\to X}, \ldots, k_{Z\to R+M}$ such that the system of differential equations (1.2) has two or more distinct equilibria at which the species concentrations are positive.

In preparation for a description of how to draw the SR graph for a reaction network, we need a little vocabulary: By the *complexes* [10] of a reaction network we mean the objects at the heads and tails of the reaction arrows. Thus, the complexes of network (1.1) are $A + B$, X, $B + N$, Y, $2A + N$, $B + X$, Z, and $R + M$.

The SR graph for a reaction network has two kinds of nodes: *species nodes* and *reaction nodes*. There is a species node for each species in the network (A, B, M, N, R, Y, and Z in (1.1)). Moreover, there is a reaction node for each reaction or reversible reaction pair in the network. That is, reversible reactions such as $A + M \rightleftharpoons X$ share the same node. *Edges* join species nodes to reaction nodes as follows: If a species (such as A) appears in a complex (such as $A + M$) at the head or tail of a reaction (such as $A + M \rightleftharpoons X$), then an (unoriented) edge joins the species node to the reaction node and is *labeled* with the name of the complex in which that species appears. (Thus, for example, an edge would join the species node corresponding to A to the reaction node corresponding to $A + M \rightleftharpoons X$, and the edge would be labeled $A + M$.) The SR graph for network (1.1) is shown in Figure 1.1.

We now need to define some features of SR graphs that are especially relevant to our problem. Pairs of edges that meet at a reaction node and have the same complex label are called *c-pairs* (complex pairs). For example, the two edges labeled $A + M$ that meet at the reaction node $A + M \rightleftharpoons X$ in Figure 1.1 form a c-pair.

Notice that *cycles* might appear in the SR graph. Cycles that contain an odd

number of c-pairs are called *o-cycles* (odd cycles). For example, the outer cycle in Figure 1.1 is an o-cycle, since it contains three c-pairs, centered at the reaction nodes $A + M \rightleftharpoons X$, $B + N \rightleftharpoons Y$, and $Y \to 2A + N$. Cycles that contain an even number of c-pairs are called *e-cycles* (even cycles). In particular, cycles that contain no c-pairs are e-cycles.

The *stoichiometric coefficient* of an edge is the coefficient of the species adjacent to that edge in the complex label of the edge. For the reader's convenience, we have labeled each edge of the SR graph in Figure 1.1 with its stoichiometric coefficient. For example, the stoichiometric coefficient of the edge from A to $A + M \rightleftharpoons X$ in Figure 1.1 is 1, and the stoichiometric coefficient of the edge from A to $Y \to 2A + N$ is 2. Cycles for which alternately multiplying and dividing the stoichiometric coefficients along the cycle gives the final result 1 are called *s-cycles* (stoichiometric cycles). For example, for the outer cycle that begins at A and visits N, B, Z M, the stoichiometric coefficients along the cycle are $2, 1, 1, 1, 1, 1, 1, 1, 1$ (see Figure 1.1). Then, by alternately multiplying and dividing the stoichiometric coefficients along the cycle, we get $2 \cdot 1^{-1} \cdot 1 \cdot 1^{-1} \cdot 1 \cdot 1^{-1} \cdot 1 \cdot 1^{-1} \cdot 1 \cdot 1^{-1} = 2$, and thus this cycle is not an s-cycle. On the other hand, for the cycle that visits N, Y we get $1 \cdot 1^{-1} \cdot 1 \cdot 1^{-1} = 1$, and thus this cycle is an s-cycle.

We say that two cycles in the SR graph *have a species-to-reaction (S-to-R) intersection* if the common edges of the two cycles constitute a path that begins at a species node and ends at a reaction node, or if they constitute a disjoint union of such paths.

For example, the common edges of the cycle that visits N, Y with the cycle that visits A, X, B, Y in Figure 1.1 form a path that begins at a reaction node and ends at a reaction node, and so they do *not* have an S-to-R intersection.

Then the main result of this article is the following.

THEOREM 1.1. *Consider a reaction network such that in its SR graph*

(i) *each cycle is an o-cycle or an s-cycle,*

(ii) *no two e-cycles have an S-to-R intersection.*

Then, taken with mass action kinetics, the reaction network does not have the capacity for multiple positive equilibria.

In particular, the theorem above implies that the reaction network (1.1) does not have the capacity for multiple equilibria. Indeed, all cycles in the SR graph of (1.1) are o-cycles, except for two cycles (the cycle that visits N, Y and the cycle that visits M, Z, X), which are s-cycles, and so condition (i) is satisfied. Also, these two e-cycles do not have an S-to-R intersection, so condition (ii) is satisfied. On the other hand, previous results in [8, 15, 16] give no information about network (1.1).

In general, if there are no cycles in the SR graph, or if all cycles are o-cycles, then conditions (i) and (ii) are satisfied. Or, if all stoichiometric coefficients in a network are one, then all cycles are s-cycles, and so condition (i) is satisfied. Also, if no species node is adjacent to three or more reaction nodes, then no two cycles have an S-to-R intersection, and so condition (ii) is satisfied. Note then that for some reaction networks it is not even necessary to draw the SR graph in order to conclude that they do not have the capacity for multiple equilibria: If all the stoichiometric coefficients are one and no species appears in three or more reactions, then the reaction network does not have the capacity for multiple equilibria. For example, if we replace $2A + N$ by $A + N$ in (1.1), then the new reaction network has all stoichiometric coefficients equal to one, and no species appears in three or more reactions. Therefore, without having to draw its SR graph, it follows that this new reaction network does not have the capacity for multiple equilibria.

In the next section we begin our proof of Theorem 1.1. In particular, we will eventually want to connect the SR graph to network *injectivity*, an idea introduced in [4].

2. Reaction networks and injectivity. Let us first give a precise definition of a reaction network in terms of the set of species, the set of complexes, and the set of reactions. Recall that the *complexes* of a reaction network are to be understood as the objects at the head or tail of reaction arrows. We denote by $\mathbb{R}$ the set of real numbers, by $\mathbb{R}_+$ the set of positive numbers, and by $\bar{\mathbb{R}}_+$ the set of nonnegative numbers. Also, given a set I, we denote by $\mathbb{R}^I$ the vector space of formal linear combinations $\sum_{i\in I}\lambda_i i$, generated by the elements of $i \in I$, with coefficients $\lambda_i \in \mathbb{R}$. By $\bar{\mathbb{R}}^I_+$ we mean the members of $\mathbb{R}^I$ with $\lambda_i \geq 0$ for all $i \in I$. By $\mathbb{R}^I_+$ we mean the members of $\mathbb{R}^I$ with $\lambda_i > 0$ for all $i \in I$. By the *support* of an element $x \in \mathbb{R}^I_+$ we mean the set $\operatorname{supp}(x) = \{i \in I :\ x_i \neq 0\}$.

As in [4], we can regard a chemical reaction network as an abstract structure given by the following definition.

DEFINITION 2.1 (see [6, 7]). *A* chemical reaction network $\mathcal{N} = (\mathscr{S}, \mathscr{C}, \mathscr{R})$ *consists of three finite sets:*

(i) *a set* $\mathscr{S}$ *of* species *of the network,*
(ii) *a set* $\mathscr{C} \subset \bar{\mathbb{R}}^{\mathscr{S}}_+$ *of* complexes *of the network,*
(iii) *a set* $\mathscr{R} \subset \mathscr{C} \times \mathscr{C}$ *of* reactions*, with the following properties:*
  (a) $(y, y) \notin \mathscr{R}$ *for any* $y \in \mathscr{C}$*,*
  (b) *for each* $y \in \mathscr{C}$ *there exists* $y' \in \mathscr{C}$ *such that* $(y, y') \in \mathscr{R}$ *or such that* $(y', y) \in \mathscr{R}$*.*

We write the more suggestive $y \to y'$ in place of (y, y') when (y, y') is a member of $\mathscr{R}$. Also, if $\{y \to y',\ y' \to y\} \subset \mathscr{R}$, we will denote the set $\{y \to y',\ y' \to y\}$ by the more suggestive $y \rightleftharpoons y'$ and will say that $y \rightleftharpoons y'$ is a *reversible reaction.* If $y \to y' \in \mathscr{R}$ and $y' \to y \notin \mathscr{R}$, we say that $y \to y'$ is an *irreversible reaction.* For example, consider the reaction network

$$A + B \rightleftharpoons C, \qquad A \to 2B. \tag{2.1}$$

In this case $\mathscr{S} = \{A, B, C\}$, $\mathscr{C} = \{A + B, C, A, 2B\}$, $\mathscr{R} = \{A + B \to C,\ C \to A + B,\ A \to 2B\}$.

DEFINITION 2.2. *By a* mass action system *we mean a reaction network* $(\mathscr{S}, \mathscr{C}, \mathscr{R})$ *taken together with an element* $k \in \mathbb{R}^{\mathscr{R}}_+$*. The number* $k_{y\to y'}$ *is the* rate constant *for the reaction* $y \to y'$*.*

In the next definition we use the following notation: for two vectors in $\bar{\mathbb{R}}^{\mathscr{S}}_+$, say $u = \sum_{s\in\mathscr{S}} u_s s$ and $v = \sum_{s\in\mathscr{S}} v_s s$, we denote by u^v the product $\prod_{s\in\mathscr{S}}(u_s)^{v_s}$. Here we use the convention that $0^0 = 1$.

Our aim now is to write the differential equation that, for a mass action system, governs the evolution of composition vector $c \in \bar{\mathbb{R}}^{\mathscr{S}}_+$.

DEFINITION 2.3. *For a mass action system* $(\mathscr{S}, \mathscr{C}, \mathscr{R}, k)$ *the associated differential equation is*

$$\dot{c} = \sum_{y\to y'\in\mathscr{R}} k_{y\to y'} c^y (y' - y). \tag{2.2}$$

In components, the differential equations associated with a mass action system derived from network (2.1) are

$$\begin{aligned}\dot{c}_A &= -k_{A+B\to C}c_Ac_B - k_{A\to 2B}c_A + k_{C\to A+B}c_C,\\ \dot{c}_B &= -k_{A+B\to C}c_Ac_B + k_{C\to A+B}c_C + 2k_{A\to 2B}c_A,\\ \dot{c}_C &= k_{A+B\to C}c_Ac_B - k_{C\to A+B}c_C.\end{aligned} \tag{2.3}$$

Note that these are not the differential equations one would write for a CFSTR, for they take no account of the effects of the feed and outflow streams. The appropriate CFSTR differential equations are

$$\begin{aligned}\dot{c}_A &= (g/V)(c_A^f - c_A) - k_{A+B\to C}c_Ac_B - k_{A\to 2B}c_A + k_{C\to A+B}c_C,\\ \dot{c}_B &= (g/V)(c_B^f - c_B) - k_{A+B\to C}c_Ac_B + k_{C\to A+B}c_C + 2k_{A\to 2B}c_A,\\ \dot{c}_C &= (g/V)(c_C^f - c_C) + k_{A+B\to C}c_Ac_B - k_{C\to A+B}c_C.\end{aligned} \tag{2.4}$$

As we indicated in [4], however, the appropriate CFSTR equations *do* derive from a mass action system associated with the augmented network

$$A + B \rightleftharpoons C,\ A \to 2B,\ 0 \to A,\ 0 \to B,\ 0 \to C,\ A \to 0,\ B \to 0,\ C \to 0. \tag{2.5}$$

Here 0 is the *zero complex*, which is understood to be the zero vector of $\bar{\mathbb{R}}_+^{\mathscr{S}}$. As explained in [4], the added *outflow reactions* $A \to 0$, $B \to 0$, and $C \to 0$ serve to model the contributions of the outflow stream to the CFSTR differential equations (taking each rate constant to be g/V), while the *feed reactions* $0 \to A$, $0 \to B$, and $0 \to C$ serve to model the contributions of the feed stream (taking the rate constants to be, respectively, gc_A^f/V, gc_B^f/V, and gc_C^f/V).

In general, to obtain the augmented network, one adds to the network of true chemical reactions an outflow reaction $s \to 0$ for each $s \in \mathscr{S}$, and a feed reaction $0 \to s$ for each species s deemed to be in the feed stream. *Hereafter, when we speak of a reaction network $(\mathscr{S},\mathscr{C},\mathscr{R})$, it will be understood that we have in mind the augmented network constructed to generate the CFSTR differential equations. In particular, the full set of reactions $\mathscr{R}$ will contain the "true" set of chemical reactions (denoted $\mathscr{R}_t$) and a reaction $s \to 0$ for each $s \in \mathscr{S}$.*

DEFINITION 2.4. *A reaction network $(\mathscr{S},\mathscr{C},\mathscr{R})$* has the capacity to admit multiple positive equilibria *if there is a $k \in \mathbb{R}_+^{\mathscr{R}}$ such that, for the mass action system $(\mathscr{S},\mathscr{C},\mathscr{R},k)$, the associated differential equation admits two distinct equilibria in $\mathbb{R}_+^{\mathscr{S}}$.*

Remark. Our aim will be to describe networks that do *not* have the capacity for multiple positive equilibria. For our study of classical CFSTRs this is apparently a little more than we need: In Definition 2.4, we permit the rate constants associated with the outflow reactions (i.e., reactions of the form $s \to 0$ for all $s \in \mathscr{S}$) to take arbitrary positive values, while for the classical CFSTRs such rate constants should all be identical (and equal to g/V).

We are interested in what we call *injective* reaction networks because *injective reaction networks do not have the capacity for multiple positive equilibria* (see [4]). The characterization of injectivity we use here is the one given by Theorem 3.3 in [4].

DEFINITION 2.5. *A reaction network $\mathcal{N} = (\mathscr{S},\mathscr{C},\mathscr{R})$ with n species is* injective *if*

$$\det(y_1,\dots,y_n)\det(y_1 - y_1',\dots,y_n - y_n') \geq 0$$

for all choices of reactions[1] *$y_1 \to y_1',\dots,y_n \to y_n'$ in $\mathscr{R}$.*

[1] Some of these reactions could be feed or outflow reactions.

Therefore, we need to study the relationship between the signs of $\det(y_1, \dots, y_n)$ and $\det(y_1 - y_1', \dots, y_n - y_n')$. For this, our main tool will be the SR graph.

3. The SR graph.

DEFINITION 3.1. *Consider some reaction network* $\mathcal{N} = (\mathscr{S}, \mathscr{C}, \mathscr{R})$. *The* SR graph $\Gamma_{\mathcal{N}}$ *of* $\mathcal{N}$ *is an unoriented graph defined as follows. Each node of* $\Gamma_{\mathcal{N}}$ *is either a* species node *or a* reaction node. *There is one species node for each species in* $\mathscr{S}$. *There is one reaction node for each reversible reaction in* $\mathscr{R}_t$, *and there is one reaction node for each irreversible reaction in* $\mathscr{R}_t$.[2] *Each edge in the graph* $\Gamma_{\mathcal{N}}$ *connects a species node to a reaction node (so* $\Gamma_{\mathcal{N}}$ *is a bipartite graph) according to the following prescription: Consider a species node* s *and a reaction node* r *given by* $y \to y'$ *or* $y \rightleftharpoons y'$. *If* $s \in \operatorname{supp}(y)$, *then there is an edge between* s *and* r *and we* label *it with the complex* y. *Similarly, if* $s \in \operatorname{supp}(y')$, *then there is an edge*[3] *between* s *and* r *and we* label *it with the complex* y'.

For example, for the reaction network (2.1) there are three species nodes and two reaction nodes, and we get the SR graph in Figure 3.1.

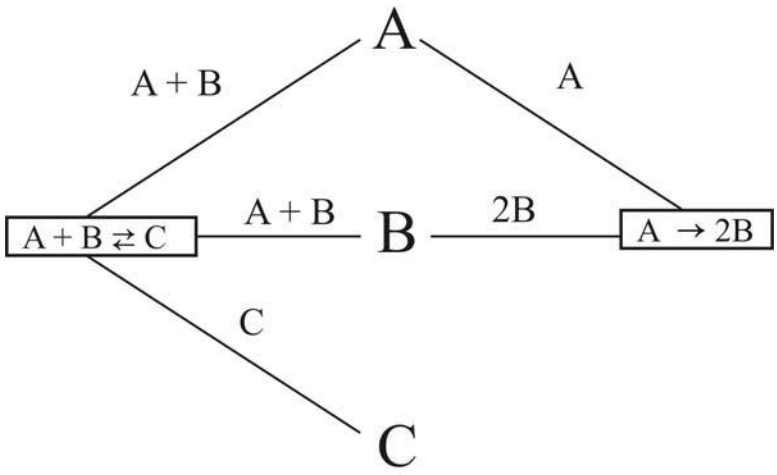


FIG. 3.1. *The SR graph of reaction network* (2.1).

In an SR graph there are some configurations of edges and cycles that are especially important to us. The following definition describes them.

DEFINITION 3.2. *Consider the SR graph* $\Gamma_{\mathcal{N}}$ *of some reaction network* $\mathcal{N}$. *A pair of edges in* $\Gamma_{\mathcal{N}}$ *that meet at a reaction node and have the same complex label is called a* c-pair. *A cycle that contains an odd number of c-pairs is called an* o-cycle. *A cycle that contains an even number of c-pairs is called an* e-cycle. *The* stoichiometric coefficient *of an edge is the coefficient of the species adjacent to that edge in the complex label of the edge. An* s-cycle *is one for which, if we alternately multiply and divide the stoichiometric coefficients of edges along the cycle, we get the final result* 1. *An* S-to-R chain *in an SR graph is a simple path from a species node*

[2]Recall that $\mathscr{R}_t$ is the set of true chemical reactions—that is, the set of reactions before the addition of reactions such as $s \to 0$ or $0 \to s$.

[3]If s is contained in both $\operatorname{supp}(y)$ and $\operatorname{supp}(y')$ (as in $A + B \to 2A$), then there are two edges joining the species node s to the reaction node $y \to y'$, one carrying the label y and the other carrying the label y'.

to a reaction node. We say that two cycles in $\Gamma_{\mathcal{N}}$ have an S-to-R intersection *if their common edges constitute an S-to-R chain or a disjoint union of two or more S-to-R chains.*

Recall that we gave another example of an SR graph in section 1.

4. The OSR graph. In this section we define the *oriented species-reaction graph (OSR graph)*, which will be the main tool for proving the results in the rest of this article. For this and the next section, we consider a fixed reaction network $\mathcal{N} = (\mathscr{S}, \mathscr{C}, \mathscr{R})$. Recall that any complex y in $\mathscr{C}$ is a linear combination $y = \sum_{s\in\mathscr{S}} y_s s$, where $y_s \geq 0$ for all $s \in \mathscr{S}$. Recall too that the *support* of y is defined by $\text{supp}(y) = \{s \in \mathscr{S} : y_s > 0\}$. In view of our interest in network injectivity (Definition 2.5), we consider a fixed ordered set of reactions[4] $R = \{y_1 \to y_1', \ldots, y_n \to y_n'\} \subset \mathscr{R}$, where n is the number of species for the network $\mathcal{N}$. *We seek to determine the conditions under which the product*

$$\det(y_1, \ldots, y_n) \det(y_1 - y_1', \ldots, y_n - y_n')$$

is positive when it is not zero. Hereafter, then, we assume that, for the ordered set of reactions R under consideration, the product above is not zero.

In this case the complexes $y_1, \ldots, y_n$ are linearly independent vectors. Then it is not difficult to see that one can make a bijective association between the n species of the network $\mathcal{N}$ and the n complexes $y_1, \ldots, y_n$, which associates with each complex a particular species in its support. In other words, there exists a (not necessarily unique) bijection $f : \{y_1, \ldots, y_n\} \to \mathscr{S}$ such that $f(y_i) \in \text{supp}(y_i)$, $i = 1, 2, \ldots, n$.

Hereafter, we choose one such bijection and denote by e_i the species $f(y_i)$. Thus, the set of species of the network $\mathcal{N}$ becomes $\{e_1, \ldots, e_n\}$, and we have $e_i \in \text{supp}(y_i)$, $i = 1, 2, \ldots, n$. For the sake of concreteness, we suppose that the determinant function is such that $\det(e_1, \ldots, e_n) > 0$. (In what follows, some readers might wish to associate $\{e_1, \ldots, e_n\}$ with the standard basis of $\mathbb{R}^n$, in which case the complexes $y_1, \ldots, y_n$ would be associated with vectors in $\mathbb{R}^n$.)

DEFINITION 4.1. *The* OSR graph of R *is an oriented graph* $\mathcal{G}_R$*, defined as follows. The set of nodes of* $\mathcal{G}_R$ *is* $\mathscr{S} \cup (R \cap \mathscr{R}_t)$*. The nodes in* $\mathscr{S}$ *are called* species nodes*, and the nodes in* $R \cap \mathscr{R}_t$ *are called* reaction nodes*. Each (oriented) edge in the graph* $\mathcal{G}_R$ *connects a species node to a reaction node or a reaction node to a species node in the following way. Consider some true reaction* $y_j \to y_j'$*. There is exactly one incoming edge toward the node* $y_j \to y_j'$ *in* $\mathcal{G}_R$*, and it comes from the node of the species* e_j*. We label this edge with the complex* y_j*. There is one outgoing edge from the reaction node* $y_j \to y_j'$ *toward each species node* $e_i \in \text{supp}(y_j)$*, except for* e_j*. We label these edges with the complex* y_j *as well. There is one outgoing edge from the reaction node* $y_j \to y_j'$ *toward each species node* $e_i \in \text{supp}(y_j')$*. We label these edges with the complex* y_j'*.*

For example, if in the reaction network (2.1) we choose the reactions that make up the set R to be $y_1 \to y_1' = A \to 2B$, $y_2 \to y_2' = A + B \to C$, $y_3 \to y_3' = C \to 0$ and we identify A, B, C with e_1, e_2, e_3, then we get the OSR graph in Figure 4.1.

Since the OSR graph is defined similarly to the SR graph, we can also refer to s-cycles, o-cycles, and e-cycles in the OSR graph, their definitions being analogous to those in the SR graph. However, whenever we mention a cycle in an OSR graph, that cycle *will have to be an oriented cycle.* In particular, the s-cycles, o-cycles, and e-cycles in an OSR graph have to be oriented cycles, and a c-pair has to be an oriented

[4]Some of these reactions might be outflow reactions.

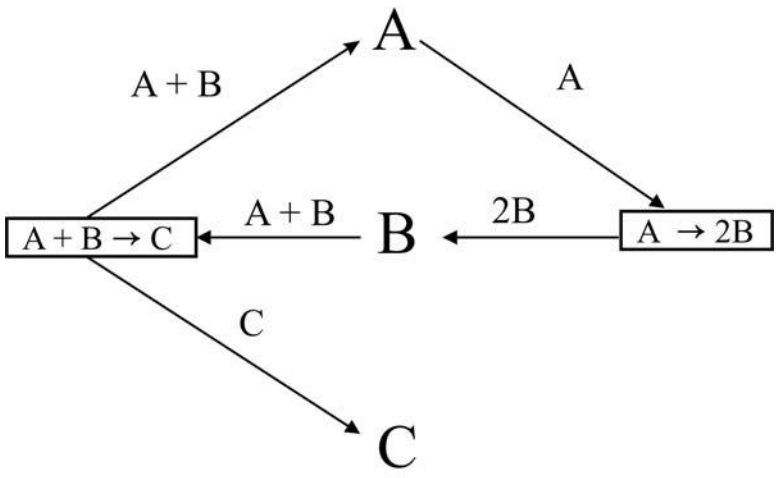


FIG. 4.1. *An OSR graph for some reactions in* (2.1).

pair of edges as well (i.e., one of the two adjacent edges that form a c-pair should point toward their common reaction vertex, and the other should point away from their common reaction vertex).

Remark. Note that each (oriented) edge in the OSR graph $\mathcal{G}_R$, connecting some species node and some reaction node, corresponds uniquely to some (unoriented) edge in the SR graph $\Gamma_{\mathcal{N}}$ of $\mathcal{N}$ connecting the same species node to the corresponding reaction node in $\Gamma_{\mathcal{N}}$, and has the same complex label. In other words, the OSR graph $\mathcal{G}_R$ is an (oriented) subgraph of the SR graph $\Gamma_{\mathcal{N}}$.

Remark. Suppose that R contains only outflow reactions, i.e., $R = \{A_1 \to 0, \ldots, A_n \to 0\}$. Then the OSR graph G_R has n species vertices, has *no reaction vertices*, and has *no edges*.

5. Properties of the OSR graph. To be able to formulate properties of the OSR graph we first need to introduce more definitions and notation.

Note that, for each $y_i \to y_i' \in R$, the complex y_i has a decomposition of the form

$$y_i = \sum_{e_k \in \mathrm{supp}(y_i)} y_{ik} e_k,$$

which defines numbers $y_{ik} > 0$. In particular, recall that $e_i \in \mathrm{supp}(y_i)$; i.e., we obtain $y_{ii} > 0$.

In view of Definition 2.5, we will now define a special multilinear expansion of $\det(y_1 - y_1', \ldots, y_n - y_n')$. For each $y_i \to y_i' \in R$ the vector $y_i - y_i'$ has a decomposition of the form

$$y_i - y_i' = \sum_{e_k \in \mathrm{supp}(y_i)} y_{ik} e_k - \sum_{e_k \in \mathrm{supp}(y_i')} y_{ik}' e_k,$$

where $y_{ik} > 0$ were mentioned above. We want now to consider the multilinear expansion of $\det(y_1 - y_1', \ldots, y_n - y_n')$ obtained by expanding each $y_i - y_i'$ in terms of the basis elements $e_1, \ldots, e_n$, with one exception: If $\mathrm{supp}(y_i) \cap \mathrm{supp}(y_i') \neq \emptyset$ for some i (as in a reaction of the form $A + B \to 2A$), we *do not* want to confuse

the contribution of y_i with the contribution of y'_i. For this reason, we represent the multilinear expansion of $\det(y_1 - y'_1, \ldots, y_n - y'_n)$ as the sum of all terms of the form

$$\det(\delta_{1k_1} e_{k_1}, \ldots, \delta_{nk_n} e_{k_n}),$$

where $e_{k_1} \in \mathrm{supp}(y_1) \cup \mathrm{supp}(y'_1), \ldots, e_{k_n} \in \mathrm{supp}(y_n) \cup \mathrm{supp}(y'_n)$, and

$$\delta_{ik_i} = \begin{cases} y_{ik_i} & \text{if } e_{k_i} \in \mathrm{supp}(y_i) \backslash \mathrm{supp}(y'_i), \\ -y'_{ik_i} & \text{if } e_{k_i} \in \mathrm{supp}(y'_i) \backslash \mathrm{supp}(y_i), \\ \text{either } y_{ik_i} \text{ or } -y'_{ik_i} & \text{if } e_{k_i} \in \mathrm{supp}(y_i) \cap \mathrm{supp}(y'_i). \end{cases}$$

DEFINITION 5.1. *By a* term in the expansion of the determinant $\det(y_1 - y'_1, \ldots, y_n - y'_n)$ *we mean a term in the multilinear expansion of* $\det(y_1 - y'_1, \ldots, y_n - y'_n)$ *described above.*

Note that a term might have the value zero. We will describe an important relationship between *nonzero* terms in the expansion of $\det(y_1 - y'_1, \ldots, y_n - y'_n)$ and the graph $\mathcal{G}_R$. Let us denote by Δ the term $\det(y_{11}e_1, \ldots, y_{nn}e_n)$. Of course, Δ is a (nonzero) term in the expansion of $\det(y_1 - y'_1, \ldots, y_n - y'_n)$ in the sense of the previous definition. Suppose that there is an edge ε in $\mathcal{G}_R$ from the reaction node $y_i \to y'_i$ to some species node e_k and having the complex label y_i. Let us denote by Δ_ε the result of replacing $y_{ii}e_i$ by $y_{ik}e_k$ in Δ and leaving everything else unchanged. Note that, according to the definition above, Δ_ε *is a (zero-valued) term in the expansion of* $\det(y_1 - y'_1, \ldots, y_n - y'_n)$. Similarly, suppose there is an edge ε' in $\mathcal{G}_R$ from the reaction node $y_i \to y'_i$ to some species node e_k and having the complex label y'_i. Let us denote by $\Delta_{\varepsilon'}$ the result of replacing $y_{ii}e_i$ by $-y'_{ik}e_k$ in Δ and leaving everything else unchanged. According to the definition above, $\Delta_{\varepsilon'}$ *is also a (zero-valued) term in the expansion of* $\det(y_1 - y'_1, \ldots, y_n - y'_n)$.

If L is a cycle in $\mathcal{G}_R$, let us denote by Δ_L the term resulting from making replacements in Δ as above, simultaneously for all edges in L that go from a reaction node to a species node. (See the example after the proof of the following lemma.) Then Δ_L is also a (nonzero) term in the expansion of $\det(y_1 - y'_1, \ldots, y_n - y'_n)$. If $\mathcal{L}$ is a set of disjoint cycles in $\mathcal{G}_R$, let us denote by $\Delta_\mathcal{L}$ the term resulting from making replacements in Δ as above, simultaneously for all edges in $\mathcal{L}$ that go from a reaction node to a species node. Then $\Delta_\mathcal{L}$ is also a (nonzero) term in the expansion of $\det(y_1 - y'_1, \ldots, y_n - y'_n)$. Lemma 5.1 will show that all nonzero terms in the expansion of $\det(y_1 - y'_1, \ldots, y_n - y'_n)$ are of the form $\Delta_\mathcal{L}$ for some set $\mathcal{L}$ of disjoint cycles in $\mathcal{G}_R$.

The case of $\det(y_1, \ldots, y_n)$ is similar, but simpler. We have

$$\det(y_1, \ldots, y_n) = \sum_{e_{k_1} \in \mathrm{supp}(y_1), \ldots, e_{k_n} \in \mathrm{supp}(y_n)} \det(y_{1k_1} e_{k_1}, \ldots, y_{nk_n} e_{k_n}),$$

and we state the following definition.

DEFINITION 5.2. *By a* term in the expansion of the determinant $\det(y_1, \ldots, y_n)$ *we mean a term in the multilinear expansion of* $\det(y_1, \ldots, y_n)$ *above.*

Note now that, according to the two definitions above, each term in the expansion of the determinant $\det(y_1, \ldots, y_n)$ is also a term in the expansion of the determinant $\det(y_1 - y'_1, \ldots, y_n - y'_n)$. Note also that Δ_ε defined as above is a term in the expansion of $\det(y_1, \ldots, y_n)$, since the complex label of ε is y_i, while $\Delta_{\varepsilon'}$ defined as above will not be a term in the expansion of $\det(y_1, \ldots, y_n)$, since the complex label of ε' is y'_i. Let us refer to edges similar to ε' as *product edges*. In other words, an edge ε' is a "product edge" if it is oriented from a reaction node $y_i \to y'_i$ to a species node, and

the complex label of the edge ε' is y'_i. Then, if a cycle L in $\mathcal{G}_R$ contains no product edges, Δ_L is also a (nonzero) term in the expansion of $\det(y_1, \dots, y_n)$. Similarly, if $\mathcal{L}$ is a set of disjoint cycles in $\mathcal{G}_R$ that contain no product edges, then $\Delta_{\mathcal{L}}$ is also a (nonzero) term in the expansion of $\det(y_1, \dots, y_n)$.

The following lemma associates with each nonzero term in the expansion of $\det(y_1 - y'_1, \dots, y_n - y'_n)$ a set of mutually disjoint cycles in the OSR graph $\mathcal{G}_R$, in a bijective way (also note the example after the proof).

LEMMA 5.1. *There is a bijective correspondence that associates with each nonzero term Δ_* in the expansion of $\det(y_1 - y'_1, \dots, y_n - y'_n)$ a set $\mathcal{L}$ of disjoint cycles in $\mathcal{G}_R$ such that $\Delta_* = \Delta_{\mathcal{L}}$. In particular, if $\mathbf{L}$ is the collection of all sets of mutually disjoint cycles in $\mathcal{G}_R$, we have*

$$\det(y_1 - y'_1, \dots, y_n - y'_n) = \sum_{\mathcal{L} \in \mathbf{L}} \Delta_{\mathcal{L}}.$$

Proof. Consider some nonzero term $\Delta_* = \det(\delta_{1k_1} e_{k_1}, \dots, \delta_{nk_n} e_{k_n})$ in the expansion of $\det(y_1 - y'_1, \dots, y_n - y'_n)$. Then $(k_1, k_2, \dots, k_n)$ is a permutation of the set $\{1, 2, \dots, n\}$. We denote this permutation by σ, i.e., $\sigma(i) = k_i$, $i = 1, \dots, na$. Recall that if $e_k \in \operatorname{supp}(y_i) \cup \operatorname{supp}(y'_i)$ and $i \neq k$, then there is an edge in $\mathcal{G}_R$ from the reaction node $y_i \to y'_i$ to the species node e_k. Also, recall that for any i there is an edge in $\mathcal{G}_R$ from the species node e_i to the reaction node $y_i \to y'_i$.

Suppose that the permutation σ has a cycle of length two, $\mathcal{C} = (ij)$, $i \neq j$. In this case $\delta_{ij} \neq 0$ and $\delta_{ji} \neq 0$. Then $e_j \in \operatorname{supp}(y_i) \cup \operatorname{supp}(y'_i)$ and $i \neq j$, so there is an edge in $\mathcal{G}_R$ from the reaction node $y_i \to y'_i$ to the species node e_j. Also, $e_i \in \operatorname{supp}(y_j) \cup \operatorname{supp}(y'_j)$ and $j \neq i$, so there is an edge in $\mathcal{G}_R$ from the reaction node $y_j \to y'_j$ to the species node e_i. These two edges together with the edge from e_i to $y_i \to y'_i$ and the edge from e_j to $y_j \to y'_j$ form an (oriented) cycle $L_{\mathcal{C}}$ of length four in $\mathcal{G}_R$. Also, note that $\Delta_{L_{\mathcal{C}}}$ is the same as Δ_* at its ith and jth entries. Similarly, with any other cycle $\mathcal{C}$ of σ of length k we associate an (oriented) cycle of length $2k$ in $\mathcal{G}_R$.

Then it is not difficult to see that we have $\Delta_* = \Delta_{\mathcal{L}}$, where $\mathcal{L}$ is the set of all cycles $L_{\mathcal{C}}$ with $\mathcal{C}$ a cycle of σ.

Finally, note that if we begin from some set $\mathcal{L}$ of disjoint cycles in $\mathcal{G}_R$, construct the term $\Delta_{\mathcal{L}}$, and then construct a set $\tilde{\mathcal{L}}$ of disjoint cycles in $\mathcal{G}_R$ from the term $\Delta_{\mathcal{L}}$, as described above, then $\mathcal{L} = \tilde{\mathcal{L}}$. This shows that the correspondence described above is bijective. ∎

Example. Consider the ordered set of five reactions

$$2A \to B, \quad A + B \to C, \quad C + D \to B + E, \quad D \to 2C, \quad E \to 0. \tag{5.1}$$

We identify the species sequence A, B, C, D, E with e_1, e_2, e_3, e_4, e_5. The corresponding OSR graph appears in Figure 5.1.

There are three oriented cycles: l_1, which passes through the species nodes A and B; l_2, which passes through the species nodes B and C; and l_3, which passes through the species nodes C and D.

We have $\det(y_1 - y'_1, \dots, y_5 - y'_5) = \det(2e_1 - e_2, e_1 + e_2 - e_3, e_3 + e_4 - e_2 - e_5, e_4 - 2e_3, e_5)$. The term $\det(y_{11}e_1, \dots, y_{nn}e_n)$ in the multilinear expansion of $\det(y_1 - y'_1, \dots, y_5 - y'_5)$ is in this case $\det(2e_1, e_2, e_3, e_4, e_5)$. This term equals $\Delta_\emptyset$; i.e., it corresponds to the set $\mathcal{L}$ of disjoint cycles being the empty set.

We will now check that there is a one-to-one correspondence between all other nonzero terms in the multilinear expansion of $\det(y_1 - y'_1, \dots, y_5 - y'_5)$ and nonempty sets of disjoint cycles in the OSR graph.

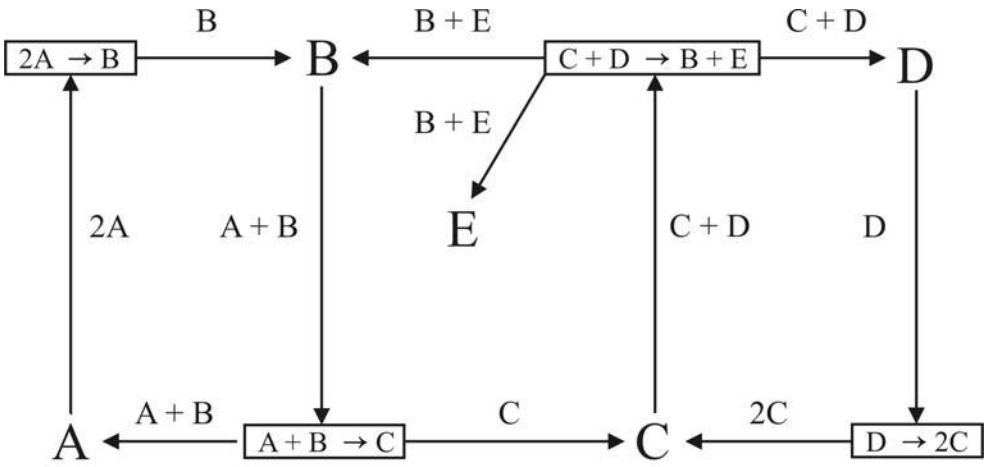


FIG. 5.1. *An OSR graph for the set of reactions* (5.1).

The cycle l_1 corresponds to replacing $2e_1$ in $\det(2e_1, e_2, e_3, e_4, e_5)$ by $-e_2$, and replacing e_2 in $\det(2e_1, e_2, e_3, e_4, e_5)$ by e_1, since the cycle l_1 visits the species node B after leaving A and visits the species node A after leaving B. The corresponding term in the multilinear expansion of $\det(y_1 - y_1', \ldots, y_5 - y_5')$ is therefore $\det(-e_2, e_1, e_3, e_4, e_5)$. Similarly, the cycle l_2 corresponds to replacing e_2 in $\det(2e_1, e_2, e_3, e_4, e_5)$ by $-e_3$, and replacing e_3 in $\det(2e_1, e_2, e_3, e_4, e_5)$ by $-e_2$, since the cycle l_2 visits the species node C after leaving B and visits the species node B after leaving C. The corresponding term in the multilinear expansion of $\det(y_1 - y_1', \ldots, y_5 - y_5')$ is therefore $\det(2e_1, -e_3, -e_2, e_4, e_5)$. Similarly, the cycle l_3 corresponds to the term $\det(2e_1, e_2, e_4, -2e_3, e_5)$. There is one more nonzero term in the expansion of $\det(y_1 - y_1', \ldots, y_5 - y_5')$. This term is $\det(-e_2, e_1, e_4, -2e_3, e_5)$, and it corresponds to the set $\{l_1, l_3\}$ of disjoint cycles.

To formulate an analogous lemma for $\det(y_1, \ldots, y_n)$ let us denote by $\mathbf{L}_{np}$ the collection of all sets of mutually disjoint cycles in $\mathcal{G}_R$ that have no product edges. Then we have the following result.

LEMMA 5.2. *There is a bijective correspondence that associates with each nonzero term* Δ_* *in the expansion of* $\det(y_1, \ldots, y_n)$ *a set* $\mathcal{L} \in \mathbf{L}_{np}$ *such that* $\Delta_* = \Delta_{\mathcal{L}_{np}}$. *In particular, we have*

$$\det(y_1, \ldots, y_n) = \sum_{\mathcal{L} \in \mathbf{L}_{np}} \Delta_{\mathcal{L}}.$$

Proof. The proof here is analogous to that of the previous lemma. ∎

Let us now look more closely at the connection between the SR graph and the OSR graph, as follows.

LEMMA 5.3. *If two (oriented) cycles* l_1 *and* l_2 *in* $\mathcal{G}_R$ *have a common vertex, then their (unoriented) versions* l_1^{SR} *and* l_2^{SR} *in* $\Gamma_{\mathcal{N}}$ *have an S-to-R intersection.*

Proof. Suppose that l_1 and l_2 have a species node s in common. Since they are oriented cycles, each one of them has to contain an outgoing edge from s. However, there is a unique outgoing edge adjacent to s in $\mathcal{G}_R$. Therefore that edge is common to the two cycles, and the corresponding edge in $\Gamma_{\mathcal{N}}$ is common to l_1^{SR} and l_2^{SR}. Analogously, if l_1 and l_2 have a reaction node r in common, each one of them has to

contain the unique incoming edge adjacent to r in $\mathcal{G}_R$. This shows that l_1^{SR} and l_2^{SR} have at least one edge in common.

Suppose now that we travel along the two cycles l_1 and l_2 in $\mathcal{G}_R$, beginning from some common edge and following the orientation of that edge. The first node where the two cycles separate from each other has to be a reaction node, since all species nodes have just one outgoing edge in $\mathcal{G}_R$. On the other hand, if we travel along the two cycles l_1 and l_2 in $\mathcal{G}_R$, beginning from some common edge, in the direction opposite to the orientation of that edge, then the first node where the two cycles separate from each other has to be a species node, since all reaction nodes have just one incoming edge in $\mathcal{G}_R$. In conclusion, the common edges of l_1^{SR} and l_2^{SR} form one or more S-to-R chains (see Definition 3.2); i.e., l_1^{SR} and l_2^{SR} have an S-to-R intersection. □

Before we can prove our main result we have to prove a few lemmas about special types of cycles in OSR graphs.

LEMMA 5.4. *Consider a set $\mathcal{L}$ of disjoint o-cycles in $\mathcal{G}_R$. Then $\Delta_{\mathcal{L}} > 0$.*

Proof. If $\mathcal{L} = \emptyset$, we have $\Delta_\emptyset = \det(y_{11}e_1, \dots, y_{nn}e_n) > 0$. Consider now the case when $\mathcal{L}$ contains exactly one cycle l. Denote by $e_{i_1}, e_{i_2}, \dots, e_{i_k}, e_{i_1}$ (in this order) the species vertices visited by the oriented cycle l. Then the term $\Delta_{\{l\}}$ is the same as $\det(y_{11}e_1, \dots, y_{nn}e_n)$ *except* that the entry $y_{i_1 i_1}e_{i_1}$ in $\det(y_{11}e_1, \dots, y_{nn}e_n)$ is replaced by $\delta_{i_1 i_2}e_{i_2}$, the entry $y_{i_2 i_2}e_{i_2}$ is replaced by $\delta_{i_2 i_3}e_{i_3}$, and so on, until the entry $y_{i_k i_k}e_{i_k}$ is replaced by $\delta_{i_k i_1}e_{i_1}$.

Since all $y_{ii} > 0$ it follows that the sign of $\Delta_{\{l\}}$ equals the sign of the cyclic permutation $(i_1 i_2 \dots i_k)$ times the sign of the product $\delta_{i_1 i_2}\delta_{i_2 i_3} \dots \delta_{i_k i_1}$. According to the standard decomposition of a cyclic permutation into a product of transpositions[5] the sign of the cyclic permutation $(i_1 i_2 \dots i_k)$ is $(-1)^{k-1}$. On the other hand, note that if the edge of l from the reaction node $y_{i_j} \to y'_{i_j}$ to the species node $e_{i_{j+1}}$ has the complex label y_{i_j}, then $\delta_{i_j i_{j+1}}$ is positive, and if the edge of l from the reaction node $y_{i_j} \to y'_{i_j}$ to the species node $e_{i_{j+1}}$ has the complex label y'_{i_j}, then $\delta_{i_j i_{j+1}}$ is negative.[6] Recall that the other edge of l adjacent to $y_{i_j} \to y'_{i_j}$ has to have the complex label y_{i_j}. In conclusion, the number of *positive* $\delta_{i_j i_{j+1}}$'s equals the number of *c-pairs* along l. However, according to the hypothesis, l has an odd number of c-pairs, say $2p+1$. Then the sign of the product $\delta_{i_1 i_2}\delta_{i_2 i_3} \dots \delta_{i_k i_1}$ is $(-1)^{k-2p-1}$, which implies that the sign of $\Delta_{\{l\}}$ is $(-1)^{k-1}(-1)^{k-2p-1}$, i.e., $\Delta_{\{l\}} > 0$.

For arbitrary $\mathcal{L}$ let us notice that since the cycles in $\mathcal{L}$ are mutually disjoint it follows that $\Delta_{\mathcal{L}}$ can be written as a product of determinants, one for each cycle in $\mathcal{L}$, and the considerations above apply for each one of these determinants. Then $\Delta_{\mathcal{L}} > 0$. □

LEMMA 5.5. *Suppose that l is an e-cycle and an s-cycle in $\mathcal{G}_R$, and $\mathcal{L}$ is a set of cycles in $\mathcal{G}_R$ that are disjoint from each other and disjoint from l. Then $\Delta_{\mathcal{L}} + \Delta_{\mathcal{L}\cup\{l\}} = 0$.*

Proof. Case 1. Suppose that the set $\mathcal{L}$ is empty. We have $\Delta_\emptyset = \det(y_{11}e_1, \dots, y_{nn}e_n)$. As in the proof of the previous lemma, the term $\Delta_{\{l\}}$ is the same as $\det(y_{11}e_1, \dots, y_{nn}e_n)$ *except* that the entry $y_{i_1 i_1}e_{i_1}$ in $\det(y_{11}e_1, \dots, y_{nn}e_n)$ is replaced by $\delta_{i_1 i_2}e_{i_2}$, the entry $y_{i_2 i_2}e_{i_2}$ is replaced by $\delta_{i_2 i_3}e_{i_3}$, and so on, until the entry $y_{i_k i_k}e_{i_k}$ is replaced by $\delta_{i_k i_1}e_{i_1}$. Note that the stoichiometric coefficient of the edge of l from the species node e_{i_j} to the reaction node $y_{i_j} \to y'_{i_j}$ is $y_{i_j i_j}$, and the stoichiometric coefficient of the edge of l from the reaction node $y_{i_j} \to y'_{i_j}$ to the species node $e_{i_{j+1}}$ is $\delta_{i_j i_{j+1}}$. Therefore if we alternately multiply and divide the stoichiometric coefficients

[5] I.e., the decomposition $(i_1 i_2 \dots i_k) = (i_1 i_2)(i_2 i_3) \dots (i_{k-1} i_k)$.

[6] Here we are using cyclic notation: By $\delta_{i_k i_{k+1}}$ we mean $\delta_{i_k i_1}$, and by $e_{i_{k+1}}$ we mean e_{i_1}.

of the edges along the cycle l, we get

$$(y_{i_1 i_1}/\delta_{i_1 i_2})(y_{i_2 i_2}/\delta_{i_2 i_3})\dots(y_{i_{k-1} i_{k-1}}/\delta_{i_{k-1} i_k})(y_{i_k i_k}/\delta_{i_k i_1}).$$

Since l is an s-cycle the product above equals 1, and we obtain $y_{i_1} y_{i_2} \dots y_{i_k} = \delta_{i_1}\delta_{i_2}\dots\delta_{i_k}$. Then the absolute value of $\Delta_{\{l\}}$ is the same as the absolute value of $\det(y_{11}e_1, \dots, y_{nn}e_n)$. Since l is an e-cycle we reason as in the proof of the previous lemma to conclude that $\Delta_{\{l\}}$ is negative. Then

$$\Delta_\emptyset + \Delta_{\{l\}} = \det(y_{11}e_1, \dots, y_{nn}e_n) + \Delta_{\{l\}} = 0.$$

Case 2. Suppose that $\mathcal{L}$ contains at least one cycle. Since l is disjoint from all cycles in $\mathcal{L}$, we can argue exactly as in Case 1 that $\Delta_\mathcal{L}$ and $\Delta_{\mathcal{L}\cup\{l\}}$ have the same absolute value and different signs. □

6. The main result. We can now prove the following theorem.

THEOREM 6.1. *Consider some reaction network $\mathcal{N}$ such that in its SR graph $\Gamma_\mathcal{N}$ all cycles are o-cycles or s-cycles, and no two e-cycles have an S-to-R intersection. Then the reaction network $\mathcal{N}$ is injective.*

Proof. Consider some set $R = \{y_1 \to y_1', \dots, y_n \to y_n'\}$ of n reactions in $\mathcal{N}$ such that $\det(y_1, \dots, y_n)\det(y_1 - y_1', \dots, y_n - y_n') \neq 0$. We want to show that

$$\det(y_1, \dots, y_n)\det(y_1 - y_1', \dots, y_n - y_n') > 0.$$

By reordering the basis vectors e_i, we can suppose that $e_i \in \mathrm{supp}(y_i)$; i.e., the set R obeys the conditions imposed in the previous sections. (We are using here the "standard" determinant, for which $\det(e_1, \dots, e_n) > 0$.)

We will show first that $\det(y_1 - y_1', \dots, y_n - y_n') > 0$. Recall from Lemma 5.1 that the determinant can be calculated as a sum of terms, one for each member of the class $\mathbf{L}$ of all possible sets (including the empty set) of disjoint cycles taken from the OSR graph $\mathcal{G}_R$:

$$\det(y_1 - y_1', \dots, y_n - y_n') = \sum_{\mathcal{L}\in\mathbf{L}} \Delta_\mathcal{L}. \tag{6.1}$$

Let $\{\mathcal{O}_1, \dots, \mathcal{O}_p\}$ be the collection of all possible sets (including the empty set) of disjoint o-cycles that can be taken from $\mathcal{G}_R$. We partition the class $\mathbf{L}$ of all possible sets of disjoint cycles into p subclasses $\{\mathbf{L}_1, \mathbf{L}_2, \dots, \mathbf{L}_p\}$, according to the particular subset of o-cycles that each cycle-set in $\mathbf{L}$ contains. That is, $\mathbf{L}_i$ might contain several sets of cycles, but for each set of cycles in $\mathbf{L}_i$ the subset of o-cycles is precisely $\mathcal{O}_i$, $i = 1, \dots, p$. In light of this partition, (6.1) can be rewritten as

$$\det(y_1 - y_1', \dots, y_n - y_n') = \sum_{i=1}^{p} \sum_{\mathcal{L}\in\mathbf{L}_i} \Delta_\mathcal{L}. \tag{6.2}$$

To show that the (presumed nonzero) $\det(y_1 - y_1', \dots, y_n - y_n')$ is in fact positive, it will suffice to show that $\sum_{\mathcal{L}\in\mathbf{L}_i} \Delta_\mathcal{L} \geq 0$ for all $i = 1, \dots, p$.

With this in mind, we consider a particular collection $\mathbf{L}_i$ of disjoint cycle-sets, with $\mathcal{O}_i$ the common subset of o-cycles for each cycle-set in $\mathbf{L}_i$. If no cycle-set in $\mathbf{L}_i$ contains an e-cycle, then, by virtue of Lemma 5.4, $\Delta_\mathcal{L} > 0$ for every $L \in \mathbf{L}_i$. (This is true even if $\mathcal{O}_i$ is empty.) Thus, it remains to consider only the case for which at least one cycle-set in $\mathbf{L}_i$ contains an e-cycle (which, by hypothesis, must be an

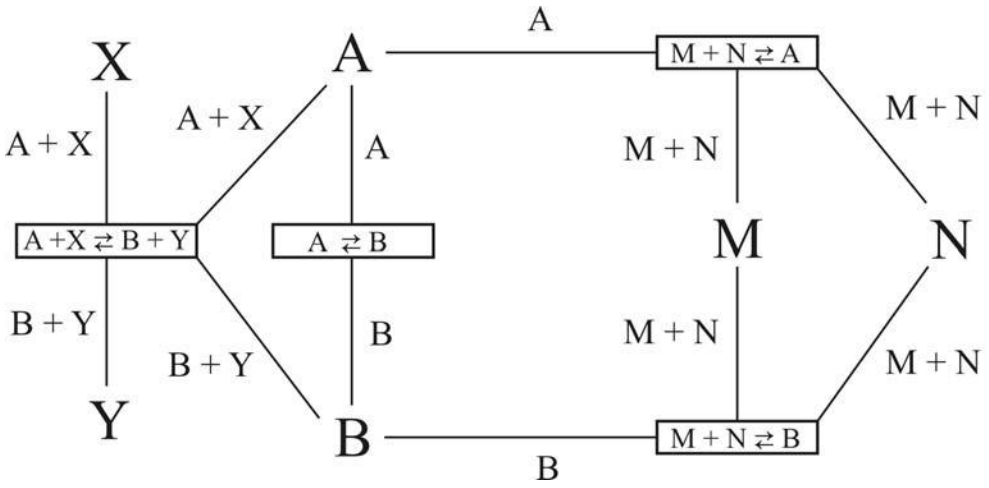


FIG. 6.1. *SR graph for the reaction network* (6.4).

s-cycle). Let l_e be some fixed e-cycle residing in a cycle-set of $\mathbf{L}_i$. Since l_e is disjoint from all members of $\mathcal{O}_i$ and from every other e-cycle (by virtue of the hypothesis and Lemma 5.3), it follows that, for each cycle-set $\mathcal{L} \in \mathbf{L}_i$ that does not have l_e as a member, the cycle-set $\mathcal{L} \cup \{l_e\}$ also belongs to the family $\mathbf{L}_i$. Note that from Lemma 5.5 we have

$$\Delta_{\mathcal{L}} + \Delta_{\mathcal{L}\cup\{l_e\}} = 0. \tag{6.3}$$

By partitioning $\mathbf{L}_i$ into such cycle-set pairs—one member distinguished from the other only by the presence of l_e—we can deduce in this case that $\sum_{\mathcal{L}\in\mathbf{L}_i} \Delta_{\mathcal{L}} = 0$.

The proof that $\det(y_1, \ldots, y_n) > 0$ is virtually identical, except that we consider only cycles containing no product edges. □

Remark. We will say that two cycles have an *orientable S-to-R intersection* if the two cycles have an S-to-R intersection and also have the following additional property: There are directions along the two cycles, consistent on their intersection, such that, for each S-to-R connected component of the intersection of the two cycles, its species end node occurs before its reaction end node. Note then that in Lemma 5.3 one can replace "S-to-R intersection" by "orientable S-to-R intersection." It is then possible to strengthen Theorem 6.1 by replacing "S-to-R intersection" by "orientable S-to-R intersection."

The following example shows a reaction network for which Theorem 6.1 gives no information, but for which this strengthened version of Theorem 6.1 does give information. Consider the reaction network

$$A + X \rightleftharpoons B + Y, \quad A \rightleftharpoons B, \quad M + N \rightleftharpoons A, \quad M + N \rightleftharpoons B. \tag{6.4}$$

The SR graph of this reaction network is shown in Figure 6.1. The middle cycle and the outer cycle are e-cycles that have an S-to-R intersection, but they do not have an orientable S-to-R intersection.

7. Split c-pairs. A different approach to showing that a reaction network does not have the capacity for multiple equilibria was described in [8, 15, 16] and is based on a different graph associated with the reaction network, called the *SCL graph*. That

approach introduced the notion of a *split c-pair*. The same notion of *split c-pair* makes sense for SR graphs as well: We say that two cycles in an SR graph *split a c-pair* if each edge of the c-pair appears in at least one of the two cycles, and if one of the two cycles contains one edge of the c-pair but not the other edge. (The other cycle might contain just the other edge, or both.) We have the following claim.

LEMMA 7.1. *Consider some reaction network $\mathcal{N}$ and its SR graph $\Gamma_{\mathcal{N}}$. Suppose that there are two cycles l_1 and l_2 in $\Gamma_{\mathcal{N}}$ that have an S-to-R intersection. Then l_1 and l_2 split a c-pair.*

Proof. Denote by r the reaction node ending of a component of the intersection of l_1 and l_2. Note that there are exactly three edges of l_1 and l_2 adjacent to the node r, precisely one of which is common to both cycles. Then at least two of these three edges have the same complex label, because there are at most two different complex labels on all edges adjacent to r. These two edges that have the same complex label (say ε_1 and ε_2) form a c-pair. It is not possible that each one of l_1 and l_2 contains both ε_1 and ε_2, since r is the ending of a component of the intersection of l_1 and l_2. On the other hand, each one of l_1 and l_2 has to contain at least one of ε_1 and ε_2, because, of their three edges adjacent to r, only one edge is common to both ε_1 and ε_2. Therefore l_1 and l_2 split a c-pair. □

Then our main result implies a criterion based on split c-pairs, but for the SR graph instead of the SCL graph, as follows.

COROLLARY 7.2. *Consider some reaction network $\mathcal{N}$ such that in its SR graph $\Gamma_{\mathcal{N}}$ all cycles are o-cycles or s-cycles and no two e-cycles split a c-pair. Then the reaction network $\mathcal{N}$ is injective.*

8. Concluding remarks and implications for biology. Theorem 1.1 provides rather easily satisfied conditions for the preclusion of multiple equilibria based on reaction network structure alone. The first condition will be satisfied in the very common situation for which every nonzero stoichiometric coefficient is 1 (in which case every cycle in the SR graph is an s-cycle). Further, violation of the second condition requires not only that there be two cycles in the SR graph but also that there be two even cycles that intersect in a prescribed way. Indeed, Theorem 1.1 goes a long way toward explaining just why, despite the great variety of reaction networks that might arise in nature, there are so few experimental reports in the chemical engineering literature of multiple stationary states in an isothermal homogeneous CFSTR context.

At the same time, we believe that the theorem provides reasons to believe that enzyme-driven biochemical reaction networks, written at the mechanistic mass action level, might be far more prone than others to exhibit multiple equilibria (and particularly bistability [2, 13, 18]). The fact is that enzyme catalysis promotes the presence of cycles in the SR graph, as might be seen by constructing the SR graph for even the simplest possible mechanism of enzyme catalysis:

$$S + E \rightleftharpoons SE \to P + E.$$

Here, E is an enzyme, S is a substrate, P is a product, and SE represents S bound to the enzyme. (The enzyme E serves as a catalyst for the "overall reaction" $S \to P$.) For more intricate enzyme-catalyzed reaction networks, written at the mechanistic level, it is easy to see how an abundance of cycles in the SR graph might arise (so that the second condition of Theorem 1.1 becomes more likely to be violated).

Extensions to biology are somewhat more complicated than might first appear, for the classical CFSTR model as described in this article might not be entirely ap-

propriate in biological settings, not even as a crude metaphor. Even if we think of the stirred reactor vessel as a surrogate for a cell and even if we imagine that substrates and products (S and P in the example above) are transported readily through the cell membrane, it might be inappropriate to suppose that high molecular weight enzyme-related molecules (E and ES in the example) are also transported through the cell membrane. That is, the heavy enzyme-related species might be regarded as "entrapped" within the cell. For the entrapped species picture, the classical homogeneous CFSTR equations, which presume an outflow of all species, might not always be suitable. (Note that this presumption played a substantive role in proofs contained in this paper and in its predecessor [4].)

In some cases, it might be appropriate to imagine that enzymes are synthesized within the cell at constant rate (i.e., constant relative to the rapid time scale of other reactions) and that all enzyme-containing species degrade within the cell at rates proportional to their concentrations. In such cases, the mathematics becomes essentially identical to the mathematics of the classical CFSTR: Constant-rate enzyme synthesis plays the role of a constant enzyme feed rate to the cell, while the degradation of enzyme-containing species replaces the outflow of these species from the cell.

In other cases, when such suppositions of enzyme supply and degradation are deemed inappropriate, the resulting mathematical structure is similar but not identical to that studied in this article; in particular, there are no outflow reactions, such as $E \to 0$ for the enzymatic species. It happens that the absence of these outflows gives rise to surprisingly delicate mathematical questions when one tries to extend the results of Theorem 1.1 to entrapped enzyme models. Indeed, *one must reframe the very question of multiple equilibria to take into account the fact that one is interested only in equilibria consistent with a fixed enzyme supply.*

Nevertheless, there is a sense in which results in this paper and its predecessor [4] carry over to the entrapped species case. Even when the kinetics is not mass action, it can be shown that if a reaction network does not have the capacity for multiple equilibria when all species are in the outflow, then, in the entrapped species case, the network cannot give rise to multiple equilibria that are, in a certain sense, nondegenerate [5].

REFERENCES

[1] R. Aris, *Elementary Chemical Reactor Analysis*, Dover Publications, Mineola, NY, 2000.

[2] C. P. Bagowski and J. E. Ferrell, *Bistability in the JNK cascade*, Curr. Biol., 11 (2001), pp. 1176–1182.

[3] G. Craciun, *Systems of Nonlinear Equations Deriving from Complex Chemical Reaction Networks*, Ph.D. thesis, Department of Mathematics, The Ohio State University, Columbus, OH, 2002.

[4] G. Craciun and M. Feinberg, *Multiple equilibria in complex chemical reaction networks:* I. *The injectivity property*, SIAM J. Appl. Math., 65 (2005), pp. 1526–1546.

[5] G. Craciun and M. Feinberg, *Multiple equilibria in complex chemical reaction networks: Extensions to entrapped species models*, IEE Proc. Systems Biol., to appear.

[6] M. Feinberg, *Lectures on Chemical Reaction Networks*, notes of lectures given at the Mathematical Research Center, University of Wisconsin, Madison, WI, 1979; available online at www.chbmeng.ohio-state.edu/~feinberg/LecturesOnReactionNetworks.

[7] M. Feinberg, *Existence and uniqueness of steady states for a class of chemical reaction networks*, Arch. Ration. Mech. Anal., 132 (1995), pp. 311–370.

[8] M. Feinberg, *Some recent results in chemical reaction network theory*, in Patterns and Dynamics in Reactive Media, IMA Vol. Math. Appl. 37, R. Aris, D. G. Aronson, and H. L. Swinney, eds., Springer, Berlin, 1991, pp. 43–70.

[9] V. Hatzimanikatis, K. H. Lee, and J. E. Bailey, *A mathematical description of regulation of the G1-S transition of the mammalian cell cycle*, Biotech. Bioengineering, 65 (1999), pp. 631–637.

[10] F. Horn and R. Jackson, *General mass action kinetics*, Arch. Ration. Mech. Anal., 47 (1972), pp. 81–116.

[11] *Law of mass action*, in Britannica Concise Encyclopedia, Encyclopedia Britannica, 2004, http://concise.britannica.com/ebc/article?eu=396792.

[12] E. Lee, A. Salic, R. Kruger, R. Heinrich, and M. W. Kirschner, *The roles of APC and axin derived from experimental and theoretical analysis of the Wnt pathway*, PLoS Biol., 1 (2003), pp. 116–132.

[13] N. I. Markevich, J. B. Hoek, and B. N. Kholodenko, *Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades*, J. Cell Biol., 164 (2004), pp. 353–359.

[14] B. Novak and J. J. Tyson, *Modeling the control of DNA replication in fission yeast*, Proc. Natl. Acad. Sci. USA, 94 (1997), pp. 9147–9152.

[15] P. Schlosser and M. Feinberg, *A theory of multiple steady states in isothermal homogeneous CFSTRs with many reactions*, Chemical Engineering Sci., 49 (1994), pp. 1749–1767.

[16] P. Schlosser, *A Graphical Determination of the Possibility of Multiple Steady States in Complex Isothermal CFSTRs*, Ph.D. thesis, Department of Chemical Engineering, University of Rochester, Rochester, NY, 1988.

[17] J. J. Tyson, K. C. Chen, and B. Novak, *Sniffers, buzzers, toggles and blinkers: Dynamics of regulatory and signaling pathways in the cell*, Curr. Opin. Cell Biol., 15 (2003), pp. 221–231.

[18] W. Xiong and J. E. Ferrell, *A positive-feedback-based bistable "memory module" that governs a cell fate decision*, Nature, 426 (2003), pp. 460–465.

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1339–1365

MULTIPARAMETRIC BIFURCATION ANALYSIS OF A BASIC TWO-STAGE POPULATION MODEL*

S. M. BAER†, B. W. KOOI‡, YU. A. KUZNETSOV§, AND H. R. THIEME¶

Abstract. In this paper we investigate long-term dynamics of the most basic model for stage-structured populations, in which the per capita transition from the juvenile into the adult class is density dependent. The model is represented by an autonomous system of two nonlinear differential equations with four parameters for a single population. We find that the interaction of intra-adult competition and intra-juvenile competition gives rise to multiple attractors, one of which can be oscillatory. A detailed numerical study reveals a rich bifurcation structure for this two-dimensional system, originating from a degenerate Bogdanov–Takens (BT) bifurcation point when one parameter is kept constant. Depending on the value of this fixed parameter, the corresponding triple critical equilibrium has either an elliptic sector or it is a topological focus, which is demonstrated by the numerical normal form analysis. It is shown that the canonical unfolding of the codimension-three BT point reveals the underlying dynamics of the model. Certain new features of this unfolding in the elliptic case, which are important in applications but have been overlooked in available theoretical studies, are established. Various three-, two-, and one-parameter bifurcation diagrams of the model are presented and interpreted in biological terms.

Key words. bifurcation analysis, Bogdanov–Takens codimension-three point, elliptic sector, homoclinic orbits to saddle, saddle-node, and neutral saddle, two-stage population model

AMS subject classifications. 34C23, 92D25, 37G05

DOI. 10.1137/050627757

1. Introduction.

Population growth models that include age, stage or body size structure often predict complex population dynamics. The models are rather sophisticated, involving partial or functional differential equations, difference equations, or integral equations [3, 6, 7, 35]. In this paper, we investigate a simple-stage structured model governed by a two-dimensional system of time-autonomous ordinary differential equations. The equations represent the juvenile and adult stage, respectively. We show that this reduced but biologically-based model predicts, qualitatively, complex population dynamics. Of course, the complexity is restricted by the Poincaré–Bendixson theory. Earlier work addressing competition between age stages [21] found that periodic orbits can surround a unique interior equilibrium. Here we show that multiple equilibria are possible, both stable and unstable periodic orbits can exist and even coexist, and homoclinic orbits can occur through the interaction of periodic orbits and multiple equilibria.

The best-known two-dimensional ODE system in population biology is the Lotka–Volterra predator/prey model where the dynamic behavior is simple but structurally

*Received by the editors March 26, 2005; accepted for publication (in revised form) November 15, 2005; published electronically April 21, 2006.
http://www.siam.org/journals/siap/66-4/62775.html

†Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287 (baer@math.la.asu.edu).

‡Department of Theoretical Biology, Vrije Universiteit, de Boelelaan 1087, 1081 HV Amsterdam, the Netherlands (kooi@bio.vu.nl).

§Mathematical Institute, Utrecht Universiteit, Budapestlaan 6, 3584 CD Utrecht, the Netherlands (kuznetsov@math.uu.nl).

¶Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287 (thieme@math.la.asu.edu). This author was partially supported by NSF grants DMS-9706787 and 0314529.

unstable. In an extension of this model, called the Rosenzweig–MacArthur model, the trophic interaction is described by a hyperbolic functional response instead of a linear functional response. If a carrying capacity for the prey is added, stable oscillations can occur, but no more complex dynamics. An extension of the Rosenzweig–MacArthur model was proposed and studied in [1], where the per capita mortality rate of the predators is replaced by a density-dependent mortality rate, and in [40] where the Holling type II functional response is replaced by a nonmonotonic Holling type IV functional response for the predator-prey interaction. These planar systems show rich asymptotic dynamic behavior including global bifurcations and a variety of codimension-two points originating from codimension-three bifurcation points.

We deal with a model for one population with two stages introduced in [21] and further studied in [34, Chapter 11]. The life history of the individuals comprises a juvenile and an adult stage. The population state is the number of juveniles and adults. Juveniles and adults die, while adults reproduce. The transition rate between the two stages, together with the per capita mortality and reproduction rates, form the parameters of the two first-order ordinary differential equations (ODEs) which specify the time derivative of the population state.

We use bifurcation analysis to study the dependence of the long-term dynamic behavior of this system on parameter variation. See [16, 36, 32, 22] for an introduction to bifurcation analysis and [1] for applications to ecosystem models. The structural stability is studied with respect to so-called free or bifurcation parameters. Parameter values at which the asymptotic dynamics change are called bifurcation points; for example, a Hopf bifurcation point represents a transition between constant and time-periodic solutions. Starting at this point one can vary two parameters simultaneously. This is sometimes called a two-parameter bifurcation study. The resulting bifurcation curves separate regions in the two-dimensional parameter space, which differ in qualitative asymptotic behavior. At so-called codimension-two points three or more parameter regions come together. In other words, different types of bifurcation points can originate from or terminate at these points, making these points useful for starting a numerical bifurcation analysis. Codimension-two points can be followed by varying three parameters simultaneously, and so on.

Our stage-structured population model is a two-dimensional system with four parameters. When one parameter is fixed, a degenerate Bogdanov–Takens bifurcation (BT point for short) with a triple critical equilibrium is found. Different types of such degenerate BT points are studied in detail in [2, 12], where three categories of topological types are distinguished: the saddle, the focus, and the elliptic case. We will derive a normal form with terms up to and including fourth order for the codimension-three BT point by using a time reparameterization combined with a smooth transformation that also includes fourth-order terms. This analysis reveals that the elliptic or focus case applies for the two-stage population model depending on the value of the fixed parameter. When this parameter is varied a transition from elliptic to focus codimension-three BT bifurcation is found.

The truncated normal form of the codimension-three BT point in the elliptic case is embedded into an appropriate three-parameter family, and we study its unfolding by performing a numerical bifurcation analysis for the neighborhood of the origin in that three-dimensional parameter space. In the three-dimensional parameter space, two branches of codimension-two BT curves emanate from the codimension-three point. These curves form the intersection of saddle-node and Hopf bifurcation planes. Furthermore, codimension-two Bautin bifurcations as well as codimension-two homoclinic orbits to neutral saddles originate from the codimension-three BT point.

In two-parameter space, codimension-one homoclinic bifurcation curves can originate from BT points. Both saddle, saddle-node and neutral saddle homoclinic orbit bifurcations actually occur. The analysis reveals unexpected similarities between the elliptic and the focus cases, which have been overlooked in earlier theoretical studies.

The normal form analysis results are used to interpret those of the numerical bifurcation analysis study of the full planar two-stage population model and consequently also those reported in [34, Chapter 11].

2. The two-stage population model. The model is introduced in [21] and further motivated in [34, Chapter 11]. The population is split into juveniles (larvae, e.g.) and adults, the numbers of which are denoted by $L(t)$ and $A(t)$, respectively. The system has the form

$$\frac{dL}{dt} = \beta(L,A)A - \mu(L,A)L - f(L,A)L\,, \tag{2.1a}$$

$$\frac{dA}{dt} = f(L,A)L - \alpha(L,A)A\,, \tag{2.1b}$$

where β is the per capita reproduction rate of an average adult individual, μ is the per capita mortality rate of an average juvenile individual (α for the adults) and f the per capita transition rate from the juvenile into the adult stage. In general, these rates can depend on both the densities of juveniles and adults.

Under realistic additional assumptions, all trajectories converge to an equilibrium (not necessarily the same) if $\frac{\partial}{\partial L}[f(L,A)L] \geq 0$ for all $L, A > 0$ [34, Thm. 11.6]. (See [21] for a convergence result under different assumptions.) So there are no complex dynamics if the transition from the juvenile to the adult stage is only weakly affected by intra-juvenile competition. However, we show in this paper that complex behavior may indeed occur if there is both a strong influence on the per capita transition rate by intra-juvenile competition and a strong effect on the per capita birth rate by intra-adult competition. We restrict our consideration to pure intrastage competition, i.e., there is no competition between juveniles and adults. So the juvenile per capita mortality rate depends on the number of juveniles, i.e., $\mu(L)$, and the adult per capita mortality rate $\alpha(A)$ and reproduction rate $\beta(A)$ depend only on the number of adults A. This may occur when juveniles and adults have different habitats. For simplicity, the per capita mortality rates for both juveniles and adults are taken as constants. Time is scaled such that $\alpha = 1$; this means that one time unit equals the expectation of adult life. For convenience we divide the other rates β, μ and f by α, but we do not introduce new variables. As in [21], a Ricker-type function is chosen for the stage transition rate,

$$f(L) = \frac{\mu}{m}e^{-L}. \tag{2.2}$$

The reproduction rate $\beta(A)$ is chosen as

$$\beta(A) = g\big(\frac{m}{\mu}A\big). \tag{2.3}$$

To interpret the parameter m, we consider the expression for the probability of surviving the juvenile stage to become an adult in the absence of competition ($L = 0$):

$$p = \frac{f(0)}{\mu + f(0)} = \frac{1}{1+m}, \tag{2.4}$$

where $f(0)$ is evaluated using (2.2). The numerator $f(0)$ is the per capita transition rate from the juvenile and the adult stage (without competition) and the denominator $\mu + f(0)$ is the total rate at which juveniles leave their stage. For $m = 0$ all juveniles survive to adults, for $m = 1$ half survive, and as $m \to \infty$ the probability of surviving approaches zero.

After introducing the scaled number of adults y as

$$y = \frac{m}{\mu} A, \tag{2.5}$$

the stage-structured population model becomes

$$\frac{dL}{dt} = \frac{\mu}{m}\big[g(y)y - mL - Le^{-L}\big], \tag{2.6a}$$

$$\frac{dy}{dt} = Le^{-L} - y. \tag{2.6b}$$

At equilibrium, (2.6b) gives a fixed relationship between y^* and L^* which is independent of all parameters. Note from (2.6a) that these values are independent of the parameter μ. We assume that the birth rate $g(y)$ is also of Ricker-type:

$$g(y) = e^{(1/b)(a-y)}. \tag{2.7}$$

When not varied, the model parameters have the following reference values $a = 0.43, b = 2.2, m = 0.01$, while $0 \le \mu \le 1.0$.

3. Organizing centers. This section is partially based on [34, section 11.9]. The system (2.6) has been scaled in such a way that the equilibria do not depend on the parameter μ. The origin $(L = 0, y = 0)$ is always an equilibrium.

The parameter m can be expressed as a function of the L-component of the interior equilibria, L^*,

$$m = \big[g(L^*e^{-L^*}) - 1\big]e^{-L^*} =: M(L^*). \tag{3.1}$$

The value

$$m_0 = M(0) = g(0) - 1 = e^{a/b} - 1 \tag{3.2}$$

is the threshold value for the existence of interior equilibria and the stability of the origin. The Jacobian of system (2.6) evaluated at the origin reads

$$\begin{pmatrix} -\frac{\mu}{m}(m+1) & \frac{\mu}{m}g(0) \\ 1 & -1 \end{pmatrix}. \tag{3.3}$$

The determinant of this Jacobian matrix switches its sign at m_0. Hence, m_0 marks a transcritical bifurcation. If $m \ge m_0$, the origin is the only equilibrium and it attracts all solutions starting in the biologically relevant nonnegative quadrant. We therefore restrict our investigation to $m \in (0, m_0)$. Then the origin is a saddle with one part of its unstable manifold lying in the positive quadrant. Moreover the origin is a strong repeller and not part of ω-limit sets of solutions starting in the nonnegative quadrant, except at the origin itself. Further, there exists at least one interior equilibrium.

Whatever the choice of the positive parameters a and b, the function M is strictly decreasing for small $L^* > 0$ and large $L^* > 0$. It can also be shown that the equation $M(L^*) = m$ has at most three interior solutions. This implies that, depending on the

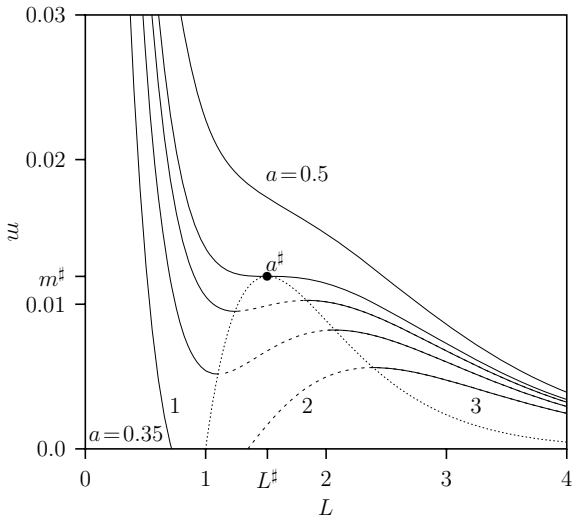


FIG. 3.1. *Graph of the function M for $b = 2.2$ and various values of a. From right to left: $a = 0.5$, $a = a^\sharp = 0.4492276697$, $a = 0.43$, $a = 0.4$, $a = 0.35$. The bullet gives the simultaneous point of inflection and critical point for $a = a^\sharp$. The dotted line connects the critical points where a is varied. The numbers label the branches where equilibrium E_i occurs at branch $i = 1, 2, 3$.*

choice of a and b, M is either strictly decreasing on $[0, \infty)$ or that there are numbers $L_2, L_1 > 0$ such that M is strictly decreasing on $[0, L_1]$ and $[L_2, \infty)$ and strictly increasing on $[L_1, L_2]$. When the transition occurs, the three positive equilibria merge into one *triple* equilibrium, and we have a unique $L^* > 0$ such that $M'(L^*) = 0 = M''(L^*) = 0$. If we fix $b > 0$, we can use this relation to determine the value of a at which the transition occurs, say $a^\sharp$. Then, M is strictly decreasing for $a > a^\sharp$ and has changing monotonicity behavior for $a < a^\sharp$. The value of L at the corresponding equilibrium $L^\sharp = L^*$, which together with (3.1) finally gives $m^\sharp = M(L^*)$ and $y^\sharp = y^* = L^* e^{-L^*}$.

Figure 3.1 shows the graph of the function M for $b = 2.2$ and various values of a. The dotted curve marks points where $M'(L^*) = 0$ when a is varied. This curve cuts the horizontal axis $m = 0 = M(L^*)$ at $L^* = 1$, where $a = a_\sharp = 0.3679$.

For the fixed choice of $a_\sharp < a < a^\sharp$ and $b > 0$, M changes its monotonicity; moreover M is strictly positive on $[0, \infty)$. Let

$$m_0 = M(0) = g(0) - 1, \qquad m_1 = M(L_1), \qquad m_2 = M(L_2). \tag{3.4}$$

The critical points L_1 and L_2 can be determined as the solutions of $M'(L) = 0$ and one finds $m_0 > m_2 > m_1$.

We have the following cases depending on m:

$m = m_1$: There are exactly two interior equilibria and their L-components satisfy $L_1^* = L_1$ and $L_2^\diamond := L_2^* > L_2$.

$m = m_2$: There are exactly two interior equilibria, and their L-components satisfy $L_1^\diamond := L_1^* \in (0, L_1)$ and $L_2^* = L_2$.

Then,

$m \geq m_0$: There is no interior equilibrium.

$m \in (m_2, m_0)$: There is exactly one interior equilibrium and its L-component satisfies $L^* \in (0, L_1^\diamond)$.

$m \in (m_1, m_2)$: There are exactly three interior equilibria and their L-components satisfy $L_1^* \in (L_1^\diamond, L_1)$, $L_2^* \in (L_1, L_2)$, $L_3^* \in (L_2, L_2^\diamond)$.

$m \in (0, m_1)$: There is exactly one interior equilibrium and its L-component satisfies $L^* > L_2^\diamond$.

$m = 0$: There is exactly one interior equilibrium $L^* = 1$.

One can show (see [34, page 173]) that the determinant of the Jacobian matrix of the system (2.6), with $y = Le^{-L}$, has the opposite sign of the derivative of M. Hence, every interior equilibrium with $L^* = L_1$ or $L^* = L_2$ has at least one eigenvalue 0. As we have seen, such equilibria occur if and only if $m = m_1$ or $m = m_2$. It can also be shown (see [34, page 172]) that the trace of the Jacobian matrix evaluated at an interior equilibrium is a linear function of μ. The trace switches its sign from negative to positive at

$$\mu = \phi(L^*) := \frac{g(y^*) - 1}{L^* - g(y^*)}, \qquad y^* = L^* e^{-L^*}. \tag{3.5}$$

At $a = a^\sharp$ we have $L^* = L_1 = L_2 = L^\sharp$ and $m = m_1 = m_2 = m^\sharp$, see Figure 3.1. At this point the determinant and the trace of the Jacobian matrix evaluated at the equilibrium $(L^\sharp, y^\sharp)$ are zero. Therefore, for fixed $b > 0$ and $\mu = \mu^\sharp, m = m^\sharp$, and $a = a^\sharp$, we have a BT point. Moreover, the equilibrium is triple at this point; thus a degenerate BT point occurs.

For $b = 2.2$, we have

$$L^\sharp = 1.513180178,\ y^\sharp = 0.33321523 \tag{3.6}$$

and

$$\mu^\sharp = 0.01179614,\ m^\sharp = 0.01192386945,\ a^\sharp = 0.4492276697. \tag{3.7}$$

4. Normal form analysis. In this section we perform a normal form analysis of the degenerate BT point and study its canonical unfolding.

4.1. Critical normal form. First, we write system (2.1) for fixed $b > 0$ at the critical parameter values (3.7) in a coordinate system where the equilibrium with the coordinates (3.6) is shifted to the origin of the phase plane by the transformation

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} L - L^\sharp \\ y - y^\sharp \end{pmatrix}. \tag{4.1}$$

The transformed planar system becomes

$$\dot{\mathbf{x}} = \mathbf{J}\mathbf{x} + \mathbf{F}(\mathbf{x}), \tag{4.2}$$

where $\mathbf{J}$ is the Jacobian matrix evaluated at the equilibrium and $\mathbf{F}(\mathbf{x}) = O(\|\mathbf{x}\|^2)$.

Next we use a similarity transformation to put the linear part in the Jordan canonical form. First we calculate two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$ such that

$$\mathbf{J}^2\mathbf{v} = 0, \quad \mathbf{J}\mathbf{v} = \mathbf{u}, \quad \|\mathbf{u}\| = 1, \tag{4.3}$$

where the vector norm is defined by $\|\mathbf{u}\|^2 = \langle \mathbf{u}, \mathbf{u} \rangle$ and $\langle \mathbf{u}, \mathbf{v} \rangle$ stands for the standard inner product in $\mathbb{R}^2$: $u_1 v_1 + u_2 v_2$. These two vectors (the eigenvector $\mathbf{u}$ and the generalized eigenvector $\mathbf{v}$), are linearly independent and form a basis in the plane. Notice that $\mathbf{v}$ is the eigenvector belonging to the zero eigenvalue of the matrix $\mathbf{J}^2$

($\mathbf{J}$ is nilpotent of index 2, that is, $\mathbf{J}^2\mathbf{v} = \mathbf{0}$ but $\mathbf{J}\mathbf{v} \neq \mathbf{0}$). One can calculate $\mathbf{v}$ as an eigenvector associated with the zero eigenvalue of the squared Jacobian matrix evaluated at the equilibrium point.

The similarity transformation is now defined by

$$\mathbf{x} = \mathbf{U}\mathbf{y}, \tag{4.4}$$

where $\mathbf{U}$ denotes the matrix the columns of which are formed by a normalized eigenvector and a generalized eigenvector. The matrix $\mathbf{U}$ is invertible, since the vectors $\mathbf{u}$ and $\mathbf{v}$ are linearly independent, and we can write

$$\mathbf{y} = \mathbf{U}^{-1}\mathbf{x}. \tag{4.5}$$

Then

$$\mathbf{U}^{-1}\mathbf{J}\mathbf{U} = \mathbf{J}_0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \tag{4.6}$$

and

$$\dot{\mathbf{y}} = \mathbf{J}_0\mathbf{y} + \mathbf{U}^{-1}\mathbf{F}(\mathbf{U}\mathbf{y}). \tag{4.7}$$

Taylor series expansion of the right-hand side of (4.7) at the equilibrium $\mathbf{y} = \mathbf{0}$ gives

$$\begin{aligned} \frac{dy_1}{dt} &= y_2 + \frac{1}{2}a_{20}y_1^2 + a_{11}y_1y_2 + \frac{1}{2}a_{02}y_2^2 \\ &\quad + \frac{1}{6}a_{30}y_1^3 + \frac{1}{2}a_{21}y_1^2y_2 + \frac{1}{2}a_{12}y_1y_2^2 + \frac{1}{6}a_{03}y_2^3 \\ &\quad + \frac{1}{24}a_{40}y_1^4 + \frac{1}{6}a_{31}y_1^3y_2 + \frac{1}{4}a_{22}y_1^2y_2^2 + \frac{1}{6}a_{13}y_1y_2^3 + \frac{1}{24}a_{04}y_2^4 + O(\|\mathbf{y}\|^5), \end{aligned} \tag{4.8a}$$

$$\begin{aligned} \frac{dy_2}{dt} &= \frac{1}{2}b_{20}y_1^2 + b_{11}y_1y_2 + \frac{1}{2}b_{02}y_2^2 \\ &\quad + \frac{1}{6}b_{30}y_1^3 + \frac{1}{2}b_{21}y_1^2y_2 + \frac{1}{2}b_{12}y_1y_2^2 + \frac{1}{6}b_{03}y_2^3 \\ &\quad + \frac{1}{24}b_{40}y_1^4 + \frac{1}{6}b_{31}y_1^3y_2 + \frac{1}{4}b_{22}y_1^2y_2^2 + \frac{1}{6}b_{13}y_1y_2^3 + \frac{1}{24}b_{04}y_2^4 + O(\|\mathbf{y}\|^5). \end{aligned} \tag{4.8b}$$

The final transformation to the normal form

$$\frac{d\xi}{d\tau} = \eta, \tag{4.9a}$$

$$\frac{d\eta}{d\tau} = A\xi^2 + B\xi\eta + C\xi^3 + D\xi^2\eta + E\xi^4 + F\xi^3\eta + O(\|(\xi,\eta)\|^5) \tag{4.9b}$$

is achieved by a time reparameterization combined with a smooth change of coordinates. In this way it is possible to remove both fourth-order terms from (4.9) (using $BC \neq 0$ as will be verified numerically in our case).

The time reparameterization introduces a new time τ as follows:

$$dt = (1 + \theta_1 y_1 + \theta_2 y_1^2)d\tau, \tag{4.10}$$

where θ_1 and θ_2 are to be defined later. Alternatively one can use $(1+\theta_1 y_1+\theta_2 y_2)$ which leads to the same results. The smooth transformation reads

$$\begin{aligned}\xi = {} & y_1 + \frac{1}{2}g_{20}y_1^2 + g_{11}y_1y_2 + \frac{1}{6}g_{30}y_1^3 + \frac{1}{2}g_{21}y_1^2y_2 + \frac{1}{2}g_{12}y_1y_2^2 \\ & + \frac{1}{24}g_{40}y_1^4 + \frac{1}{6}g_{31}y_1^3y_2 + \frac{1}{4}g_{22}y_1^2y_2^2 + \frac{1}{6}g_{13}y_1y_2^3, \end{aligned} \tag{4.11a}$$

$$\begin{aligned}\eta = {} & y_2 + \frac{1}{2}h_{20}y_1^2 + h_{11}y_1y_2 + \frac{1}{6}h_{30}y_1^3 + \frac{1}{2}h_{21}y_1^2y_2 + \frac{1}{2}h_{12}y_1y_2^2 \\ & + \frac{1}{24}h_{40}y_1^4 + \frac{1}{6}h_{31}y_1^3y_2 + \frac{1}{4}h_{22}y_1^2y_2^2 + \frac{1}{6}h_{13}y_1y_2^3, \end{aligned} \tag{4.11b}$$

where g_{ij} and h_{ij} are unknown coefficients. Differentiating (4.11a) and (4.11b) with respect to τ yields

$$\frac{d\xi}{d\tau} = (1+\theta_1 y_1+\theta_2 y_1^2)\left(\frac{\partial \xi}{\partial y_1}\frac{dy_1}{dt} + \frac{\partial \xi}{\partial y_2}\frac{dy_2}{dt}\right), \tag{4.12a}$$

$$\frac{d\eta}{d\tau} = (1+\theta_1 y_1+\theta_2 y_1^2)\left(\frac{\partial \eta}{\partial y_1}\frac{dy_1}{dt} + \frac{\partial \eta}{\partial y_2}\frac{dy_2}{dt}\right). \tag{4.12b}$$

Substituting (4.8a) and (4.8b) into (4.12a) and (4.12b), and then equating coefficients (4.9a) and (4.9b), gives the equations to find the coefficients $g_{ij}, h_{ij}, \theta_1, \theta_2$ in (4.11) and (4.10), as well as A, B, C, D in (4.9), where θ_1 and θ_2 are used to enforce $E = F = 0$. This gives

$$A = \frac{1}{2}b_{20}, \tag{4.13a}$$

$$B = a_{20} + b_{11}, \tag{4.13b}$$

$$C = \frac{1}{6}b_{30} - \frac{1}{2}a_{20}b_{11}, \tag{4.13c}$$

$$\begin{aligned} D = {} & \frac{1}{2}\,b_{21} + \frac{1}{4}\,b_{02}(b_{11}-a_{20}) + \frac{1}{2}\,b_{11}a_{11} + \frac{1}{2}\,a_{30} + \frac{3(a_{20}+b_{11})}{20(3\,a_{20}b_{11}-b_{30})} \\ & \times \big(b_{40} - 6\,a_{20}b_{21} - 3\,b_{11}b_{02}a_{20} + 3\,b_{02}{a_{20}}^2 \\ & \qquad -4\,b_{11}a_{30} + 6\,b_{30}a_{11} + b_{30}b_{02} - 6\,b_{11}a_{20}a_{11}\big)\,. \end{aligned} \tag{4.13d}$$

Note that the third-order coefficient D in (4.9) depends (via b_{40}) on the fourth-order terms of (4.8).

The Taylor coefficient A in (4.9) equals zero, since $b_{20} = 0$ because the critical equilibrium is triple. This implies that one of the nondegeneracy conditions for a classical codimension-two BT point is violated [22, p. 272]. This leads to a bifurcation point with codimension three (or higher) and could, together with the requirement $\det \mathbf{J} = 0$ and $\operatorname{tr} \mathbf{J} = 0$, be used as defining functions to determine the critical parameter values corresponding to this bifurcation point.

If $CD \neq 0$, then the truncated critical normal form

$$\frac{d\xi}{dt} = \eta, \tag{4.14a}$$

$$\frac{d\eta}{dt} = B\xi\eta + C\xi^3 + D\xi^2\eta \tag{4.14b}$$

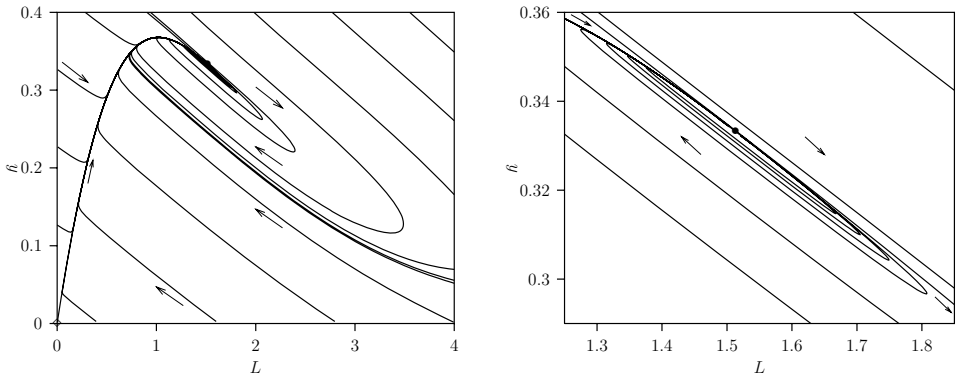


FIG. 4.1. *Phase portrait of* (2.1) *at the critical parameter values with* $b = 2.2$ *(left) and an enlargement of a neighborhood of the critical equilibrium (right), where the elliptic sector is clearly visible.*

can be simplified further by a linear coordinate and time scaling:

$$\frac{d\xi}{dt} = \eta, \tag{4.15a}$$

$$\frac{d\eta}{dt} = \beta\xi\eta + \epsilon_1\xi^3 + \epsilon_2\xi^2\eta, \tag{4.15b}$$

where $\epsilon_1 = \pm 1, \epsilon_2 = \pm 1$, and

$$\beta = \frac{B}{\sqrt{|C|}}.$$

In [2, 12], three topologically different cases are distinguished:

- Saddle case $\epsilon_1 = 1$, any ϵ_2 and β;
- Focus case $\epsilon_1 = -1$ and $0 < \beta < 2\sqrt{2}$;
- Elliptic case $\epsilon_1 = -1$ and $2\sqrt{2} < \beta$.

When $b = 2.2$, the calculated values of the coefficients of the truncated critical normal form (4.14) are $B = 1.0538275511$, $C = -0.110108078$, and $D = -1.23163654$. This gives $\epsilon_1 = -1$ (because $C < 0$), $\epsilon_2 = -1$ (because $D < 0$), and

$$\beta = \frac{B}{\sqrt{-C}} = 3.175849820.$$

We conclude that with $b = 2.2$ the *elliptic case* applies, for $\beta = 3.175849820 > 2\sqrt{2}$. Direct numerical integration of (2.1) at the critical parameter values (3.7) confirms this conclusion (see Figure 4.1).

Calculations showed that in an extended range of parameter values $b > 0$ the parameters C and D defined in (4.14b) are negative, implying $\epsilon_1 = -1$ and $\epsilon_2 = -1$. In Figure 4.2 we give the dependence of β on the parameter b. At $b^\natural = 1.7300228$, where $\beta = 2\sqrt{2}$, there is a transition from the elliptic case to the *focus* case. The phase portrait at the critical parameter values with $b = 1.5$ (when the focus case applies) is depicted in Figure 4.3. Compared with Figure 4.1, where $b = 2.2$, we see that the elliptic sector disappeared.

The transition at $b = b^\natural$ is a bifurcation of codimension four (or higher), which could be considered as the ultimate organizing center in model (2.6). However, in the following we will deal with $b > b^\natural$ corresponding to the elliptic case because it is

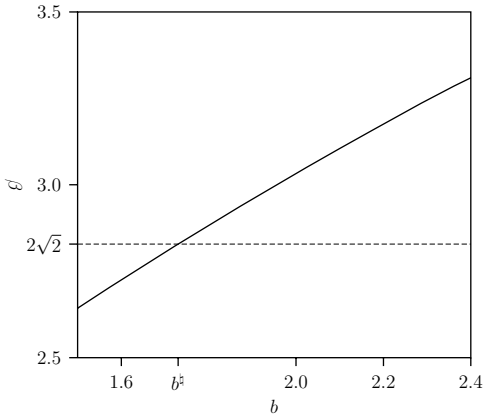


FIG. 4.2. *The coefficient β defined in* (4.15b) *as a function of the parameter b. There is a transition from the elliptic case to the focus case at the codimension-four bifurcation point, where $b = b^\natural = 1.7300228$.*

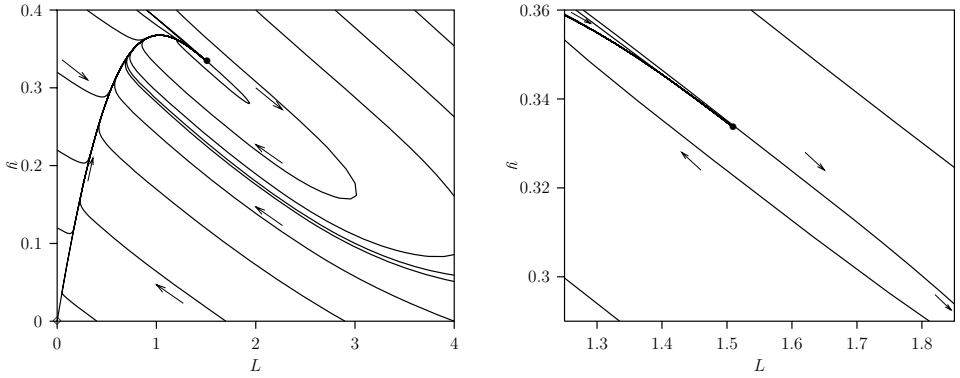


FIG. 4.3. *Phase portrait of* (2.1) *at the critical parameter values with $b = 1.5$ (left) and an enlargement of a neighborhood of the critical equilibrium (right), where no elliptic sector is present.*

the most interesting one. The focus case is discussed extensively in applied literature (for example, in [1, 22] with the analysis of the Rosenzweig–MacArthur predator-prey model having density-dependent mortality rate for the predators, as well as in [13], where an enzyme-catalyzed reaction model is studied). The elliptic case is much less understood, although it has been found in a mathematical model of a reaction of catalytic oxidation in [37].

4.2. Bifurcation diagram of the canonical unfolding. The local bifurcation diagram for the focus and elliptic case has been studied theoretically in [12], where the truncated critical normal form (4.15) is embedded in the following three-parameter family

$$\frac{d\xi}{dt} = \eta, \tag{4.16a}$$

$$\frac{d\eta}{dt} = -\mu_1 - \mu_2\xi + \nu\eta + \beta\xi\eta - \xi^3 - \xi^2\eta, \tag{4.16b}$$

with μ_1, μ_2, and ν serving as the unfolding parameters. Below we reconstruct the

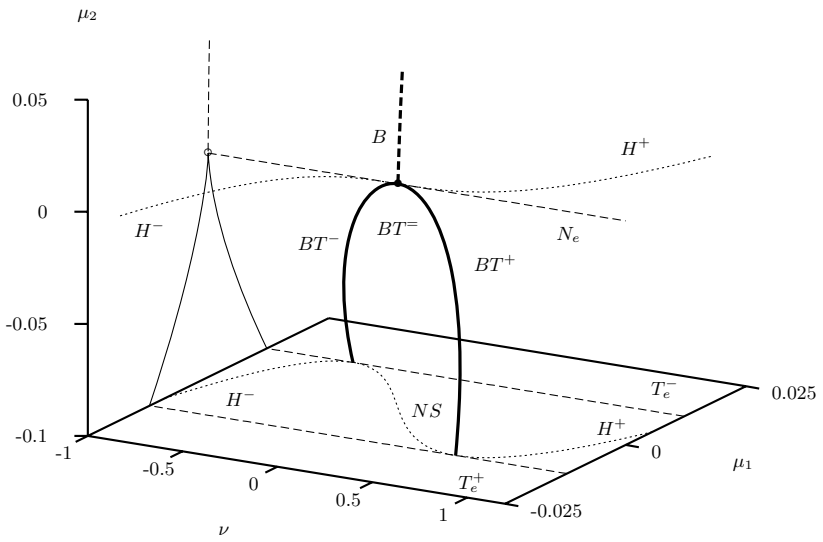


FIG. 4.4. *Three-parameter bifurcation diagram with μ_1, μ_2, and ν as bifurcation parameters for the normal form* (4.16)*, where $\beta = 3.175849820$. Only equilibrium bifurcations are shown. The tangent bifurcation curves $T_e^{\pm}$ (dashed curves) and Hopf $H^{\pm}$ and neutral saddle NS curve (dotted curve) in the (μ_1, ν)-plane for $\mu_2 = -0.1$ are plotted.*

bifurcation diagram of (4.16) in the elliptic case using the same numerical continuation methods as applied in the next section to compute global bifurcation diagrams of (2.1).

The equilibria of (4.16) satisfy $\eta = 0$ and $-\mu_1 - \mu_2\xi - \xi^3 = 0$. Therefore, depending on the parameter values μ_1 and μ_2, there is one or there are three real solutions. For example, when $\mu_1 = 0$, we have $\xi = 0$ and $\xi = \pm\sqrt{-\mu_2}$ besides $\eta = 0$.

In Figure 4.4, a partial three-parameter bifurcation diagram near the codimension-three bifurcation point $\mu_1 = \mu_2 = \nu = 0$ is shown for the normal form (4.16), where $\beta = 3.175849820$ (elliptic case). Two BT curves of different type emanate from the codimension-three point denoted by $BT^=$. Locally attracting limit cycles bifurcate from the first curve which is called a *supercritical Bogdanov–Takens bifurcation curve* and is denoted by BT^-, while repelling limit cycles bifurcate from the curve of second type which is a *subcritical Bogdanov–Takens bifurcation curve*, denoted by BT^+. Another codimension-two bifurcation curve, namely, a cusp curve denoted by N_e in the figure, passes through the point $BT^=$. Finally, from $BT^=$, a codimension-two Bautin (generalized or degenerated Hopf, see [22]) bifurcation curve B emanates. Software packages LOCBIF [18, 22] and CONTENT [25, 15] were used for the numerical continuation of these curves related to equilibrium bifurcations.

Figure 4.5 presents two-parameter slices of the complete bifurcation diagram of (4.16) with $\beta = 3.175849820$ for $\mu_2 = -0.1$ and $\mu_2 = -1$. The left diagram with $\mu_2 = -0.1$ gives a clear picture of the unfolding near the codimension-three BT point. The same diagram (together with corresponding phase portraits) is sketched in Figure 4.6, which we advise to consult while reading the rest of this section. The codimension-two $BT^{\pm}$ points and four saddle-node homoclinic bifurcation points $D_i, i = 1, \ldots, 4$ (analyzed theoretically in [26]) are indicated in all diagrams. These points lie all on the tangent bifurcation curves for equilibria $T_e^{\pm}$. At the supercritical BT^- point, a supercritical Hopf bifurcation curve H^- originates, and similarly a subcritical Hopf bifurcation curve H^+ emanates from the subcritical BT^+. From each BT point, BT^-

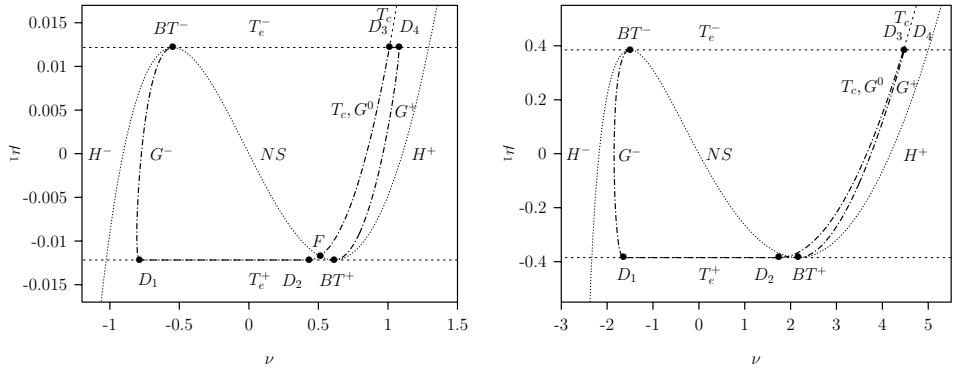


FIG. 4.5. *Two-parameter bifurcation diagrams of* (4.16) *with* μ_1 *and* ν *as bifurcation parameters and* $\mu_2 = -0.1$ *(left) and* $\mu_2 = -1$ *(right) for normal form* (4.16)*, where* $\beta = 3.175849820$*. The labels are explained in the text.*

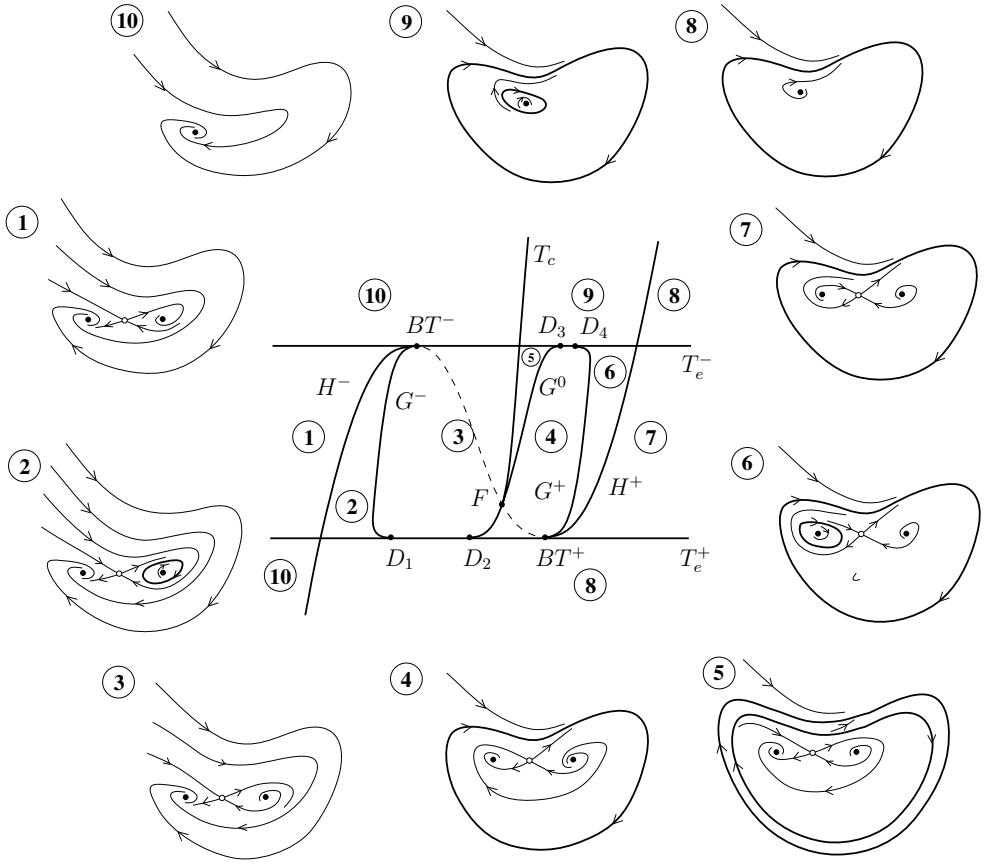


FIG. 4.6. *Schematic bifurcation diagram of* (4.16) *for small* $\mu_2 < 0$ *and* $\beta > 2\sqrt{2}$.

and BT^+, a global bifurcation curve emanates, indicated by G^+ or G^-, respectively. These are saddle homoclinic bifurcation curves. The homoclinic orbits corresponding to them are "small", i.e., they go around one equilibrium only. There exists another

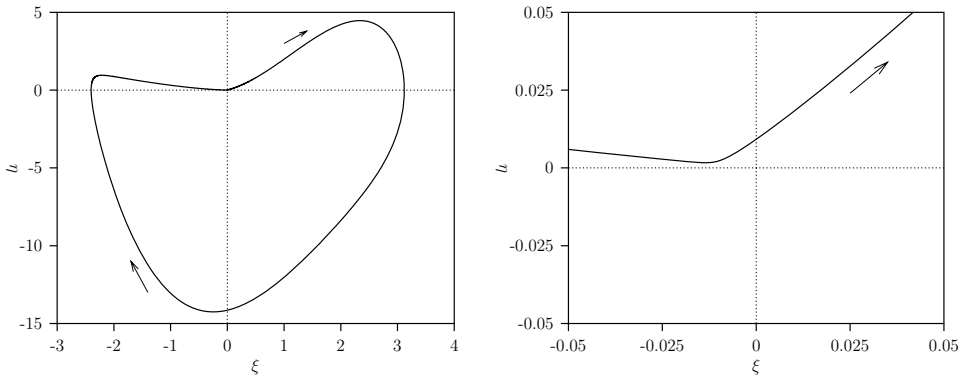


FIG. 4.7. *A nonhyperbolic limit cycle of* (4.16) *at the critical parameter value* $\nu = 0.8458472970$ *corresponding to* T_c *(left) and an enlargement of a neighborhood of the origin, where a saddle equilibrium is located (right). Other parameters are* $\mu_1 = 0, \mu_2 = -0.1, b = 3.175849820$. *We recall that the three equilibria are:* $\eta = 0$ *and* $\xi = 0$ *or* $\xi = \pm\sqrt{0.1}$. *The equilibrium in the origin is a saddle, the left one is stable and the right one unstable.*

saddle homoclinic bifurcation curve, denoted by G^0 in the figure. In contrast with G^+ and G^-, the homoclinic orbit corresponding to G^0 is "big", i.e., it surrounds two equilibria. The homoclinic curves were calculated using the package HOMCONT, a part of AUTO [11]. The implemented theory and numerical procedures are described in [4, 5].

A tangent bifurcation curve for limit cycles T_c, where two limit cycles collide and disappear, crosses the curve T_e^- and ends at a point F in the intersection of the homoclinic curve G^0 and the neutral saddle curve NS connecting the Bogdanov–Takens points BT^- and BT^+. This point F corresponds to a codimension-two "big" homoclinic orbit to a neutral saddle, where the trace of the Jacobian matrix is zero [22]. The tangent bifurcation curve for limit cycles T_c and the homoclinic curve G^0 have an infinite-order contact at F [29]. It should be noted that T_c is indistinguishable from G^0 in Figure 4.5. Between the equilibrium bifurcation curves T_e^+ and T_e^-, the curve T_c is located just above G^0. This can be verified by accurate computations in AUTO or CONTENT with many mesh points (e.g., `NTST=1000`). Figure 4.7 demonstrates that the critical limit cycle corresponding to T_c is located at a small but clearly visible distance from the saddle equilibrium at the origin. This nonhyperbolic limit cycle bifurcates into (outer) stable and (inner) unstable limit cycles, shown in Figure 4.8 for parameter values between the curves T_c and G^0. When we cross the homoclinic bifurcation curve G_0 above point F, the inner unstable limit cycle "collides" with the saddle and disappears via the saddle homoclinic orbit that is unstable from the outside, in accordance with the positive sign of the trace of the Jacobian matrix above the curve NS (see Figure 4.5). Crossing G^0 below F results in the appearance of a stable "big" cycle.

The saddle homoclinic curve G^-, originating at the point BT^-, terminates tangentially at a point D_1 on the bifurcation curve T_e^+. Between the two codimension-two points D_1 and D_2 the saddle-node equilibrium (existing along the tangent bifurcation curve T_e^+) has a smooth homoclinic orbit. On this line segment D_1D_2, the global bifurcation and the local bifurcation occur simultaneously. The homoclinic orbit is asymptotic to a saddle-node rather than to a saddle. In point D_2, the "big" saddle homoclinic curve G^0 departs tangentially from T_e^+. This curve G^0 ends also tangen-

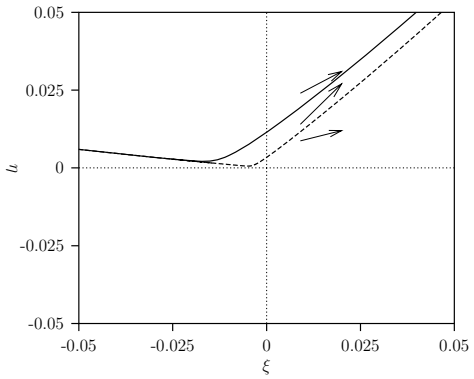


FIG. 4.8. *A stable limit cycle (outer, solid) and an unstable (inner, dashed) limit cycle of* (4.16) *at* $\nu = 0.8458473000$. *The other parameters are* $\mu_1 = 0, \mu_2 = -0.1, b = 3.175849820$.

tially at a codimension-two point D_3 on T_e^-, which is similar to D_2. Another saddle homoclinic curve, G^-, originating at BT^+ on the bifurcation curve T_e^+, terminates tangentially at a point D_4 on T_e^-. In all points D_i, the corresponding homoclinic orbits are nonsmooth at the saddle-node. The right diagram with $\mu_2 = -1$ shows that already in the vicinity of the codimension-three point the two saddle-node homoclinic bifurcation points D_3 and D_4 are also close to each other and hardly distinguishable in Figure 4.5.

4.3. Elliptic versus focus case. The bifurcation diagram of the normal form (4.16) presented in Figure 4.6 differs drastically from the theoretical bifurcation diagram for the elliptic case given in [12, p. 8]. The reason for this discrepancy is that the diagram in [12] concerns phase portraits in a *fixed* small neighborhood (in fact, an elliptic disk) of the origin. Therefore, additional bifurcation curves associated with boundary tangencies appear, while some parts of the global bifurcation curves described above become "invisible," since the corresponding bifurcations happen outside the neighborhood. This approach is absolutely legitimate in theoretical studies, but is of little use in applied analysis, where artificial boundaries have no meaning and global phase portraits in the whole phase plane must be considered. Only such portraits could provide a good understanding of the long-term dynamics of the model.

Figure 4.5 gives such global bifurcation diagrams of the normal form (4.16). It turns out that the two-parameter slices are topologically equivalent to those corresponding to the focus case (see [2, p. 36] or [12, p. 7], where an intersection of bifurcation surfaces with a small sphere centered at the origin in the three-dimensional parameter space is shown). However, the inner limit cycle demonstrates rapid amplitude changes ("canard-like" behavior) near the bifurcation curve T_c. It is this phenomenon that makes the continuation of limit cycles near T_c difficult. One could also observe that, in contrast with the focus case, the "big" homoclinic orbit to the neutral saddle (see point F) does not shrink to the origin of the phase plane, when we approach the codimension-three elliptic BT point in the parameter space. Instead, this homoclinic orbit tends to the boundary of the elliptic sector that has a finite size in (4.16). The similarity between the focus and the elliptic cases implies that the transition between them at $b = b^\natural$, although interesting from a theoretical point of view, is of minor importance in applications, since it does not affect dynamics away from the degenerate BT bifurcation points.

These similarities and differences between the focus and the elliptic cases were overlooked in all theoretical studies. Of course, global phase portraits arising from an elliptic BT case in a specific model could differ from the global bifurcation diagrams of the canonical unfolding (4.16) reported above. However, if other phase objects (equilibria, cycles, etc.) do not interact with the objects bifurcating around the codimension-three BT point, one would encounter the described bifurcation diagrams in his/her system, as it happens in our ecological model (2.1).

5. Bifurcation diagrams of (2.1). To facilitate our understanding of the bifurcation structure, we construct three-, two-, and one-parameter bifurcation diagrams of (2.1) for a representative parameter value $b = 2.2$ (elliptic case). We also show explicitly various homoclinic orbits.

5.1. Three-parameter bifurcation diagram. In Figure 5.1, the three parameters μ, m, a are changed simultaneously. There is a codimension-three point, denoted by $BT^{=}$ at the critical parameter values $(\mu^\sharp, m^\sharp, a^\sharp)$ given by (3.7). The bifurcation pattern resembles that of the normal form (4.16) presented in Figure 4.4. There is now another codimension-two Bautin bifurcation curve B which is not connected to a codimension-three point $BT^{=}$. Varying the fourth parameter b showed that such a connection never occurs in the region of the parameter space of interest in this paper.

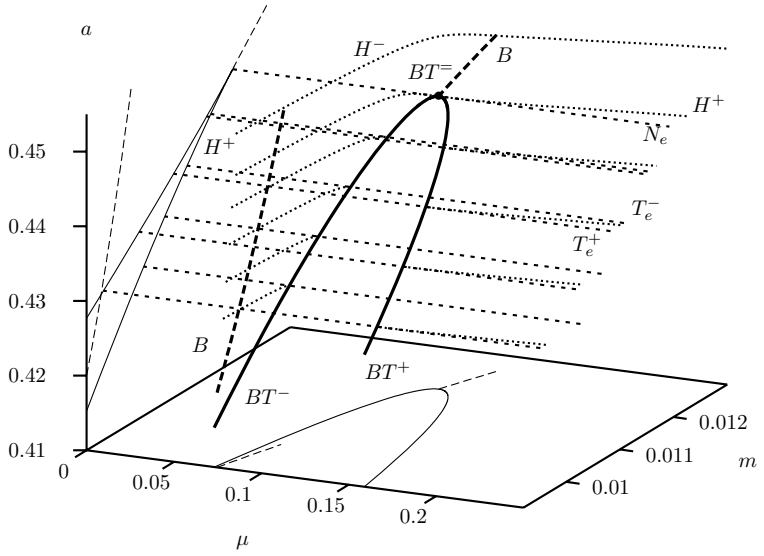


FIG. 5.1. *Three-parameter bifurcation diagram for equilibria of* (2.1) *with* μ, m *and* a *as bifurcation parameters. The codimension-three Bogdanov–Takens point is denoted by* $BT^{=}$. *From this point two codimension-two BT curves,* BT^{-} *and* BT^{+}, *emanate. They are shown along with their projections. The codimension-two Bautin bifurcation curves* B *are long dashed. The Hopf bifurcation curves (dotted) and tangent bifurcation curves (short dashed) are shown for* $a = 0.455$, $a = 0.4492277$, $a = 0.44$ *and* $a = 0.43$. *The curve* N_e *passing through* $BT^{=}$ *is a cusp bifurcation curve for equilibria. The two curves* $T_e^{\pm}$ *for* $a = 0.44$ *are the intersections of the* $a = 0.44$ *plane with the tangent bifurcation surfaces. Similarly the* H^{+} *and* H^{-} *curves from the intersections for the two-dimensional Hopf bifurcation surfaces. For a-values above the point* $BT^{=}$ *these sheets are separated by the codimension-two Bautin bifurcation curves* B. *These sheets are separated for a-values below the point* $BT^{=}$, *where the two codimension-two BT curves,* BT^{-} *and* BT^{+}, *form the end curves of the separated sheets. For smaller* μ*-values there is always a codimension-two Bautin bifurcation curve* B *that separates super- and subcritical Hopf bifurcation surfaces.*

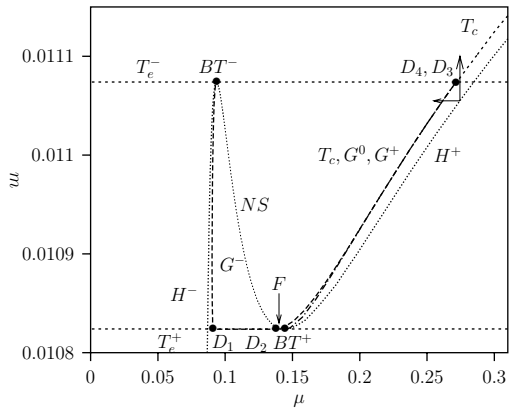


FIG. 5.2. *Two-dimensional bifurcation diagram with μ and m as bifurcation parameters, where $a = 0.44$.*

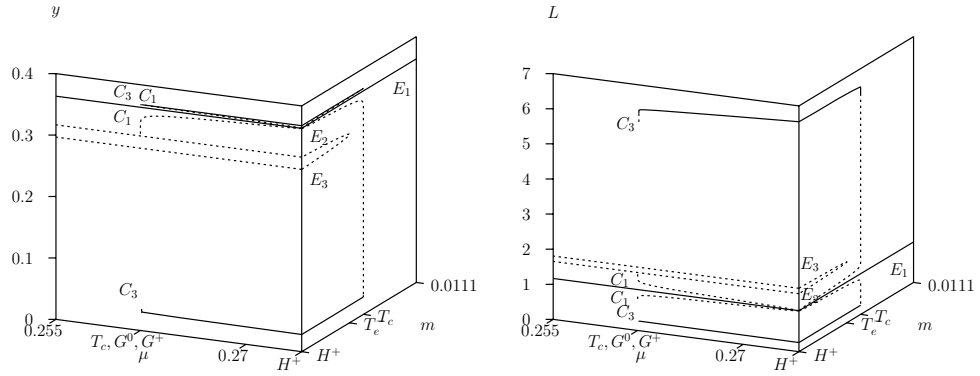


FIG. 5.3. *Combined bifurcation diagrams for region around bifurcation point D_3, where $a = 0.44$. The bottom right vertex, for both y and L, corresponds to the point $\mu = 0.2743923$ and $m = 0.01105501$ on the Hopf bifurcation H^+ in Figure 5.2.*

Codimension-one bifurcation curves for $a = 0.44$ and for the reference value $a = 0.43$ are investigated in the next section using two-parameter bifurcation diagrams.

5.2. Two-parameter bifurcation diagram. The two-parameter bifurcation diagram for $a = 0.44$ in Figure 5.2 strongly resembles that given in Figure 4.5 (right), where $\mu_2 = -1$. The Hopf bifurcation curves H^- (origination to the left in BT^-) and H^+ (origination to the right in BT^+), the neutral-saddle curve NS (between BT^- and BT^+), and the tangent bifurcation curves $T_e^{\pm}$ (straight lines parallel to the ν-axis) as well as the tangent bifurcation curve for limit cycles T_c are shown. Observe that although the parameter value $a = 0.44$ is rather close to $a^\sharp$, the two saddle-node homoclinic bifurcation points D_3 and D_4 are indistinguishable in the figure. In two one-parameter bifurcation diagrams with respect to μ and m combined in Figure 5.3, the extrema of the cycles in the region close to points D_3 and D_4 are plotted. The two homoclinic bifurcations for the cycles C_1 and C_3 are connected via the tangent bifurcation of these cycles at T_c. This explains how the large amplitude cycle C_3 surrounding all three equilibria, E_1, E_2, and E_3, appears. The tangent bifurcation

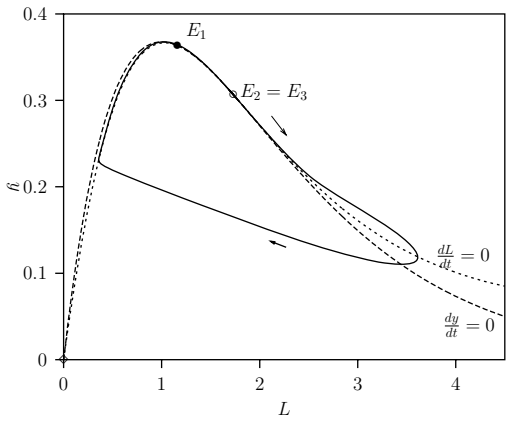


FIG. 5.4. *Phase-plane plot for parameter values between the codimension-two points D_3 and D_4, $\mu = 0.2714740$ and $m = 0.01107377$ for $a = 0.44$. Also the two isoclines are shown. These intersect in the two equilibria, namely, the saddle-node, where two points E_2 and E_3 coincide, and the unstable equilibrium E_1.*

curve T_c for limit cycles goes all the way down to the point F corresponding to the homoclinic orbit to a neutral saddle. Between the lines $T_e^{\pm}$, T_c is very close to the "big" homoclinic curve G^0.

A smooth homoclinic orbit to a saddle-node on the tangent bifurcation curve T_e^- between the points D_3 and D_4 is depicted in Figure 5.4. Also the isoclines are drawn. The homoclinic orbits locally coincide with a center manifold W^c of the saddle-node $E_2 = E_3$ and is therefore smooth. The orbit approaches and leaves the saddle-node via the center manifold, that is, it is tangent to the eigenvector belonging to the zero eigenvalue of the saddle-node on T_e^-. At the points D_3 and D_4 the homoclinic bifurcation curves G^+ and G^0 terminate at the equilibrium tangent bifurcation curve T_e^-, where the equilibria E_3 and E_2 coincide. At these points, the homoclinic orbit to the saddle-node is nonsmooth.

Figure 5.5 shows a two-parameter bifurcation diagram where μ and m are again the bifurcation parameters, while now $a = 0.43$. The two-parameter bifurcation diagram in Figure 5.5 resembles that for $a = 0.44$ given in Figure 5.2 except that there are no points D_3, D_4. If the values of the parameter a are lower, the Hopf bifurcation H^+ does not intersect the curve T_e^-. One can show that H^+ has a horizontal asymptote as $\mu \to \infty$, namely, the line $m = m_\sharp = 0.01026241193$. Furthermore, the curve G^- originating at BT^- and terminating at D_1 follows closely the Hopf bifurcation curve H^-. Figure 5.6 is an expanded view of Figure 5.5 in the region of the the two points D_1 and D_2. The tangent bifurcation curve T_c for limit cycles is indistinguishable from the segment of G^0 located above the point F also in Figure 5.6.

5.3. One-parameter bifurcation diagrams. In Figures 5.7 and 5.8, where $m = 0.01, a = 0.43$, the bifurcation parameter is μ. For all μ-values, there are three interior equilibria, denoted by E_1, E_2 and E_3, the positions of which are independent of μ. For the (scaled) numbers of juveniles and adults at these equilibria, $E_j = (L_j^*, y_j^*)$, we have $L_1^* < L_2^* < L_3^*$ and $y_1^* > y_2^* > y_3^*$, respectively. Hopf bifurcations of E_1 and E_3 occur at H^+ and H^-, respectively, giving rise to limit cycles. Their maximum and minimum values are plotted in the figure, a stable one, denoted by C_3 and generated by the supercritical Hopf bifurcation from the equilibrium E_3 and an unstable one,

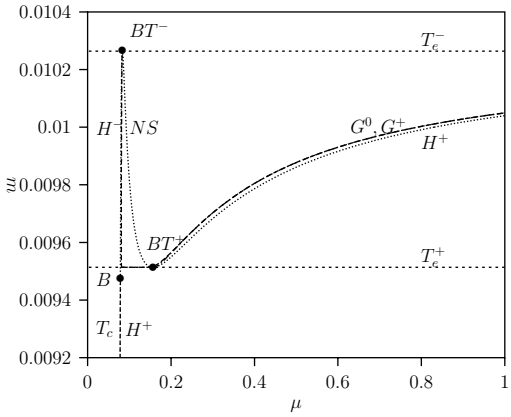


FIG. 5.5. *Two-parameter bifurcation diagram with μ and m as bifurcation parameters, where $a = 0.43$. At the Bogdanov–Takens codimension-two point BT^-, a tangent (T_e^-), supercritical Hopf (H^-) and a homoclinic curve (G^-) meet. Similarly, at the point BT^+, a tangent (T_e^+), subcritical Hopf (H^+) and a homoclinic curve (G^+) meet. At the Bautin codimension-two point B, a tangent bifurcation for the limit cycle T_c originates. For more details, see Figure* 5.6.

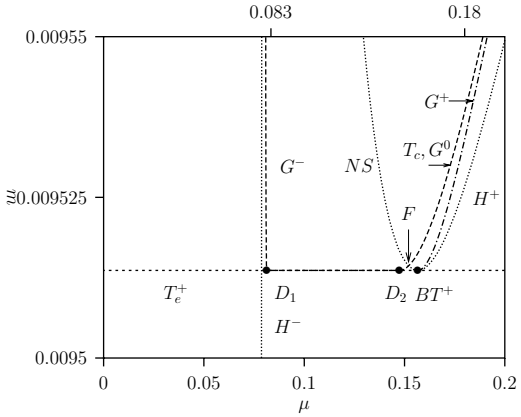


FIG. 5.6. *The Hopf bifurcation curve H^- (dotted) and the tangent bifurcation T_e^+ (short dashed) are shown. Further, the homoclinic bifurcation curve G^- which originates from BT^- (shown in Figure* 5.5*), and the homoclinic bifurcation curve G^+ (dashed-dotted) which originates from BT^+ ($\mu = 0.1566$) together with T_e^+ are shown. Between the codimension-two points D_1 ($\mu = 0.1474$) and D_2 ($\mu = 0.0813$), there is a smooth saddle-node homoclinic orbit. The tangent bifurcation curve T_c for limit cycles is indistinguishable from the segment of the homoclinic curve G^0 located above the point F.*

denoted by C_1 and generated by the subcritical Hopf bifurcation from the equilibrium E_1. For large μ, there is a branch of "big" stable periodic orbits (also denoted by C_3) which surround all three equilibria. This branch also exists for μ values for which the unstable period orbits C_1 surrounding the equilibrium E_1 exist. There are homoclinic orbits at the points where these limit cycles touch the saddle equilibrium E_2.

In Figure 5.9, we show the one-parameter diagram for fixed $\mu = 0.08337$ with m as the bifurcation parameter. The parameter μ has been chosen in such a way that it lies between the μ-coordinates of the codimension-two points D_1 and D_2 in Figure 5.6, where a saddle-node homoclinic bifurcation occurs. In this global bifurcation curve,

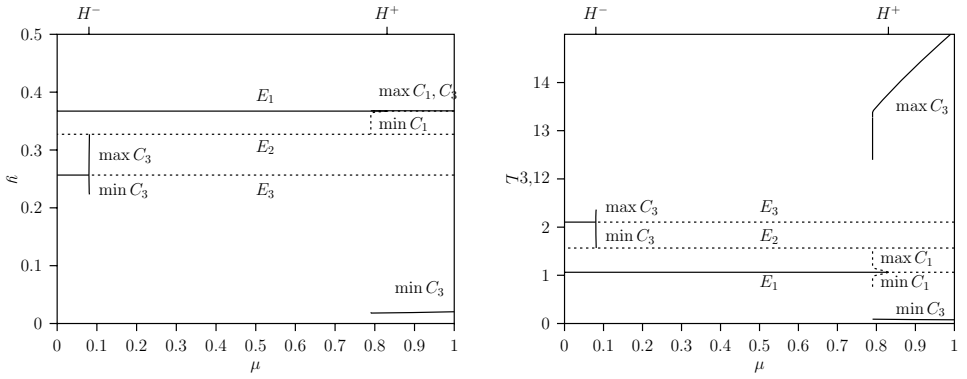


FIG. 5.7. *Bifurcation diagram, where $m = 0.01$ and $a = 0.43$, and μ is the single bifurcation parameter. Stable equilibria are represented by solid lines and unstable equilibria by dashed lines which are horizontal because the y and L equilibrium values are independent of μ. A branch of stable limit cycles C_3 (solid curves) originate through the supercritical Hopf (H^-), and a branch of unstable limit cycles C_1 (dashed curves) through the subcritical Hopf bifurcation (H^+). A homoclinic orbit occurs where these limit cycles touch the saddle equilibrium E_2.*

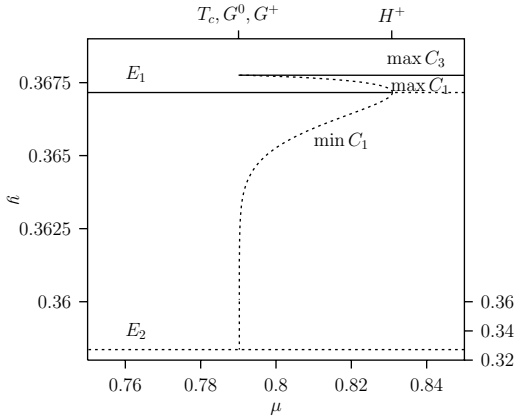


FIG. 5.8. *Detail of Figure 5.7. The solid curve is the maximum of the large amplitude cycle C_3. The dashed curve is the unstable limit cycle C_1 that emanates from the Hopf bifurcation point H^+. In order to get a better plot, the scaling of the y-axis below $y = 0.36$ is changed. The two homoclinic bifurcations G^0 and G^+, as well as the tangent bifurcation T_c of limit cycles, occur at almost the same value of μ.*

also a local tangent bifurcation T_e^+, where the two equilibria E_1 and E_2 coincide, is involved. The time-averages of the limit cycles are plotted. The stable limit cycle C_3 disappears when the average reaches the tangent point at T_e^+. This is explained in Figure 5.10, where the homoclinic orbit in the phase plane is shown for parameter values of m at the saddle-node homoclinic bifurcation. The homoclinic orbit is similar to that shown in Figure 5.4 for a point between the two codimension-two points D_3 and D_4.

A periodic solution $(L(t), y(t))$ near the saddle-node homoclinic orbit stays close to a point where the saddle-node will appear for most of its period, then completes the orbit by making a rapid and large excursion in the phase plane. Hence the time-average for one period is approximately the equilibrium value. This explains why it

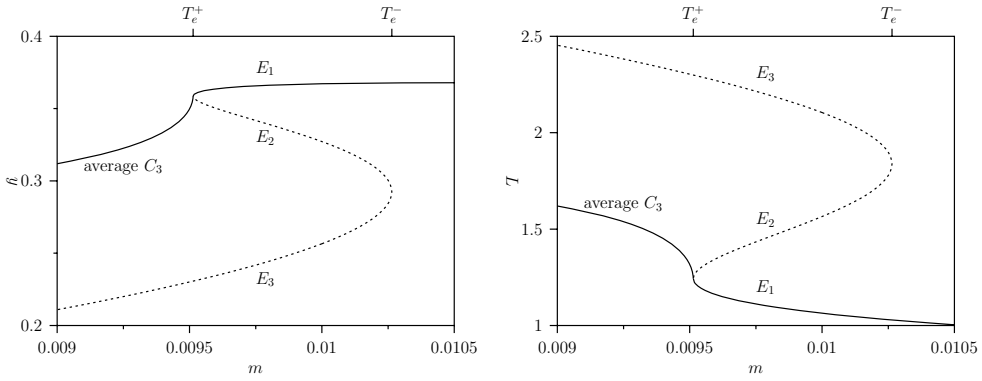


FIG. 5.9. *Bifurcation diagrams for* $y = \frac{\mu}{m}A$ *and* L *with* m *as bifurcation parameter for* $\mu = 0.08336837$ *and* $a = 0.43$. *Stable equilibria (solid curve) and unstable equilibria (dashed curve) as well as the time-averages for the stable limit cycle (solid curve) are plotted. Stable limit cycles originate in the supercritical Hopf* (H^-) *bifurcation shown in Figures* 5.5 *and* 5.6. *These cycles terminate at a smooth saddle-node homoclinic orbit.*

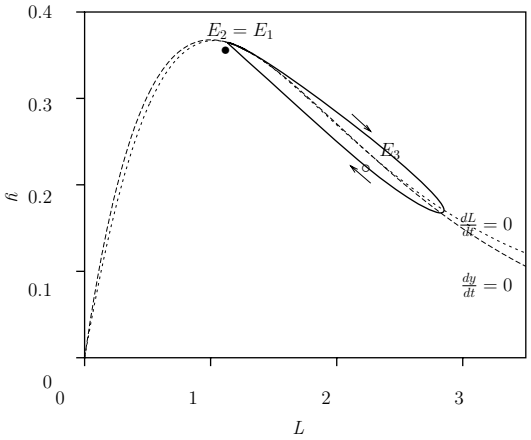


FIG. 5.10. *Phase portrait for* $\mu = 0.083$, $a = 0.43$ *and* $m = 0.009513629$. *The smooth homoclinic orbit to the saddle-node equilibrium* $E_1 = E_2$ *coincides locally with a center manifold of this point. Also, the unstable equilibrium* E_3 *inside the homoclinic orbit and two isoclines are shown.*

is advantageous to plot the time-averages in addition to the maximum and minimum values. Observe that y becomes small during the excursion but this occurs over a relatively (the period T goes to infinity) short time interval. Note that the unstable manifold of the saddle E_2 can go to E_1 directly or via a "big" loop around E_3 similar to the saddle-node homoclinic orbit shown in Figure 5.10.

Figure 5.11 is a one-parameter diagram, where m is the bifurcation parameter for $\mu = 1.0$ and $a = 0.43$. It shows how the three interior equilibria are connected. There are two tangent bifurcations $T_e^{\pm}$, where two equilibria coincide. A subcritical Hopf bifurcation of E_1 occurs at H^+, generating a branch of unstable limit cycles, denoted by C_1, while a supercritical Hopf bifurcation of E_3 occurs at H^-, generating a branch of stable limit cycles, denoted by C_3. The maximum and minimum values for both branches have been plotted in the figure. Varying m, this unstable limit cycle touches the saddle E_2 in a homoclinic orbit.

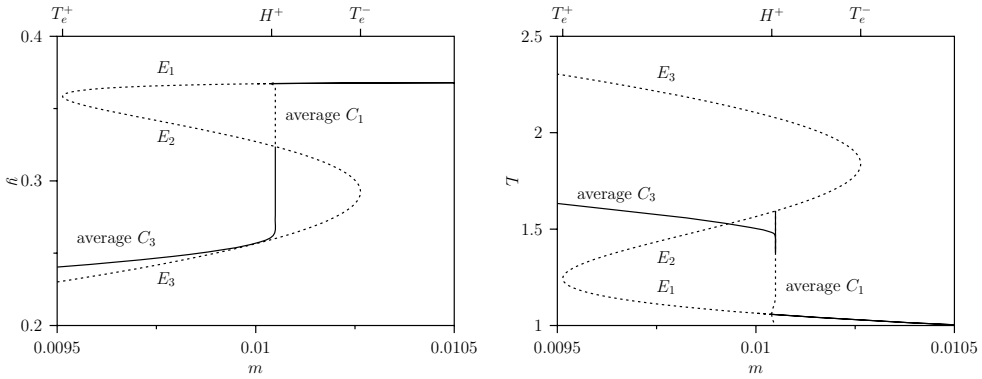


FIG. 5.11. *Bifurcation diagrams for $y = \frac{\mu}{m}A$ and L with m as bifurcation parameter, where $\mu = 1.0$ and $a = 0.43$. Stable equilibria (solid curve) and unstable equilibria (dashed curve) as well as the time-averages for the stable limit cycles (solid curve) and for the unstable limit cycles (dashed curve) are plotted.*

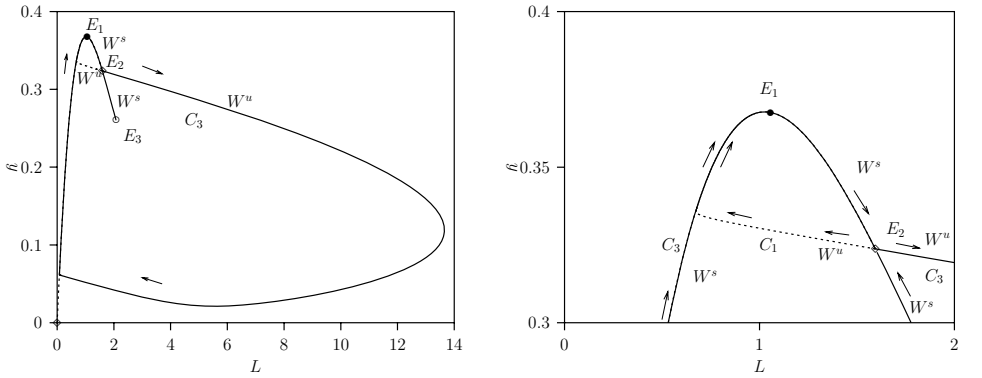


FIG. 5.12. *Two homoclinic orbits and time-evolution for $\mu = 1.0$ and $a = 0.43$, where $m = 0.01004953$ and $m = 0.01004952$ for large amplitude and small amplitude orbits, respectively. The right figure is an expanded view of the left figure. The stable W^s and unstable W^u manifolds of the saddle point indicated by a $\diamond$ are shown. The stable focus E_1 is indicated by a bullet and the source E_3 by a circle. Here the flow direction is indicated by arrows for the stable homoclinic orbit C_3 at the outside and for the unstable homoclinic orbit C_1 at the inside.*

Furthermore, the average values for the limit cycle C_3 are plotted. The unstable cycle C_3 originates in the supercritical Hopf bifurcation H^- of equilibrium E_3. Both average curves touch the saddle curve E_2. The average curve as a function of m for the stable limit cycles C_3 is very steep close to the bifurcation point. The average L-values first decrease before increasing up to the saddle-point value. This is most clear from the graphs $y(m)$ and $L(m)$. Actually, the cycle C_3 becomes unstable via a tangent limit cycle bifurcation not show in the figure. Both homoclinic bifurcations occur in very close proximity.

Figure 5.12 shows approximations of these two homoclinic orbits, one "big" (solid) C_3 and one "small" (dashed) C_1. The period of the plotted limit cycles is very large, indicating a homoclinic orbit. There are two rather sharp bends in the "big" homoclinic orbit. The top one is close to the saddle point as expected. The lower bend is associated with the zero equilibrium $(L = 0, y = 0)$ which is a saddle point for the parameter values used. Both orbits pass the stable focus E_1 closely.

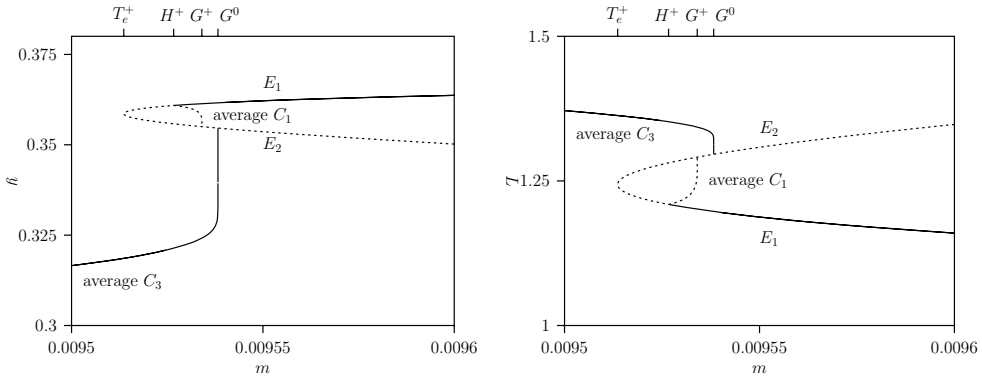


FIG. 5.13. *Bifurcation diagrams for $y = \frac{\mu}{m}A$ and L with m as bifurcation parameter, where $\mu = 0.18$ and $a = 0.43$. Stable equilibria (solid curve) and unstable equilibria (dashed curve) as well as the time-averages for the stable limit cycle (solid curve) and for the unstable limit cycle (dashed curve) are plotted.*

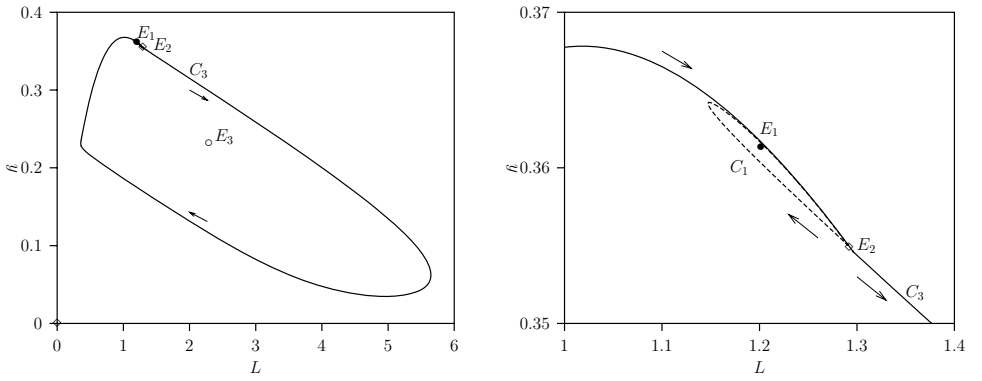


FIG. 5.14. *Phase portrait for $\mu = 0.18$ and $a = 0.43$. Left: Where $m = 0.009534053$, there is a "big" homoclinic orbit C_3. Right: An expanded view of the homoclinic orbit C_3 (solid curve) close to saddle E_2. For $m = 0.009538250$, in the same plot the "small" homoclinic orbit C_1 (dashed curve) is shown.*

In order to get more detailed information on the dynamics close to the homoclinic orbits, we study the one-parameter diagram for fixed $\mu = 0.18$, where m is varied (see also Figure 5.6).

In Figure 5.13, the average values of the variables for the limit cycles are plotted together with their equilibrium values. For m values below the Hopf bifurcation H^+, the large amplitude stable limit cycles, whose averaged values form the curve C_3, are globally attracting. For m values above the curve H^+, the unstable limit cycles C_1 which emanate from E_1 at the Hopf bifurcation H^+ are separating boundaries. Starting outside the limit cycle results in convergence to the stable limit cycle C_3, but starting inside the limit cycle gives convergence to the stable equilibrium E_1. Increasing m, this bistability persists until the unstable limit cycle (averaged C_1) disappears abruptly at the global bifurcation G^+ via the homoclinic orbit shown in Figure 5.14. For parameter values m above G^+, there is still bistability of the stable equilibrium E_1 and the stable limit cycle C_3, but now the stable manifold of the saddle equilibrium E_2 acts as the separatrix.

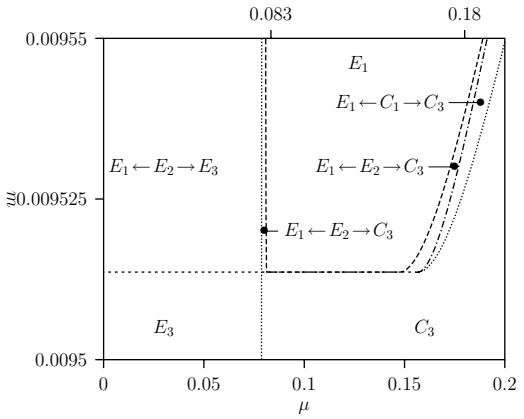


FIG. 5.15. *Attractors in some regions of the two-parameter bifurcation diagram shown in Figure 5.6.* $E_1 \leftarrow E_2 \rightarrow E_3$ *means bistability of the two equilibria* E_1 *and* E_3 *and the stable manifold of the saddle* E_2 *is the separatrix.*

In Figure 5.15 several attractors for typical points in the two-parameter bifurcation diagram Figure 5.6 are indicated. The regimes are all separated by codimension-one curves as shown in Figure 5.6. For small μ values, there is bistability of the two equilibria E_1 and E_3 and the stable manifold of the saddle E_2 is the separatrix. This region is labeled $E_1 \leftarrow E_2 \rightarrow E_3$. For larger μ values there is bistability of the equilibrium E_1 and the limit cycle C_3, while the stable manifold of either the saddle E_2 ($E_1 \leftarrow E_2 \rightarrow C_3$) or the unstable limit cycle C_1 ($E_1 \leftarrow C_1 \rightarrow C_3$) is a separatrix. Note that C_3 can surround either just E_3 or all three equilibria, E_1, E_2, and E_3.

6. Discussion. We have considered the most basic model for stage-structured populations in which per capita transition from the juvenile into the adult class is density dependent, namely, a system of two ordinary differential equations. This model was originally suggested in [21]. Birth pulses have been added in [33], but we are only interested in the time-autonomous case here. Alternative more sophisticated models involve differential-delay equations with density-dependent delays [17], transport equations with density-dependent speed [27, 8], difference equations [3, 6, 7], and integral equations with density-dependent transition kernels [10]. We refer to [20] for a review of the different model formulations and their relationships. While more complex models also exhibit more complex behavior than simple models, the complex behavior may be more difficult to detect and track due to the large number of parameters and initial conditions inherent in such systems.

In [21], periodic orbits were found which surround a unique interior equilibrium. Up to three interior equilibria, two of which can undergo Hopf bifurcation, were discovered in [34, sect. 11]. In order to determine more systematically the dynamics of our planar system (2.6), we have performed a four-parameter bifurcation analysis under the assumption of pure intrastage competition. To report our results, we fix a representative value of parameter b that characterizes the birth rate of adults and then presents the three-parameter bifurcation diagrams. These are the three parameters which we have varied. The parameter a is related to the average number of offspring produced by one typical adult if there is no competition. The parameter μ represents the per capita mortality rate of juveniles. The parameter m has no direct interpretation, but $p_0 = \frac{1}{1+m}$ is the probability of surviving the juvenile stage under

no competition. The parameter a determines the shape of a function M involved in the equation $M(L^*) = m$ which determines the juvenile coordinate L^* of an interior equilibrium. Depending on a, M can be either strictly monotone decreasing in $L^* \geq 0$ or decreasing for small L^*, increasing for intermediate L^*, and decreasing again for large L^*. The $a = a^\sharp$ value at which M undergoes the transition is the a-coordinate of a codimension-three BT bifurcation point which is the organizing center of our three-parameter bifurcation diagrams. There exists a unique $L^* > 0$ such that $M'(L^*) = M''(L^*) = 0$. L^* is the juvenile coordinate of a saddle-node. $m = M(L^*)$ is the m coordinate of the degenerate BT point. Its μ-coordinate is determined by making 0 a double eigenvalue of the Jacobian matrix. This leads to the following procedure for calculating the position of the codimension-three BT bifurcation point. When b is fixed, there are five unknowns, namely, two equilibrium values for the state variables, $L^\sharp, y^\sharp$ and three parameters $a^\sharp, m^\sharp, \mu^\sharp$. There are two equilibrium equations (right-hand sides of (2.6) zero), the requirements that the determinant and trace of the Jacobian matrix evaluated at the equilibrium point are zero and the additional requirement that $M''(L^\sharp) = 0$. The first four are necessary for a codimension-two BT point, while the latter is satisfied at a codimension-three BT point. Surprisingly, the latter condition is a geometrical one (the position of the inflection point of a function $M(L^*)$ defined in (3.1)). The function $M(L^*)$ is the expression for a parameter, m, derived from the equilibrium equation for one state variable, L, where the solution of the other state variable, $y^*(L^*)$, from the second equilibrium equation is substituted. We use the fact that $M(L^*)$ and $y^*(L^*)$ formulations are explicit as is the case with the population model (2.6). Commonly it is the requirement that one coefficient of a higher-order term of a Taylor series expansion equals zero (see [22, p. 272] for details). In this paper we established existence of a codimension-three BT bifurcation in our model and have verified its nondegeneracy by computing its normal form numerically.

The normal form is computed using a preliminary linear transformation to simplify the linear terms, in this case to put the linear part in the Jordan canonical form. In [22, 23] this step is omitted by using a representation of any vector in the state space as a linear combination of the eigenvector and the adjoint eigenvector. That method is appropriate for higher-order dynamics systems where normalization on the center manifold is needed [24]. The nonlinear smooth transformation in this paper is combined with a time reparameterization. This facilitates the removal of all fourth-order terms in the Taylor series expansion. Therefore the expression for the coefficient D of the $\xi^2\eta$ term given by (4.13d) differs from the classical expression [19], where only third-order terms are taken into account. The expressions (4.13) agree with those reported in [14]. In this way, we have computed the relevant normal form coefficients and established that the critical equilibrium at the codimension-three BT point is triple and has an elliptic sector for $b > b^\natural = 1.7300228$. Apparently, this case has never been observed in ecological modeling. At $b = b^\natural$, a transition from the elliptic codimension-three BT to the focus codimension-three BT point occurs.

Using known theoretical results [2, 12], we have concluded that two BT codimension-two curves emanate from the codimension-three BT point. Along these curves a Hopf bifurcation surface, as well as a tangent bifurcation and a homoclinic orbit surface, meet. There is also a codimension-two Bautin bifurcation curve B in the Hopf bifurcation surface, where the supercritical Hopf bifurcation becomes subcritical or vice versa. This happens in a parameter region where only one interior equilibrium exists. In Bautin bifurcation points, a surface of tangent bifurcations of limit cycles originates.

Our numerical analysis of the global bifurcation diagram of the canonical unfolding of the normal form, Figure 4.5, has revealed other global bifurcation curves. On codimension-two bifurcation curves D_i, transitions of the homoclinic orbit to a saddle-node homoclinic orbit occur. There is also a bifurcation surface G^0 corresponding to a "big" homoclinic orbit to the saddle that surrounds two equilibria. In this surface, a line F of codimension-two homoclinic orbits to a neutral saddle exists. Figure 5.2 resembles that of the Bazykin's predator-prey model [1, Fig. 3.5.3] and [22, Fig. 8.10], despite the fact that the degenerate BT point is here of the elliptic type while in Bazykin's model it is of the focus type. The diagrams given in [12] for both types differ a lot on first sight. However, in [12] additional bifurcation curves associated with boundary tangencies are reported. In this paper, we do not have such artificial bifurcations, since we do not restrict our attention to a small neighborhood of the critical equilibrium. Instead, we consider global phase portraits, which allows us to obtain a better understanding of the long-term dynamic behavior of the model. It should be stressed that without a preliminary analysis of the canonical unfolding (4.16) it would have been practically impossible to understand the numerical continuation results for (2.6).

A large-time solution behavior of similar complexity has been found for certain planar predator-prey systems [1, 22, 38, 31, 39] and epidemic models [30]. Predator-prey and host-parasite systems involve two species which influence each other in opposite ways with one suffering while benefiting the other. The predator-prey models which show complex behavior take into account competition among prey and predator competition for prey. Our model involves two stages of one species which compete among themselves but influence each other positively because we assume pure intra-stage competition. (Adding interstage competition actually counteracts the complexity.) Mathematically, this is reflected in the fact that the nonlinear terms in our system only depend on one variable each, whereas predator-prey and epidemic models always have nonlinear terms which involve both variables. In this paper, we find complex behavior in a class of planar systems that is different from predator-prey models both biologically and mathematically.

The complex behavior only occurs in a small parameter region (see [34, Chapter 11] for a discussion of other parameter ranges). In this parameter region, the per capita mortality rate of juveniles is smaller than the one of adults ($\mu \leq 1 = \alpha$), and almost every juvenile makes it into the adult stage if there is no competition (the small values of m result in $p_0 \approx 1$). This would hold for species where the juveniles are less prone to enemies than the adults, and the juvenile stage is very short if there are plenty of resources available. We expect that the complex behavior can be observed in a larger parameter region if more sophisticated nonlinearities than the classical Ricker function are chosen to model the competition among juveniles and adults. A thorough understanding of the basic stage-structured model is important because its simplicity makes it a useful building block in models for several species. Recently it has been used (extended by an intermediate stage of subadult individuals) to explain emergent Allee effects in predators that feed on structured prey populations [9].

The results show that despite the low dimension of the system, the dependence of the resulting long-term dynamics on parameter values can be rather complex. The rich behavior of our model is caused by the interaction of intra-adult competition and intra-juvenile competition. None of the two alone can generate multiple positive attractors. It is essential that the intra-juvenile competition does not only affect juvenile mortality, but also juvenile maturation (the per capita transition rate). A

density-dependent per capita juvenile mortality rate alone, even if combined with a density-dependent per capita birth rate, can neither generate Hopf bifurcation nor multiple interior equilibria. This feature seems to be related to the fact that in our ODE model the length of the juvenile period is exponentially distributed. If the length of the juvenile period is the same for all individuals (at least for those that are born at the same time), then a density-dependent per capita birth rate can lead to periodic solutions without any other nonlinear model ingredients [21, 28].

Acknowledgments. B.W.K. would like to thank Martin Boer for valuable discussions.

REFERENCES

[1] A. D. Bazykin, *Nonlinear Dynamics of Interacting Populations*, World Scientific, Singapore, 1998.

[2] A. D. Bazykin, Yu. A. Kuznetsov, and A. I. Khibnik, in *Bifurcation Diagrams of Planar Dynamical Systems*, Ser. Math. Cybernetics, Research Computing Centre, USSR Academy of Sciences, Pushchino, Moscow Region, 1985 (in Russian).

[3] H. Caswell, *Matrix Population Models, Construction, Analysis, and Interpretation*. Sinauer Associates Inc., Sunderland, MA, 2001.

[4] A. R. Champneys and Yu. A. Kuznetsov, *Numerical detection and continuation of codimension-two homoclinic bifurcations*. Internat. J. Bifur. Chaos Appl. Sci. Engrg., 4 (1994), pp. 795–822.

[5] A. R. Champneys, Yu. A. Kuznetsov, and B. Sandstede, *A numerical toolbox for homoclinic bifurcation analysis*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 6 (1996), pp. 867–887.

[6] J. M. Cushing, *An Introduction to Structured Population Dynamics*, Volume 71, Society for Industrial and Applied Mathematics, Philadelphia, 1998.

[7] J. M. Cushing, R. F. Costantino, B. Dennis, R. A. Desharnais, and S. M. Henson, in *Chaos in Ecology: Experimental Nonlinear Dynamics*, Theoretical Ecology Series, Academic Press, San Diego, 2003.

[8] A. M. de Roos, *A gentle introduction to physiologically structured population models*, in Structured-Population Models in Marine, Terrestrial, and Freshwater Systems, S. Tuljapurkar and H. Caswell, eds., Chapman & Hall, New York, 1997, pp. 119–204.

[9] A. M. de Roos, L. Persson, and H. R. Thieme, *Emergent Allee effects in top predators feeding on structured prey populations*, Proc. Roy. Soc. Lond. Ser. B, 270 (2003), pp. 611–618.

[10] O. Diekmann, M. Gyllenberg, H. Huang, M. Kirkilionis, J. A. J. Metz, and H. R. Thieme, *On the formulation and analysis of general deterministic structured population models. II. Nonlinear theory*, J. Math. Biol., 43 (2001), pp. 157–189.

[11] E. J. Doedel, A. R. Champneys, T. F. Fairgrieve, Y. A. Kuznetsov, B. Sandstede, and X. Wang, *Auto 97: Continuation and bifurcation software for ordinary differential equations*, Technical report, Concordia University, Montreal, Canada, 1997.

[12] F. Dumortier, R. Roussarie, J. Sotomayor, and H. Żoladek, *Bifurcations of Planar Vector Fields. Nilpotent Singularities and Abelian Integrals*, Springer-Verlag, Berlin, 1991.

[13] E. Freire, L. Pizarro, A. J. Rodríguez-Luis, and F. Fernández-Sánchez, *Multiparametric bifurcations in an enzyme-catalyzed reaction model*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 905–947.

[14] E. Gamero, E. Freire, and E. Ponce, *On the normal forms for planar systems with nilpotent linear parts*, in Bifurcation and Chaos: Analysis, Algorithms, Applications, R. Seydel, F. W. Schneider, T. Küpper, and H. Troger, eds., Birkhäuser, Basel, Switzerland, 1991, pp. 123–127.

[15] W. Govaerts, Yu. A. Kuznetsov, and B. Sijnave, *Numerical methods for the generalized Hopf bifurcation*, SIAM J. Numer. Anal., 38 (2000), pp. 329–346.

[16] J. Guckenheimer and P. Holmes, in *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, 2nd ed., Appl. Math. Sci. 42, Springer-Verlag, New York, 1985.

[17] W. S. C. Gurney and R. M. Nisbet, *Ecological Dynamics*, Oxford University Press, New York, 1998.

[18] A. I. Khibnik, Yu. A. Kuznetsov, V. V. Levitin, and E. V. Nikolaev, *Continuation techniques and interactive software for bifurcation analysis of ODEs and iterated maps*, Phys. D, 62 (1993), pp. 360–371.

[19] E. Knobloch, *Normal forms for bifurcations at a double zero eigenvalue*, Phys. Lett. A, 115 (1986), pp. 199–201.
[20] B. W. Kooi and F. D. L. Kelpin, *Physiologically structured population dynamics, a modeling perspective*, Comments on Theoret. Biol., 8 (2003), pp. 125–168.
[21] T. Kostova, J. Li, and M. Friedman, *Two models for competition between age classes*, Math. Biosci., 157 (1999), pp. 65–89.
[22] Yu. A. Kuznetsov, in *Elements of Applied Bifurcation Theory*, 2nd ed., Appl. Math. Sci. 112, Springer-Verlag, New York, 1998.
[23] Yu. A. Kuznetsov, *Numerical normalization techniques from all codim 2 bifurcations of equilibria in ODE's*, SIAM J. Numer. Anal., 36 (1999), pp. 1104–1124.
[24] Yu. A. Kuznetsov, *Practical computation of normal forms on center manifolds at codim 3 Bogdanov–Takens bifurcations*, in preparation.
[25] Yu. A. Kuznetsov and V. V. Levitin, *CONTENT: Integrated environment for the analysis of dynamical systems*, version 1.5, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, The Netherlands, 1997.
[26] V. I. Lukyanov, *Bifurcations of dynamical systems with a saddle-point separatrix loop*, Differential Equations, 18 (1982), pp. 1049–1059.
[27] J. A. J. Metz and O. Diekmann, in *The Dynamics of Physiologically Structured Populations*, Lect. Notes in Biomath. 68, Springer-Verlag, Berlin, 1986.
[28] R. M. Nisbet, *Delay-differential equations for structured populations*, in Structured-Population Models in Marine, Terrestrial, and Freshwater Systems, S. Tuljapurkar and H. Caswell, eds., New York, Chapman & Hall, 1997, pp. 89–118.
[29] V. P. Nozdracheva, *Bifurcation of structurally unstable separatrix loop*, Differentsial'nyye Uravneniya, 18 (1982), pp. 1551–1558.
[30] S. Ruan and W. Wang, *Dynamical behavior of an epidemic model with a nonlinear incidence rate*, J. Differential Equations, 188 (2003), pp. 135–163.
[31] M. Scheffer, S. Rinaldi, and Y. A. Kuznetsov, *Effects of fish on plankton dynamics: A theoretical analysis*, Canad. J. Fisheries and Aquatic Sci., 57 (2000), pp. 1208–1219.
[32] R. Seydel, *Practical Bifurcation and Stability Analysis-From Equilibrium to Chaos*, Springer-Verlag, New York, 1994.
[33] S. Tang and L. Chen, *Multiple attractors in stage-structured population models with birth pulses*, Bull. Math. Biol., 65 (2003), pp. 479–495.
[34] H. R. Thieme, *Mathematics in Population Biology*, Princeton University Press, Princeton, NJ, 2003.
[35] S. Tuljapurkar and H. Caswell, eds., *Structured Population Models in Marine, Freshwater, and Terrestrial Systems*, Chapman & Hall, New York, 1997.
[36] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Texts Appl. Math. 2, Springer-Verlag, Berlin, 1990.
[37] E. P. Volokitin and S. A. Treskov *Qualitative analysis of a mathematical model for the reaction of catalytic oxidation*, in Mathematical Problems in Chemical Kinetics, Novosibirsk, pp. 149–175. K. I. Zamaraev and G. S. Yablonskii, eds., "Nauka" Sibirsk. Otdel, 1989, (in Russian).
[38] D. Xiao and S. Ruan, *Bogdanov-Takens bifurcations in predator-prey systems with constant harvesting*, in Differential Equations with Applications to Biology, S. Ruan, G. S. K. Wolkowitz, and J. Wu, eds., Fields Inst. Commun. 21, AMS, Providence, RI, 1999, pp. 493–506.
[39] D. Xiao and S. Ruan, *Codimension two bifurcations in a predator-prey system with group defense*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 11 (2001), pp. 2123–2131.
[40] H. Zhu, S. A. Campbell, and G. S. K. Wolkowicz, *Bifurcation analysis of a predator–prey system with nonmonotonic functional response*, SIAM J. Appl. Math., 63 (2002), pp. 636–682.

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1366–1382

ON THE EVOLUTION OF DISPERSAL IN PATCHY LANDSCAPES*

STEPHEN KIRKLAND†, CHI-KWONG LI‡, AND SEBASTIAN J. SCHREIBER‡

Abstract. To better understand the evolution of dispersal in spatially heterogeneous landscapes, we study difference equation models of populations that reproduce and disperse in a landscape consisting of k patches. The connectivity of the patches and costs of dispersal are determined by a $k \times k$ column substochastic matrix S, where S_{ij} represents the fraction of dispersing individuals from patch j that end up in patch i. Given S, a dispersal strategy is a $k \times 1$ vector whose ith entry gives the probability p_i that individuals disperse from patch i. If all of the p_i's are the same, then the dispersal strategy is called unconditional; otherwise it is called conditional. For two competing populations of unconditional dispersers, we prove that the slower dispersing population (i.e., the population with the smaller dispersal probability) displaces the faster dispersing population. Alternatively, for populations of conditional dispersers without any dispersal costs (i.e., S is column stochastic and all patches can support a population), we prove that there is a one parameter family of strategies that resists invasion attempts by all other strategies.

Key words. population dynamics, evolution of dispersal, monotone dynamics

AMS subject classifications. 15A48, 39A11, 92D25

DOI. 10.1137/050628933

1. Introduction.

Plants and animals often live in landscapes where environmental conditions vary from patch to patch. Within patches, these environmental conditions may include abiotic factors such as light, space, and nutrient availability or biotic factors such as prey, competitors, and predators. Since the fecundity and survivorship of an individual depends on these factors, an organism may decrease or increase its fitness by dispersing across the environment. Depending on their physiology and their ability to accumulate information about the environment, plants and animals can exhibit two modes of dispersals and a variety of dispersal strategies. Plants and animals can be active dispersers that move by their own energy or passive dispersers that are moved by wind, water, or other animals. Passive dispersers alter their dispersal rates by varying the likelihood of dispersing and the time spent dispersing [20]. Dispersal strategies can vary from unconditional strategies in which the probability of dispersing from a patch is independent of the local environmental conditions to conditional strategies in which the likelihood of dispersing depends on local environmental factors. Understanding how natural selection acts on these different modes and strategies of dispersal has been the focus of much theoretical work [2, 5, 8, 10, 12, 15, 16, 17]. For instance, using coupled ordinary differential equation models for populations passively dispersing between two patches, Holt [8] showed that slower dispersing populations could always invade equilibria determined by faster

*Received by the editors April 10, 2005; accepted for publication (in revised form) November 15, 2005; published electronically April 21, 2006.
http://www.siam.org/journals/siap/66-4/62893.html

†Department of Mathematics and Statistics, University of Regina, Regina, SK S4S 0A2, Canada (kirkland@math.uregina.ca). The research of this author was supported in part by NSERC.

‡Department of Mathematics, The College of William and Mary, Williamsburg, VA 23187-8795 (ckli@math.wm.edu, sjschr@wm.edu). The research of the second author was partially supported by an NSF grant and an HK RCG grant. The second author is an honorary professor of the Heilongjiang University and an honorary professor of the University of Hong Kong. The research of the third author was partially supported by NSF grants EF-0436318 and DMS-0517987.

dispersing populations. Hastings [5] and Dockery et al. [2] considered evolution of dispersal in continuous space using reaction diffusion equations. Dockery et al. proved that for two competing populations differing only in their diffusion constant, the population with the larger diffusion constant is excluded. In contrast, McPeek and Holt [17] using a two patch model consisting of coupled difference equations found that "dispersal between patches can be favored in spatially varying but temporally constant environment, if organisms can express conditional dispersal strategies."

In this article, we consider the evolution of conditional and unconditional dispersers for a general class of multipatch difference equations. For these difference equations, individuals in each patch disperse with some probability. When these probabilities are independent of location, the population exhibits an unconditional dispersal strategy; otherwise it exhibits a conditional dispersal strategy. For dispersing individuals, the nature of the landscape determines the likelihood S_{ji} that a disperser from patch i ends up in patch j. Unlike previous studies of the evolution of unconditional and conditional dispersal [2, 5, 8, 17], we allow for an arbitrary number of patches and place no symmetry conditions on S. For active dispersers, asymmetries in S may correspond to geographical and ecological barriers that inhibit movement from one patch to another. For passive dispersers, these asymmetries may correspond to asymmetries in the abiotic or biotic currents in which they drift.

Our main goal is to determine what types of theorems can be proved about the evolution of dispersal for this general class of difference equation models. To achieve these goals, the remainder of the article is structured as follows. In section 2, we introduce the models. Under monotonicity assumptions about the growth rates, we prove that either populations playing a single dispersal strategy go extinct for all initial conditions or approach a positive fixed point for all positive initial conditions. We also introduce models of competing populations that differ only in their dispersal ability and prove a result about invasiveness. In section 3, we prove that for two competing populations of unconditional dispersers, the slower dispersing population displaces the faster dispersing population. The proof relies heavily on proving, in section 4, monotonicity of the principal eigenvalue for a one-parameter family of nonnegative matrices. In section 5, we prove that, provided there is no cost to dispersal and all patches can support a population, there is a one-parameter family of conditional dispersal strategies that resists invasion from other types of dispersal strategies. Numerical simulations suggest that these strategies can displace all other strategies, and we prove that these strategies can weakly coexist. In section 6, we discuss our findings and suggest directions for future research.

2. The models and basic results. Consider a population exhibiting discrete reproductive and dispersal events and living in an environment consisting of k patches. The vector of population densities is given by $x = (x_1, \ldots, x_k)^T \in \mathbf{R}_+^k$, where $\mathbf{R}_+^k$ is the nonnegative cone of $\mathbf{R}^k$. To describe reproduction and survival in each patch, let $\lambda_i : \mathbf{R}_+ \to \mathbf{R}_+$ denote the per-capita growth rate of the population in the ith patch as a function of the population density in the ith patch. For these per-capita growth rates we make the following assumptions.

A1: λ_i are positive continuous decreasing functions.

A2: $\lim_{x_i \to \infty} \lambda_i(x_i) < 1$.

A3: $x_i \mapsto x_i \lambda_i(x_i)$ is increasing.

Assumption A1 corresponds to the population exhibiting increasing levels of intra specific competition or interference as population densities increase. Assumption A2 implies that at high densities the population tends to decrease in size. Assumption A3

implies that the population does not exhibit overcompensating density dependence: higher densities in the current generation yield higher densities in the next generation. Many models in the population ecology literature satisfy these three assumptions. For instance, see the Beverton–Holt model [1] in which $\lambda_i(x_i) = \frac{a_i}{1+b_i x_i}$ and the Ivlev model [14] in which $\lambda_i(x_i) = a_i(1 - \exp(-b\, x_i))$.

To describe dispersal between patches, we assume that each individual in patch i disperses with a probability p_i and S_{ji} is the probability that a dispersing individual from patch i arrives in patch j. About the matrix S we make the following assumption.

A4: S is a $k \times k$ primitive column substochastic matrix.

S can be column stochastic if all dispersing individuals migrate successfully or substochastic if some dispersing individuals experience mortality. The primitive assumption ensures that individuals (possibly after several generations) can move from any patch to any patch. S characterizes how connected the landscape is for dispersing individuals. For example, for a fully connected metapopulation, S could be the matrix whose entries all equal $\frac{1}{k}$; i.e., an individual is equally likely to end up in any patch after dispersing. Alternatively, in a landscape with a one-dimensional lattice structure with individuals able only to move to neighboring patches in one time step S is a column substochastic tridiagonal matrix that is primitive, provided it has a positive entry on the diagonal. From these p and S, the following matrix describes how the population redistributes itself across the environment in one time step:

$$S_p = I - \operatorname{diag}(p) + S \operatorname{diag}(p),$$

where $\operatorname{diag}(p)$ denotes a diagonal matrix with diagonal entries $p_1, \ldots, p_k$.

If a census of the population is taken before reproduction and after dispersal, then the dynamics of the population are given by

$$x' = S_p \Lambda(x) x =: F(x), \tag{1}$$

where x' denotes the population state in the next time step and $\Lambda(x)$ is the $k \times k$ diagonal matrix whose ith diagonal entry equals $\lambda_i(x_i)$.

Our first result characterizes the global dynamics of (1). To state this result, let $F^n(x)$ denote F composed with itself n times. Given $x, y \in \mathbf{R}^k_+$, we write $x \geq y$ if $x_i \geq y_i$ for all $1 \leq i \leq k$, $x > y$ if $x \geq y$ and $x \neq y$, and $x \gg y$ if $x_i > y_i$ for all $1 \leq i \leq k$. For a matrix A, let $r(A)$ denote the spectral radius of A.

THEOREM 2.1. *Assume that Assumptions* A1–A4 *hold and* $p \in (0,1]^k$. *If* $r(S_p \Lambda(0)) \leq 1$, *then*

$$\lim_{n\to\infty} F^n(x) = 0$$

for all $x \geq 0$. *Alternatively, if* $r(S_p\Lambda(0)) > 1$, *then there exists a fixed point* $\hat{x} \gg 0$ *for* F *such that*

$$\lim_{n\to\infty} F^n(x) = \hat{x}$$

for all $x > 0$.

Proof. Let $A(x) = S_p\Lambda(x)$. Assumptions A1, A4, and $p \gg 0$ imply that $A(x)$ is primitive for all $x \geq 0$. Assumption A3 implies that $F(x) \geq F(y)$ (resp., $F(x) > F(y)$, $F(x) \gg F(y)$) whenever $x \geq y$ (resp., $x > y$, $x \gg y$). In other words, F is a strongly monotone map.

Suppose that $r(A(0)) \leq 1$. Let $w^T \gg 0$ be a left Perron vector of $A(0)$, i.e., $r(A(0))w^T = w^T A(0)$. Define the function $L : \mathbf{R}^k_+ \to \mathbf{R}_+$ by $L(x) = w^T x$. For $x > 0$, Assumption A1 implies that $w^T A(0) \gg w^T A(x)$. Hence, for any $x > 0$,

$$\begin{aligned} L(F(x)) &= w^T A(F(x))x \\ &= w^T A(0)x + w^T(A(F(x)) - A(0))x \\ &< r(A(0))w^T x \leq L(x). \end{aligned}$$

Since L is strictly decreasing along nonzero orbits of F, $L(0) = 0$, and $L(x) > 0$ for $x > 0$, it follows that $\lim_{n\to\infty} F^n(x) = 0$ for all $x \geq 0$.

Suppose $r(A(0)) > 1$. First, we show that there exists a positive fixed point $\hat{x}$. Let $v \gg 0$ be a right Perron eigenvector for $A(0)$, i.e., $A(0)v = r(A(0))v$. Since $A(0)v \gg v$, continuity of $A(x)$ implies that there exists $\epsilon > 0$ such that $A(y)y \gg y$, where $y = \epsilon v$. Since $F(x) \gg F(y)$ whenever $x \gg y$, induction implies $y \ll F(y) \ll F^2(y) \ll F^3(y) \ll \cdots$. Assumption A2 implies that the increasing sequence $F^n(y)$ is bounded. Hence, there exists $\hat{x}$ such that $\lim_{n\to\infty} F^n(y) = \hat{x}$. Continuity of F implies that $F(\hat{x}) = \hat{x}$. Second, we show that $\lim_{n\to\infty} F^n(x) = \hat{x}$ whenever $\hat{x} > x > 0$. In particular, $\hat{x}$ is a unique positive fixed point. Let w^T be the left Perron eigenvector of $A(\hat{x})$ that satisfies $w^T\hat{x} = 1$. Since $\hat{x}$ is a positive fixed point, $r(A(\hat{x})) = 1$. Define $L : \mathbf{R}^k_+ \to \mathbf{R}_+$ by $L(x) = w^T x$. Let $\hat{x} > x > 0$. Then $\hat{x} > F(x) > 0$ and

$$\begin{aligned} L(F(x)) &= w^T A(F(x))x \\ &= w^T A(\hat{x})x + w^T(A(F(x)) - A(\hat{x}))x \\ &> r(A(\hat{x}))w^T x = L(x). \end{aligned}$$

Hence, $L(x), L(F(x)), L(F^2(x)), \ldots$ is a positive increasing sequence bounded above by $L(\hat{x}) = 1$. Since $L(x) < 1$ for all $x < \hat{x}$, it follows that $\lim_{n\to\infty} F^n(x) = \hat{x}$ for all $0 < x < \hat{x}$. Third, it can be shown similarly that $\lim_{n\to\infty} F^n(x) = \hat{x}$ for all $x > \hat{x}$. Fourth, consider any $x \gg 0$. Choose $\overline{x} > x$ such that $\overline{x} > \hat{x}$ and choose $\underline{x} < x$ such that $0 < \underline{x} < \hat{x}$. Since $F^n(\underline{x}) < F^n(x) < F^n(\overline{x})$ for all n and $\lim_{n\to\infty} F^n(\underline{x}) = \lim_{n\to\infty} F^n(\overline{x}) = \hat{x}$, continuity of F implies that $\lim_{n\to\infty} F^n(x) = \hat{x}$. Finally, consider any $x > 0$. Assumptions A3–A4 imply that there exists $n \geq 1$ such that $F^n(x) \gg 0$. Hence, $\lim_{n\to\infty} F^n(x) = \hat{x}$. □

To understand the evolution of dispersal, we shall consider two populations that differ only in their dispersal ability. Let $x, y \in \mathbf{R}^k_+$ denote the vector of densities of the two populations and $p, \widetilde{p}$ denote their dispersal strategies. Since the populations differ only in their dispersal abilities, their dynamics are given by

$$\begin{aligned} x' &= S_p\Lambda(x+y)x =: G_1(x,y), \\ y' &= S_{\widetilde{p}}\Lambda(x+y)y =: G_2(x,y). \end{aligned} \tag{2}$$

From Assumption A2 it follows that (2) is dissipative i.e., there exists a compact set K such that for any $(x, y) \geq (0, 0)$, $G^n(x, y) \in K$ for n sufficiently large. Regarding the dynamics of (2) near equilibria, we need the following result about invasiveness. Since we have not assumed that $G(x, y)$ is continuously differentiable, this result does not follow immediately from the standard unstable manifold theory.

PROPOSITION 2.2. *Assume that $p, \widetilde{p} \in (0,1]^k$, S and Λ satisfy Assumptions* A1–A4, *and $r(S_p\Lambda(0)) > 1$. Let $\hat{x} \gg 0$ be the fixed point satisfying $G_1(\hat{x}, 0) = (\hat{x}, 0)$. If $r(S_{\widetilde{p}}\Lambda(\hat{x})) > 1$, then there exists a neighborhood $U \subset \mathbf{R}^k_+ \times \mathbf{R}^k_+$ of $(\hat{x}, 0)$ such that for any $(x, y) \in U$ with $y > 0$, $G^n(x, y) \notin U$ for some $n \geq 1$.*

Proof. Let $A(x) = S_{\widetilde{p}}\Lambda(x)$. Assume that $r(A(\hat{x})) > 1$. Let $w^T \gg 0$ be a left Perron eigenvector of $A(\hat{x})$. Since $w^T A(\hat{x}) \gg w^T$, continuity of $x \mapsto A(x)$ implies that there exists a compact neighborhood $U \subset \mathbf{R}^k_+ \times \mathbf{R}^k_+$ of $(\hat{x}, 0)$ and $c > 1$ such that $w^T A(x + y) \gg cw^T$ for all $(x, y) \in U$. Define $L : \mathbf{R}^k_+ \times \mathbf{R}^k_+ \to \mathbf{R}_+$ by $L(x, y) = w^T y$. Let (x, y) be in U with $y > 0$. We have $L(G(x, y)) = w^T A_{\widetilde{p}}(x + y)y > cL(x, y)$. Hence, if $(x, y), \ldots, G^n(x, y) \in U$, then $L(G^n(x, y)) > c^n w^T y$. Since U is compact and $y > 0$, it follows that there exists $n \geq 1$ such that $G^n(x, y) \notin U$. ∎

3. The slower unconditional disperser wins. In this section, we consider only an unconditional dispersal strategy p: a strategy that satisfies $p_1 = \cdots = p_k$ for some common value d. Equivalently, $p = d\,\mathbf{1}$, where $\mathbf{1} = (1, \ldots, 1)$. Our key result is the following theorem concerning the monotonicity of the dominant eigenvalue with respect to the parameter d.

THEOREM 3.1. *Let S be an irreducible column substochastic matrix and Λ be a diagonal matrix. If Λ is not a scalar matrix, then $d \mapsto r(((1 - d)I + dS)\Lambda)$ is decreasing on $[0, 1]$.*

The proof of Theorem 3.1 is given in section 4, where we also characterize the function $d \mapsto r(S_{d\mathbf{1}})$ when S is reducible. The following corollary follows immediately from Theorems 2.1 and 3.1.

COROLLARY 3.2. *Assume that F, S, and $\Lambda(x)$ satisfy Assumptions* A1–A4 *and $p = d\,\mathbf{1}$. Then there exists $d^* \geq 0$ such that we have the following.*

Persistence: *If $d \in [0, d^*)$, then there exists $\hat{x} \gg 0$ satisfying $\lim_{n\to\infty} F^n(x) = \hat{x}$ for all $x \gg 0$.*

Extinction: *If $d \in [d^*, 1]$, then $\lim_{n\to\infty} F^n(x) = 0$ for all $x \geq 0$.*

Moreover, $d^* = 0$ if $\max_i \lambda_i(0) \leq 1$, $d^* \in (0, 1)$ if $\max_i \lambda_i(0) > 1$ and $r(S\Lambda(0)) < 1$, and $d^* \geq 1$ if $r(S\Lambda(0)) \geq 1$.

Corollary 3.2 implies that whenever $r(S\Lambda(0)) < 1$, unconditional dispersers have a critical dispersal rate below which the population persists and above which the population is deterministically driven to extinction.

To characterize the dynamics of competing unconditional dispersers, we need an additional assumption on (2) to avoid degenerate cases. Let $v \gg 0$ be a right Perron eigenvector of S, i.e., $Sv = r(S)v$. We make the following assumption.

A5: $\Lambda(tv)$ is not a scalar matrix for any $t \geq 0$.

This assumption assures that the model exhibits a minimal amount of spatial heterogeneity in the per-capita growth rates at fixed points.

THEOREM 3.3. *Let $G = (G_1, G_2)$ satisfy Assumptions* A1–A5. *Assume that $p = d\mathbf{1}$, and $\widetilde{p} = \widetilde{d}\mathbf{1}$, where $0 < d < \widetilde{d} \leq 1$. If $r(S_p\Lambda(0)) > 1$, then for all $x > 0$ and $y \geq 0$,*

$$\lim_{n\to\infty} G^n(x, y) = (\hat{x}, 0),$$

where $\hat{x}$ is the positive fixed point of $x \mapsto G_1(x, 0)$.

Theorem 3.3 implies that the slower disperser always displaces the faster disperser. This occurs despite the fact that the faster disperser is initially able to establish itself more rapidly, as illustrated in Figures 1 and 2.

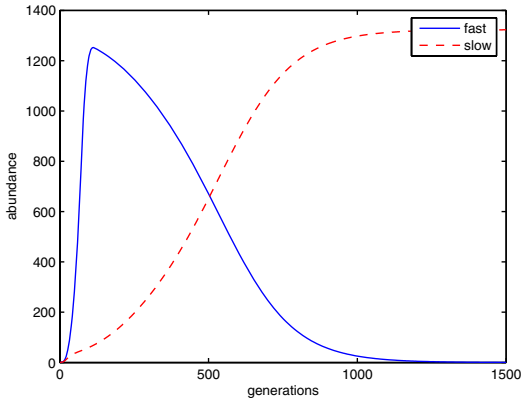
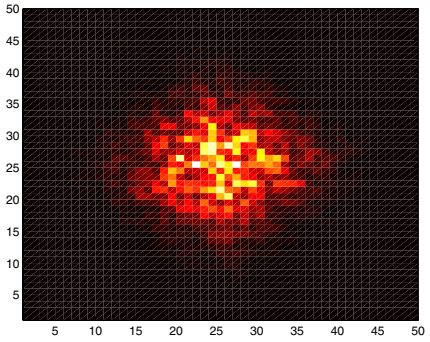
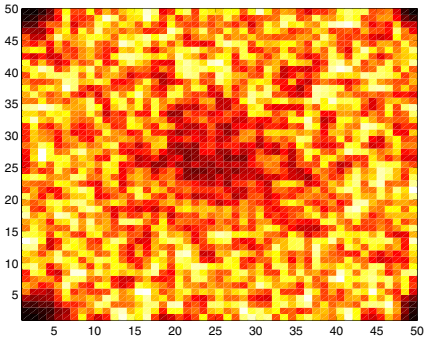


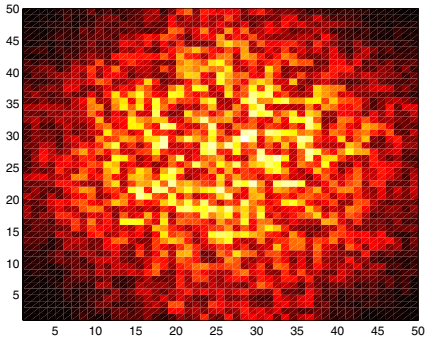
FIG. 1. *A simulation of* (2) *with* $k = 50 \times 50$ *(i.e., a two-dimensional spatial grid),* $\lambda_i(x_i) = \frac{a_i}{1+x_i}$ *with* a_i *randomly chosen from* $[1, 2]$*,* $d = 0.2$*,* $\widetilde{d} = 0.3$*, and* S *given by movement with equal likelihood to east, west, north, and south, and periodic boundary conditions. The initial condition corresponds to a density one of both populations in the center patch. The dotted and solid curves correspond to the abundances of the slower and faster dispersing populations, respectively.*



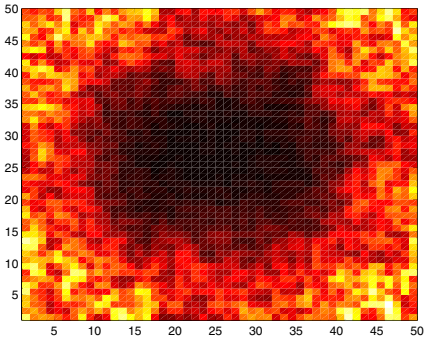
(a) slower disperser at generation 100



(b) faster disperser at generation 100



(c) slower disperser at generation 500



(d) faster disperser at generation 500

FIG. 2. *Spatial distributions of the slower disperser in* (a) *and* (c) *and the faster dispersers in* (b) *and* (d)*. The model and parameters are as in Figure* 1*. Darker (resp., lighter) shading correspond to lower (resp., higher) densities.*

Proof. The proof of this theorem relies on a result of Hsu, Smith, and Waltman [11, Theorem A] and Theorems 2.1 and 3.1. Let $A_d(x) = S_{d\mathbf{1}}(x)$. We start the proof with an important implication of Assumption A5. Suppose (x, y) satisfies $G(x, y) = (x, y)$. We claim that $\Lambda(x + y)$ is not a scalar matrix. Indeed, suppose to the contrary that $\Lambda(x + y) = tI$ for some $t > 0$. Then

$$x = S_p\Lambda(x+y)x = (1-d)tx + dtSx,$$
$$y = S_{\widetilde{p}}\Lambda(x+y)y = (1-\widetilde{d})ty + \widetilde{d}tSy.$$

Consequently, x and y (and hence $x+y$) are scalar multiples of v. Since this contradicts Assumption A5, $\Lambda(x + y)$ is not a scalar matrix for any fixed point (x, y) of G.

Assuming that $r(A_d(0)) > 1$, Theorem 2.1 implies that $x \mapsto G_1(x, 0)$ has a unique positive fixed point $\hat{x}$ that is globally stable. We prove the theorem in two cases. In the first case, assume that $r(A_{\widetilde{d}}(0)) > 1$. Theorem 2.1 implies that there is a unique $\hat{y} \gg 0$ such that $G(0, \hat{y}) = (0, \hat{y})$ and $\lim_{n\to\infty} G^n(0, y) = (0, \hat{y})$ whenever $y \gg 0$. To employ Theorem A in [11] we need to verify two things: G has no positive fixed point and $(0, \hat{y})$ is unstable. First, suppose to the contrary there exists $x \gg 0$ and $y \gg 0$ such that $G(x, y) = (x, y)$. Then $x = A_d(x + y)x$, $y = A_{\widetilde{d}}(x + y)y$, and $r(A_d(x + y)) = 1$. Since $\Lambda(x + y)$ is not a scalar matrix, Theorem 3.1 implies that $1 = r(A_d(x + y)) > r(A_{\widetilde{d}}(x + y)) = 1$. Hence, there can be no positive fixed point. Second, to show that $(0, \hat{y})$ is unstable, we use Theorem 3.1, which implies that $1 = r(A_{\widetilde{d}}(\hat{y})) < r(A_p(\hat{y}))$, and apply Proposition 2.2. Applying Theorem A of [11] implies that $\lim_{n\to\infty} G^n(x, y) = (\hat{x}, 0)$ whenever $x \gg 0$ and $y \gg 0$.

Suppose that $r(A_{\widetilde{d}}(0)) \leq 1$. Let $w^T \gg 0$ be a left Perron vector of $A_{\widetilde{d}}(0)$. Define the function $L : \mathbf{R}^k_+ \to \mathbf{R}_+$ by $L(y) = w^T y$. Let $\pi(x, y) = y$. Since $L(G^n(x, y))$ is strictly decreasing whenever $y > 0$, $L(0) = 0$, and $L(y) > 0$ for $y > 0$, it follows that $\lim_{n\to\infty} \pi(G^n(x, y)) = 0$ for all $x \geq 0$. Hence, for any $(x, y) \in \mathbf{R}^k_+ \times \mathbf{R}^k_+$, the limit points of $G^n(x, y)$ as $n \to \infty$ lie in $\mathbf{R}^k_+ \times \{0\}$. By Theorem 1.8 in [18], the closure of these limit points form a connected chain recurrent set (see [18] for the definition). Since the only connected chain recurrent sets in $\mathbf{R}^k_+ \times \{0\}$ are $(0, 0)$ and $(\hat{x}, 0)$, instability of $(0, 0)$ implies that $\lim_{n\to\infty} G^n(x, y) = (\hat{x}, 0)$ whenever $x > 0$. ∎

4. Proof of Theorem 3.1. We begin with the following preliminary result.

LEMMA 4.1. *Let v and w^T be positive k-vectors so that $w^T v = 1$. Let $\mathbf{P}$ be the polytope of nonnegative matrices A such that $w^T A = w^T$ and $Av = v$. For each $A \in \mathbf{P}$, let D_A denote the diagonal matrix of column sums of A. Then*

$$\min\{w^T D_A v | A \in \mathbf{P}\} = 1.$$

A matrix $A \in \mathbf{P}$ attains the minimum value for $w^T D_A v$ if and only if $D_A = I$.

Proof. Without loss of generality, assume that $w^T = (w_1, \ldots, w_k)$ is such that $w_1 \leq \cdots \leq w_k$. Note also that if all of the entries in w^T are equal, then each matrix in $\mathbf{P}$ is a column stochastic matrix, and the statement of the lemma follows immediately. We suppose henceforth that w^T has at least two distinct entries.

Suppose that $A \in \mathbf{P}$ and that there are indices i, j, p, q satisfying the following conditions:

$$w_i < w_j, w_p < w_q, \quad \text{and} \quad a_{ip}, a_{jq} > 0. \tag{3}$$

We claim that in this case, the matrix A does not satisfy

$$w^T D_A v \leq w^T D_B v \quad \text{for all } B \in \mathbf{P}. \tag{4}$$

To see the claim, note that from (3), it follows that for sufficiently small $\epsilon > 0$, the matrix

$$\hat{A} = A + \epsilon(-e_i/w_i + e_j/w_j)(e_p/v_p - e_q/v_q)^T$$

is nonnegative, and satisfies $w^T\hat{A} = w^T$ and $\hat{A}v = v$, so that $\hat{A} \in \mathbf{P}$. Further,

$$D_{\hat{A}} = D_A + \epsilon\frac{w_j - w_i}{w_i w_j}\,\mathrm{diag}\left(\frac{-e_p}{v_p} + \frac{e_q}{v_q}\right),$$

so that

$$w^T D_{\hat{A}} v = w^T D_A v - \epsilon\frac{(w_j - w_i)(w_p - w_q)}{w_i w_j} < w^T D_A v.$$

Thus $w^T D_A v$ does not yield the minimum, as claimed.

Suppose the minimum entry in w is repeated a times, i.e., $w_1 = \cdots = w_a < w_{a+1}$. Partition out the first a entries of w^T, to write w^T as $[w_1\mathbf{1}^T|\tilde{w}^T]$, and partition v conformally as

$$v = \left[\frac{\hat{v}}{\tilde{v}}\right].$$

Let $A \in \mathbf{P}$ satisfy (4). Suppose first that there are indices i and p with $1 \leq i \leq a$ and $a+1 \leq p$, such that $a_{ip} > 0$. Since A is a minimizer, we see from the claim above that for any indices j, q with $j \geq a+1$ and $1 \leq q \leq a$, we must have $a_{jq} = 0$. But then A has the form

$$A = \left[\begin{array}{c|c} A_1 & X \\ \hline 0 & A_2 \end{array}\right],$$

where A_1 is $a \times a$. From the facts that $w^T A = w^T$ and that the first a entries of w^T are equal and the partitioned form for A, we find that $\mathbf{1}^T A_1 = \mathbf{1}^T$. Also, $A_1\hat{v} + X\tilde{v} = \hat{v}$, so that $\mathbf{1}^T(A_1\hat{v} + X\tilde{v}) = \mathbf{1}^T\hat{v}$. Since $\mathbf{1}^T A_1 = \mathbf{1}^T$, we conclude that $X = 0$, a contradiction.

Consequently, we conclude that for any indices i and p with $1 \leq i \leq a$ and $a+1 \leq p$, we must have $a_{ip} = 0$. Thus we see that A has the form

$$A = \left[\begin{array}{c|c} A_1 & 0 \\ \hline Y & A_2 \end{array}\right],$$

where A_1 is $a \times a$ and $A_1\hat{v} = \hat{v}$. Using the fact that $w^T A = w^T$, we thus find that $w_1\mathbf{1}^T A_1 + \tilde{w}^T Y = w_1\mathbf{1}^T$. Hence we have $w_1\mathbf{1}^T A_1\hat{v} + \tilde{w}^T Y\hat{v} = w_1\mathbf{1}^T\hat{v}$, from which we deduce that $Y = 0$.

We conclude that if $A \in \mathbf{P}$ satisfies (4), then A can be written as

$$\left[\begin{array}{c|c} A_1 & 0 \\ \hline 0 & A_2 \end{array}\right],$$

where A_1 is column stochastic. The lemma is now readily established by a deflation argument. ∎

Our next result lends some insight into the irreducible case.

LEMMA 4.2. *Suppose that A is an irreducible nonnegative matrix, and let D_A be the diagonal matrix of column sums of A. Let Λ be a diagonal matrix such that $\Lambda \geq D_A$. For each $d \in [0,1]$ let $h(d) = r((1-d)\Lambda + dA)$. Then for any $d \in (0,1), h'(d) \leq 0$, with equality holding if and only if $\Lambda = D_A = aI$ for some $a > 0$. In that case, $h(d) = r(A) = a$ for each $d \in [0,1]$.*

Proof. Throughout, we suppose without loss of generality that $r(A) = 1$.

First, suppose that A is a primitive matrix; we claim that in this case, $h'(1) \leq 0$ with equality holding if and only if $\Lambda = D_A = I$. Let v be a right Perron vector for A. Since A is primitive, its spectral radius is a simple eigenvalue that strictly dominates the modulus of any other eigenvalue; it follows that in a sufficiently small neighborhood of 1, $h(d)$ is an eigenvalue of $(1-d)\Lambda + dA$ that is differentiable in d. For d in such a neighborhood of 1, let $w(d)^T$ be a left $h(d)$-eigenvector of $(1-d)\Lambda + dA$, normalized so that $w(d)^T v = 1$. Since $Av = v$, we have

$$\begin{aligned} h(d) &= w(d)^T((1-d)\Lambda + dA)v \\ &= (d-1)(w(d)^T(A-\Lambda)v) + w(d)^T Av \\ &= (d-1)(1 - w(d)^T \Lambda v) + 1. \end{aligned}$$

Since $\lim_{d\to 1} w(d)^T = w^T$, it follows that

$$\begin{aligned} \lim_{d\to 1} \frac{h(d)-h(1)}{d-1} &= \lim_{d\to 1}(1 - w(d)^T \Lambda v) = 1 - w^T \Lambda v \\ &= -(w^T D_A v - 1) - (w^T(\Lambda - D_A)v). \end{aligned}$$

Since $\Lambda \geq D_A$, we have $w^T(\Lambda - D_A)v \geq 0$, and by Lemma 4.1, we have $w^T D_A v - 1 \geq 0$, so certainly $h'(1) \leq 0$. Further, we see that $h'(1) = 0$ if and only if $w^T D_A v = 1$ and $w^T(\Lambda - D_A)v = 0$. It now follows from Lemma 4.1 that the former holds if and only if $D_A = I$, and since w^T and v are positive vectors, we see that the latter holds if and only if $\Lambda = D_A$. This completes the proof of the claim.

Next, suppose that A is an irreducible nonnegative matrix, and fix $d \in (0,1)$. Observe that the matrix $B = (1-d)\Lambda + dA$ is primitive and that $\Lambda \geq D_B$. For each $c \in [0,1]$, let $k(c) = r((1-c)\Lambda + cB)$, and note that $k(c) = h(cd)$. Applying the claim above to the function k, we see that $k'(1) \leq 0$, with equality holding if and only if $\Lambda = D_B = I$. But from the chain rule, we find that $k'(1) = dh'(d)$, so that $h'(d) \leq 0$, with equality if and only if $\Lambda = D_B = I$. That last condition is readily seen to be equivalent to $\Lambda = D_A = I$.

Finally, we note that if $\Lambda = D_A = I$, it is straightforward to see that for each $d \in [0,1]$, the matrix $(1-d)\Lambda + dA$ is column stochastic, so that $h(d) = 1 = r(A)$ for all such d. ∎

The following, which evidently yields Theorem 3.1 immediately, follows from Lemma 4.2.

COROLLARY 4.3. *Suppose that A is an irreducible nonnegative matrix, and let D_A be the diagonal matrix of column sums of A. Let Λ be a diagonal matrix such that $\Lambda \geq D_A$. For each $d \in [0,1]$ let $h(d) = r((1-d)\Lambda + dA)$. Then either*

(a) *$h(d)$ is a strictly decreasing function of $d \in [0,1]$ or*

(b) *for some $a > 0, \Lambda = D_A = aI$ and $h(d) = a$ for each $d \in [0,1]$.*

We have the following generalization of Corollary 4.3.

THEOREM 4.4. *Let S be a column substochastic matrix and Λ be a diagonal matrix with positive diagonal entries. Define the function $f(d) = r(((1-d)I + dS)\Lambda)$ for $d \in [0,1]$. Then there is a $\hat{d} \in [0,1]$ such that f is strictly decreasing on $[0,\hat{d}]$ and f is constant on $[\hat{d},1]$. Specifically, let P be a permutation matrix such that*

$$P^T S P = \left[\begin{array}{cccc|c} S_1 & 0 & \dots & 0 & X_1 \\ 0 & S_2 & \dots & 0 & X_2 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & S_k & X_k \\ \hline 0 & 0 & \dots & 0 & S_{k+1} \end{array}\right], \quad \textit{and} \quad P^T \Lambda P = \begin{bmatrix} \Lambda_1 & & & \\ & \Lambda_2 & & \\ & & \ddots & \\ & & & \Lambda_{k+1} \end{bmatrix},$$

where (i) *$P^T S P$ and $P^T \Lambda P$ are partitioned conformally,* (ii) *for each $i = 1, \dots, k$, S_i is an irreducible column stochastic matrix, and* (iii) *S_{k+1} is a column substochastic matrix such that $r(S_{k+1}) < 1$. (Note that such a permutation matrix P exists and that one part of this partitioning of $P^T S P$ may be vacuous.) Let $r(\Lambda) = \rho$. Exactly one of the following cases holds.*

(a) *For some $i = 1, \dots, k$, $\Lambda_i = \rho I$. In this case, $f(d) = \rho$ for all $d \in [0,1]$.*

(b) *There is an index $i_0 = 1, \dots, k$ and an $a < \rho$ such that $\Lambda_{i_0} = aI$ and in addition, for each $j = 1, \dots, k+1$, we have that either $r(S_j \Lambda_j) < a$ or $r(((1-d)I + dS_j)\Lambda_j) = a$ for all $d \in [0,1]$. In this case, there is a $\hat{d} \in (0,1)$ such that $f(d)$ is a strictly decreasing function of d for $d \in [0,\hat{d}]$, while for each $d \in [\hat{d},1]$, $f(d) = a$.*

(c) *If $\Lambda_i \neq \rho I$ for $i = 1, \dots, k$ and there is no index i_0 and value a satisfying the hypotheses of* (b)*, then $f(d)$ is strictly decreasing for $d \in [0,1]$.*

Proof. Throughout, we assume without loss of generality that $\rho = 1$. First, note that $f(d) = \max\{r(((1-d)I + dS_i)\Lambda_i) : i = 1, \dots, k+1\}$. Further, since $r(S_{k+1}) < 1$ it follows that no principal submatrix of S_{k+1} (including the entire matrix S_{k+1} itself) can have all of its column sums equal to 1; we then deduce from Corollary 4.3 that $r(((1-d)I + S_{k+1})\Lambda_{k+1})$ is strictly decreasing as a function of $d \in [0,1]$. Note further that if none of $\Lambda_1, \dots, \Lambda_k$ is a scalar matrix, then for each $i = 1, \dots, k$ the function $r(((1-d)I + dS_i)\Lambda_i)$ is strictly decreasing in d, from which we conclude that $f(d)$ is strictly decreasing.

Suppose next that for some $i = 1, \dots, k$, we have $\Lambda_i = I$. From Corollary 4.3 we see that $r(((1-d)I + dS_i)\Lambda_i) = 1$ for all $d \in [0,1]$, and we conclude readily that $f(d) = 1$ for all $d \in [0,1]$.

It remains only to consider the case that $\Lambda_i \neq I$ for $i = 1, \dots, k$ but that for one or more indices $i = 1, \dots, k$, Λ_i is a scalar matrix. For concreteness, we suppose that $\Lambda_i = a_i I$ for $i = 1, \dots, j$ and that for $i = j+1, \dots, k$, Λ_i is not a multiple of the identity matrix. Again without loss of generality, we can assume that $1 > a_1 \geq \cdots \geq a_j$. In this situation, we find that for each $i = 1, \dots, j$, $r(((1-d)I + dS_i)\Lambda_i) = a_i$, while for each $i = j+1, \dots, k+1$, $r(((1-d)I + dS_i)\Lambda_i)$ is a strictly decreasing function of d. It follows from the above considerations that $f(d) = \max\{a_1, r(((1-d)I + dS_{j+1})\Lambda_{j+1}), \dots, r(((1-d)I + dS_{k+1})\Lambda_{k+1})\}$.

Evidently two cases arise: either $\max\{r(S_{j+1}\Lambda_{j+1}), \dots, r(S_{k+1}\Lambda_{k+1})\} \geq a_1$ or $\max\{r(S_{j+1}\Lambda_{j+1}), \dots, r(S_{k+1}\Lambda_{k+1})\} < a_1$. In the former case we see that in fact $f(d) = \max\{r(((1-d)I + dS_{j+1})\Lambda_{j+1}), \dots, r(((1-d)I + dS_{k+1})\Lambda_{k+1})\}$ for all $d \in [0,1]$, from which we conclude that f is strictly decreasing in d. Now suppose that the latter

case holds. Since $a_1 < 1$, we see that when d is near to 0, $f(d) = \max\{r(((1-d)I + dS_{j+1})\Lambda_{j+1}), \ldots, r(((1-d)I + dS_{k+1})\Lambda_{k+1})\} > a_1$. Thus, from the intermediate value theorem it follows that there is a value $\hat{d} \in (0,1)$ such that $\max\{r(((1-d)I + dS_{j+1})\Lambda_{j+1}), \ldots, r(((1-d)I + dS_{k+1})\Lambda_{k+1})\} \geq a_1$ for $d \in [0, \hat{d}]$ and $\max\{r(((1-d)I + dS_{j+1})\Lambda_{j+1}), \ldots, r(((1-d)I + dS_{k+1})\Lambda_{k+1})\} < a_1$ for $d \in [\hat{d}, 1]$. It now follows that $f(d)$ is strictly decreasing for $d \in [0, \hat{d}]$ and $f(d) = a_1$ for $d \in [\hat{d}, 1]$. □

5. Competing conditional dispersers. In this section, we extend our study to conditional dispersers in which p need not be a constant vector. The following theorem coupled with Proposition 2.2 indicates which dispersal strategies are subject to invasion by other dispersal strategies.

THEOREM 5.1. *Assume that $\Lambda(x)$ and S satisfy Assumptions* A1–A4, *$p \in (0,1]^k$, and $r(S_p\Lambda(0)) > 1$. Let $\hat{x} \gg 0$ be the unique positive fixed point of F, and let $v \gg 0$ be a right Perron vector for S. Then $r(S_{\widetilde{p}}\Lambda(\hat{x})) \leq 1$ for all $\widetilde{p} \in (0,1]^k$ if and only if $\lambda_i(0) > 1$ for all i, S is column stochastic, and*

$$p = t\left(\Lambda^{-1}(I)\right)^{-1} v \tag{5}$$

for some $t \in (0, 1/\max\{\Lambda^{-1}(I)^{-1}v\}]$. Moreover, if p is given by (5), *then $\Lambda(\hat{x}) = I$.*

In our proof of Theorem 5.1, we show that if either S is strictly substochastic or p is not given by (5), then there are strategies $\widetilde{p}$ arbitrarily close to p that can invade, i.e., $r(S_{\widetilde{p}}\Lambda(\hat{x})) > 1$. When S is stochastic and p is given by (5), we also show that $\Lambda(\hat{x}) = I$ and, consequently, $r(S_{\widetilde{d}}\Lambda(\hat{x})) = 1$ for all $\widetilde{p} \in [0,1]^k$. The populations playing one of these strategies exhibit an *ideal-free distribution at equilibrium* [3]; i.e., the per-capita fitness in all occupied patches are equal. Theorem 5.1 suggests the possibility that strategies of the form (5) can displace all other strategies. By [11, Theorem A] a sufficient condition for this displacement is verifying that (5) can invade any strategy $\widetilde{p}$ not given by (5) and cannot coexist at equilibrium with strategy $\widetilde{d}$. This turns out not to be true in general. For example, let $\lambda_i(x_i)$ with $i = 1, 2$ be functions such that $\lambda_1(1.2) = \lambda_2(1) = 1$, $\lambda_1(1.19) = \frac{20}{9+\sqrt{41}} \approx 1.29844$, $\lambda_2(9.52/(3+\sqrt{41})) = \frac{10}{9+\sqrt{41}} \approx 0.642919$, where $9.52/(3+\sqrt{41}) \approx 1.01234$, and Assumptions A1–A3 are satisfied. Define

$$S = \begin{pmatrix} 0.5 & 0.6 \\ 0.5 & 0.4 \end{pmatrix},$$

which has right Perron vector

$$v = \begin{pmatrix} 1 \\ 5/6 \end{pmatrix}.$$

Then $p = \mathbf{1}$ is a strategy of the form (5). Define

$$\widetilde{p} = \begin{pmatrix} 0.8 \\ 2/3 \end{pmatrix}.$$

The unique positive fixed point of $y \mapsto S_{\widetilde{p}}\Lambda(y)y = G(0,y)$ is by construction given by

$$\hat{y} = \begin{pmatrix} 1.19 \\ \frac{9.52}{3+\sqrt{41}} \end{pmatrix}.$$

Since a computation reveals that

$$r(S\Lambda(\widehat{y})) = 0.993735\ldots < 1 = r(S_{\widetilde{d}}\Lambda(\widehat{y})),$$

the strategy $p = \mathbf{1}$ cannot invade and displace the strategy $\widetilde{p}$. Hence, for a general $\Lambda(x)$, we cannot expect that strategies of the form (5) will displace all other strategies. However, extensive simulations with the Beverton–Holt growth functions (i.e., $\lambda_i(x_i) = \frac{a_i}{1+b_i x_i}$) suggest that the strategies given by (5) can displace any other strategy (see Figure 3). Thus we make the following conjecture.

CONJECTURE 5.1. *If $\lambda_i(x_i) = \frac{a_i}{1+b_i x_i}$, S is primitive and column stochastic, p is given by* (5)*, $\widetilde{p}$ is not given by* (5)*, and $r(S_p\Lambda(0)) > 1$, then*

$$\lim_{n\to\infty} G^n(x,y) = (\hat{x}, 0)$$

whenever $x > 0$.

Proof of Theorem 5.1 The key proposition (which gives us more than we need) is the following.

PROPOSITION 5.2. *Suppose that A is an irreducible nonnegative matrix with column sums c_i such that $c_1 = \min_i c_i < \max_i c_i = c_k$. If $\widetilde{A}$ is a nonnegative matrix obtained from A by changing its first column from*

$$\begin{pmatrix} a_{11} \\ \cdot \\ \vdots \\ a_{k1} \end{pmatrix} \quad \text{to} \quad \begin{pmatrix} a_{11} \\ \cdot \\ \vdots \\ a_{k1} \end{pmatrix} + \gamma \begin{pmatrix} -\sum_{i=2}^{k} a_{i1} \\ a_{21} \\ \cdots \\ a_{k1} \end{pmatrix}$$

for some positive $\gamma > 0$, then $r(A) < r(\tilde{A})$. Alternatively, if $\hat{A}$ is a nonnegative matrix obtained from A by changing its last column from

$$\begin{pmatrix} a_{1k} \\ \cdot \\ \vdots \\ a_{kk} \end{pmatrix} \quad \text{to} \quad \begin{pmatrix} a_{1k} \\ \cdot \\ \vdots \\ a_{kk} \end{pmatrix} - \gamma \begin{pmatrix} -\sum_{i=2}^{k} a_{ik} \\ a_{2k} \\ \cdots \\ a_{kk} \end{pmatrix}$$

for some $\gamma \in (0,1]$, then $r(A) > r(\hat{A})$.

Proof. Note that $c_k > r(A) > c_1$. Let w^T be the left Perron vector for A such that $w_1 = 1$, and let $\tilde{v}$ be the right Perron vector for $\tilde{A}$ normalized so that $w^T\tilde{v} = 1$. Observe that for any γ such that $\tilde{A}$ is nonnegative, $\tilde{A}$ is irreducible and, consequently, v is a positive vector. Set $W = \operatorname{diag}(w_1, \ldots, w_n)$. Then WAW^{-1} has all the column sums equal to $r(A)$. Consider the first column of WAW^{-1}. We see that

$$a_{11} + \sum_{i=2}^{k} w_i a_{i1} = r(A) > c_1 = \sum_{i=1}^{k} a_{i1}.$$

Thus,

$$\sum_{i=2}^{k} w_i a_{i1} > \sum_{i=2}^{k} a_{i1}.$$

It follows that $r(\tilde{A}) = w^T\tilde{A}\tilde{v} = w^TA\tilde{v} + \gamma\tilde{v}_1(-\sum_{i=2}^{k} a_{i1} + \sum_{i=2}^{k} w_i a_{i1}) > w^TA\tilde{v} = r(A)$.

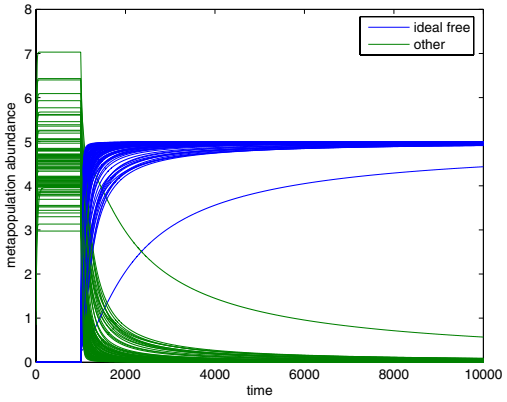


FIG. 3. *One hundred realizations of an ideal-free disperser competing against a random dispersal strategy. In the simulations,* $k = 10$ *and* $\lambda_i(x_i) = \frac{a_i}{1+b_i x_i}$. *For each simulation, the values of* a_i *are randomly selected from the interval* $[1, 2]$, p *is defined by* (5), *where* t *is randomly selected from the interval* $[0, \max(\Lambda^{-1}(I))^{-1} v]$, *and* $\widetilde{p}$ *is randomly selected from* $[0, 1]^{10}$. *To normalize the local population abundances to a value of* 1, *in each simulation* b_i *is set equal to* $\frac{1}{a_i - 1}$.

A similar argument applies to the matrix $\hat{A}$ when $\gamma < 1$, while if $\gamma = 1$, we see that the first column of $\hat{A}$ is $c_k e_k$ and $r(\hat{A}) \geq c_k > r(A)$. □

Now assume $p \in (0, 1]^k$, $r(S_p \Lambda(0)) > 1$, $\hat{x} \gg 0$ is the unique positive fixed point of F, and $v \gg 0$ is a right Perron vector for S. Let $A = S_p \Lambda(\hat{x})$. We begin by showing that $r(S_{\widetilde{p}} \Lambda(\hat{x})) \leq 1$ for all $\widetilde{p} \in [0, 1]^k$ implies that S is stochastic and p is given by (5). First, we show that A must have constant column sums c_i. Suppose to the contrary that there exists $1 \leq j \leq k$ such that $c_j = \max_i c_i > \min_i c_i$. Let $\widetilde{p}$ be any strategy where $\widetilde{p}_i = p_i$ for $i \neq j$ and $\widetilde{p}_j \in (0, p_j)$. Then $S_{\widetilde{p}} \Lambda(\hat{x})$ is given by replacing the jth column of A by a column which is

$$\geq \begin{pmatrix} a_{1j} \\ \cdot \\ \vdots \\ a_{kj} \end{pmatrix} - \gamma \begin{pmatrix} -\sum_{i=2}^{k} a_{ij} \\ a_{2j} \\ \cdots \\ a_{kj} \end{pmatrix},$$

where $\gamma = 1 - \frac{\widetilde{p}_j}{p_j} > 0$. Proposition 5.2 implies that $r(S_{\widetilde{p}} \Lambda(\hat{x})) > r(A) = 1$, contrary to our assumption about p. Therefore A must have constant column sums $c = c_1 = \cdots = c_k$. Second, suppose to the contrary that S is substochastic. Let $\widetilde{p}$ be any strategy where $\widetilde{p}_i \in (0, p_i)$. Since S is substochastic, every column sum $S_{\widetilde{p}} \Lambda(\hat{x})$ is greater than or equal to c and at least one column sum is strictly greater than c. Hence, $r(S_{\widetilde{p}} \Lambda(\hat{x})) > r(A) = 1$, contrary to our assumption about p. Therefore, S is stochastic. Finally, since S is stochastic, it follows that $c = 1$ and $\Lambda(\hat{x}) = I$. Since $\hat{x} \gg 0$, we have $\lambda_i(0) > 1$ and $\hat{x}_i = \lambda_i^{-1}(1)$ for all i. Since $\hat{x}$ is a fixed point, we get that $\hat{x} = (I - \operatorname{diag}(p) + S \operatorname{diag}(p))\hat{x}$. Equivalently, $S \operatorname{diag}(p)\hat{x} = \operatorname{diag}(p)\hat{x}$. Hence, $\operatorname{diag}(p)\hat{x} \gg 0$ is a right Perron vector for S and p is given by (5).

Now suppose that S is stochastic and p is given by (5). Then $\Lambda(\hat{x}) = I$ and $r(S_{\widetilde{p}} \Lambda(\hat{x})) = r(S_{\widetilde{p}}) = 1$ for all $\widetilde{p} \in [0, 1]^k$. □

Conjecture 5.1 suggests that for populations with Beverton–Holt local dynamics, the evolution of conditional dispersers will favor strategies on the ray defined by (5).

Hence, it is natural to ask what happens when two strategies on this ray compete against one another.

PROPOSITION 5.3. *Assume that* $\Lambda(x)$ *and* S *satisfy Assumptions* A1–A4, $\lambda_i(0) > 1$ *for all* i, *and* S *is stochastic. Let* p *and* $\widetilde{p}$ *be strategies given by* (5) *with* $t = d$ *and* $t = \widetilde{d}$, *where* $0 < d < \widetilde{d} \le 1/\max\{\left(\Lambda^{-1}(I)\right)^{-1} v\}$. *Then the set of fixed points of* G *are* $(0,0)$ *and*

$$L = \{(\alpha\hat{x}, (1-\alpha)\hat{x} : \alpha \in [0,1]\},$$

where $\hat{x} = \Lambda^{-1}(I)\mathbf{1}$. *Moreover, if* $\Lambda(x)$ *is continuously differentiable with* $\lambda_i'(x_i) < 0$ *for all* i, *and* $\frac{d}{dx_i} x_i \lambda_i(x_i) > 0$ *for all* i, *then there exists a neighborhood* $U \subset \mathbf{R}_+^k \times \mathbf{R}_+^k$ *of* L *and a homeomorphism* $h : [0,1] \times D \to U$ *with* $D = \{z \in \mathbf{R}^{2k-1} : \|z\| < 1\}$ *such that* $h(\alpha, 0) = \alpha\hat{x} + (1-\alpha)\hat{x}$, $h(0, D) = \{(0, y) \in U\}$, $h(1, D) = \{(x, 0) \in U\}$, *and* $\lim_{n\to\infty} G^n(x,y) = (\alpha\hat{x}, (1-\alpha)\hat{x})$ *for all* $(x,y) \in h(\{\alpha\} \times D)$.

Proof. By the change of variables $x \mapsto \Lambda^{-1}(I)^{-1}\mathrm{diag}\,(v)x$, we can assume without any loss of generality that $p = d\mathbf{1}$ and $\widetilde{p} = \widetilde{d}\mathbf{1}$. Thus, a point $(x,y) > 0$ is a fixed point of G if and only if

$$\begin{aligned}((1-d)I + dS)\Lambda(x+y)x &= x,\\ ((1-\widetilde{d})I + \widetilde{d}S)\Lambda(x+y)y &= y.\end{aligned}$$

Since $r(((1-d)I + dS)\Lambda(x+y))) = r((1-\widetilde{d})I + \widetilde{d}S)\Lambda(x+y)) = 1$ and $d \neq \widetilde{d}$, Theorem 3.1 implies that $\Lambda(x+y) = I$. Therefore, (x,y) needs to satisfy $x + y = \Lambda^{-1}(I)\mathbf{1}$, $Sdx = dx$, and $S\widetilde{d}y = \widetilde{d}y$. Since S is primitive, we get that x must be a scalar multiple of y. Hence, the fixed points of G are given by $(0,0)$ and L.

Now assume that $x \mapsto \Lambda(x)$ is continuously differentiable, $\lambda_i'(x) < 0$ for all i, and $\frac{d}{dx_i} x_i \lambda_i(x_i) > 0$ for all i. We will show that L is a normally hyperbolic attractor in the sense of Hirsch, Pugh, and Shub [7]. Let $(x,y) \in L$. We have

$$DG(x,y) = \begin{pmatrix} S_d(\Lambda'(x+y)\mathrm{diag}\,(x) + \Lambda(x+y)) & S_d\Lambda'(x+y)\mathrm{diag}\,(x) \\ S_d\Lambda'(x+y)\mathrm{diag}\,(y) & S_d(\Lambda'(x+y)\mathrm{diag}\,(y) + \Lambda(x+y)) \end{pmatrix}.$$

Since $0 < \lambda_i'(x_i+y_i)(x_i+y_i) + \lambda_i(x_i+y_i) < \lambda_i'(x_i+y_i)x_i + \lambda_i(x_i+y_i)$ for all i, the diagonal blocks, $S_d(\Lambda'(x+y)\mathrm{diag}\,(x) + \Lambda(x+y))$ and $S_d(\Lambda'(x+y)\mathrm{diag}\,(y) + \Lambda(x+y))$ of $DG(x,y)$, are nonnegative primitive matrices. Since $\lambda_i'(x_i+y_i) < 0$ for all i, the off-diagonal blocks, $S_d\Lambda'(x+y)\mathrm{diag}\,(x)$ and $S_d\Lambda'(x+y)\mathrm{diag}\,(y)$, of $DG(x,y)$ are negative scalar multiples of primitive matrices. Hence, $DG(x,y)$ is a primitive matrix with respect to the competitive ordering on $\mathbf{R}_+^k \times \mathbf{R}_+^k$; i.e., $(\tilde{x}, \tilde{y}) \ge_K (x,y)$ if $\tilde{x} \ge x$ and $\tilde{y} \le y$. Since L is a line of fixed points, $DG(x,y)$ has an eigenvalue of one associated with the eigenvector $(\Lambda^{-1}(I)\mathbf{1}, -\Lambda^{-1}(I)\mathbf{1})$. The Perron–Frobenius theorem implies that all the other eigenvalues of $DG(x,y)$ are strictly less than one in absolute value. Hence, L is a normally hyperbolic one-dimensional attractor. Theorem 4.1 of [7] implies that there is a neighborhood $U \subset \mathbf{R}_+^k \times \mathbf{R}_+^k$ of L and a homeomorphism $h : [0,1] \times D \to U$ with $D = \{z \in \mathbf{R}^{2k-1} : \|z\| < 1\}$ such that $h(\alpha, 0) = \alpha\hat{x} + (1-\alpha)\hat{x}$, $h(0, D) = \{(0, y) \in U\}$, $h(1, D) = \{(x, 0) \in U\}$, and $\lim_{n\to\infty} G^n(x,y) = (\alpha\hat{x}, (1-\alpha)\hat{x})$ for all $(x,y) \in h(\{\alpha\} \times D)$. ∎

Proposition 5.3 implies that once a "resident" population playing a strategy of the form (5) has established itself, a "mutant" strategy of the form (5) can invade only in a weak sense: if the mutants enter at low density, deterministically they will converge to an equilibrium with a low mutant density. After the invasion, one would

expect that demographic or environmental stochasticity would with greater likelihood result in the displacement of the mutants. Hence, once a strategy of the form (5) has established itself, it is likely to resist invasion attempts from other strategies of the form (5). Proposition 5.3 also suggests the following conjecture, which is supported by simulations using the Beverton–Holt growth function.

CONJECTURE 5.2. *Under the conditions of Proposition* 5.3*, for every* $(x, y) > 0$ *there exists* $\alpha \in [0, 1]$ *such that*

$$\lim_{n\to\infty} G^n(x, y) = (\alpha\Lambda^{-1}(I)\mathbf{1}, (1-\alpha)\Lambda^{-1}(I)\mathbf{1}).$$

6. Discussion. For organisms that disperse unconditionally, we proved that a slower dispersing population competitively excludes a faster dispersing population. Similar results have been proven for reaction diffusion equations where the dispersal kernel is self-adjoint [2], observed in a partial analysis of two patch differential equations [8] and illustrated with simulations of two patch difference equations [17]. Our proofs apply to difference equations with an arbitrary number of patches and without any symmetry assumptions about the dispersal matrix S. Since geographical and ecological barriers often create asymmetries in the movement patterns of active dispersers and create asymmetries in abiotic and biotic currents that carry passive dispersers, accounting for these asymmetries is crucial and results in a significantly more difficult mathematical problem than the symmetric case. Theorem 3.1 provides the solution to this problem by proving for any given environmental condition (i.e., the choice of Λ and S), the principal eigenvalue for the growth dispersal matrix is a decreasing function of the dispersal rate. Hence, under all environmental conditions, populations that disperse more slowly spectrally dominate populations that disperse more quickly. Despite this spectral dominance, simulations (e.g., Figure 1) illustrate that for appropriate initial conditions, faster dispersers can be numerically dominant as they initially spread across a landscape. This initial phase of numerical dominance has empirical support in studies of northern range limits of butterflies: dispersal rates increase as species move north to newly formed favorable habitat [6]. Presumably over a long period of time, selection will favor slower dispersal rates commensurate with their ancestral rates of movement (R. Holt, *personal communication*). However, since all initial conditions do not lead to an initial phase of numerical dominance for the faster dispersers (e.g., if the initial condition is a Perron vector for the slower disperser), we still require a detailed understanding of how the local intrinsic rates of growth, the dispersal matrix, and initial conditions determine whether the faster or slower disperser is numerically dominant in the initial phase of establishment.

For conditional dispersers experiencing no dispersal costs (i.e., S is column stochastic and $\lambda_i(0) > 1$ for all i), we provide proofs that generalize previous findings in two patch models [9, 17]. We prove that all dispersal strategies outside of a one-parameter family are not evolutionarily stable: when a population adopts one of these strategies, there are nearby strategies that can invade. For populations playing strategies in this exceptional one-parameter family, the populations exhibit an ideal-free distribution at equilibrium: the per-capita growth rate is constant across the landscape [3]. Contrary to prior expectations [17], we show that are growth functions for which these ideal-free strategies cannot displace all other strategies. However, numerical simulations with the biologically plausible Beverton–Holt growth functions suggest that populations playing these ideal-free strategies can displace populations playing any other strategy. Moreover, when a population at equilibrium plays an ideal-free strategy, we prove that a population playing another ideal-free strategy cannot increase from being rare

and, consequently, is likely to be driven to extinction by stochastic forces. For populations playing these ideal-free strategies, the dispersal likelihood in a patch is inversely proportional to the equilibrium abundance in a patch. Hence, enriching one patch may result in the evolution of lower dispersal rates in that patch. Conversely, habitat degradation of a patch may result in the evolution of higher dispersal rates in that patch. These predictions about ideal-free strategies, however, have to be viewed with caution, as they are sensitive to the assumption of no dispersal costs. The inclusion of the slightest dispersal costs destroys this one-parameter family of evolutionary stable strategies and leaves only the nondispersal strategy as a candidate for an evolutionary stable strategy.

Our models make several simplifying assumptions, and relaxing these assumptions provides several mathematical problems of biological interest. Most importantly, our models do not include temporal heterogeneity, which is an important ingredient in the evolution of dispersal [17]. Temporal heterogeneity can be generated exogenously or endogenously and when combined with spatial heterogeneity can promote the evolution of faster dispersers [10, 13, 17]. For instance, Hutson, Mischaikow, and Poláčik [13] proved that a faster disperser can displace or coexist with a slower disperser for periodically forced reaction diffusion equations. Whether similar results can be proven for periodic or, more generally, random difference equations requires answering mathematically challenging questions about spectral properties of periodic and random products of nonnegative matrices. Similar challenges arise when replacing increasing growth functions with unimodal growth functions [4, 10, 19] that can generate temporal heterogeneity via periodic and chaotic population dynamics.

Acknowledgments. The authors thank participants of the William and Mary "matrices in biology sessions" for their valuable feedback on this evolving work. In particular, we thank Greg Smith and Marco Huertas for finding counterexamples to an earlier conjectured form of Theorem 3.1. We also thank Bob Holt, Vivian Hutson, Konstantin Mischaikow, and an anonymous referee for their encouraging comments and helpful suggestions.

REFERENCES

[1] R. J. H. Beverton and S. J. Holt, *On the Dynamics of Exploited Fish Populations*, Fish. Invest. Ser. II 19, Ministry of Agriculture, Fisheries and Food, London, UK, 1957.

[2] J. Dockery, V. Hutson, K. Mischaikow, and M. Pernarowski, *The evolution of slow dispersal rates: a reaction diffusion model*, J. Math. Biol., 37 (1998), pp. 61–83.

[3] S. D. Fretwell and H. L. Lucas, *On territorial behavior and other factors influencing patch distribution in birds*, Acta Biotheoretica, 19 (1970), pp. 16–36.

[4] W. M. Getz, *A hypothesis regarding the abruptness of density dependence and the growth rate of populations*, Ecology, 77 (1996), pp. 2014–2026.

[5] A. Hastings, *Can spatial variation alone lead to selection for dispersal?*, Theor. Pop. Biol., 24 (1983), pp. 244–251.

[6] J. K. Hill, C. D. Thomas, R. Fox, M. G. Telfer, S. G. Willis, J. Asher, and B. Huntley, *Responses of butterflies to twentieth century climate warming: Implications for future ranges*, Roy. Soc. Lond. Proc. Ser. Biol. Sci., 269 (2002), pp. 2163–2171.

[7] M. W. Hirsch, C. C. Pugh, and M. Shub, *Invariant Manifolds*, Springer-Verlag, Berlin, 1977.

[8] R. D. Holt, *Population dynamics in two-patch environments: Some anomalous consequences of an optimal habitat distribution*, Theor. Pop. Biol., 28 (1985), pp. 181–208.

[9] R. D. Holt and M. Barfield, *On the relationship between the ideal-free distribution and the evolution of dispersal*, in Dispersal, J. Clobert, E. Danchin, A. Dhondt, and J. Nichols, eds., Oxford University Press, Oxford, UK, 2001, pp. 83–95.

[10] R. D. Holt and M. A. McPeek, *Chaotic population dynamics favors the evolution of dispersal*, Amer. Nat., 148 (1996), pp. 709–718.

[11] S. B. Hsu, H. L. Smith, and P. Waltman, *Competitive exclusion and coexistence for competitive systems on ordered Banach spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 4083–4094.
[12] V. Hutson, S. Martinez, K. Mischaikow, and G. T. Vickers, *The evolution of dispersal*, J. Math. Biol., 47 (2003), pp. 483–517.
[13] V. Hutson, K. Mischaikow, and P. Poláčik, *The evolution of dispersal rates in a heterogeneous time-periodic environment*, J. Math. Biol., 43 (2001), pp. 501–533.
[14] V. S. Ivlev, *Experimental Ecology of the Feeding of Fishes*, Yale University Press, New Haven, CT, 1955.
[15] M. L. Johnson and M. S. Gaines, *Evolution of dispersal: Theoretical models and empirical tests using birds and mammals*, Annu. Rev. Ecol. Syst., 21 (1990), pp. 449–480.
[16] S. A. Levin, D. Cohen, and A. Hastings, *Dispersal strategies in patchy environments*, Theoret. Pop. Biol., 26 (1984), pp. 165–191.
[17] M. A. McPeek and R. D. Holt, *The evolution of dispersal in spatially and temporally varying environments*, Amer. Nat., 6 (1992), pp. 1010–1027.
[18] K. Mischaikow, H. Smith, and H. R. Thieme, *Asymptotically autonomous semiflows: Chain recurrence and Lyapunov functions*, Trans. Amer. Math. Soc., 347 (1995), pp. 1669–1685.
[19] W. E. Ricker, *Stock and recruitment*, J. Fish. Res. Board. Can., 11 (1954), pp. 559–623.
[20] A. L. Shanks, *Mechanisms of cross-shelf dispersal of larval invertebrates and fish*, in Ecology of Marine Invertebrate Larvae, CRC, Boca Raton, FL, 1995, pp. 324–367.

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1383–1406

VARIABLE EXPONENT, LINEAR GROWTH FUNCTIONALS IN IMAGE RESTORATION*

YUNMEI CHEN†, STACEY LEVINE‡, AND MURALI RAO†

Abstract. We study a functional with variable exponent, $1 \leq p(x) \leq 2$, which provides a model for image denoising, enhancement, and restoration. The diffusion resulting from the proposed model is a combination of total variation (TV)-based regularization and Gaussian smoothing. The existence, uniqueness, and long-time behavior of the proposed model are established. Experimental results illustrate the effectiveness of the model in image restoration.

Key words. image restoration, linear growth functionals, variable exponent, BV-space, Dirichlet boundary condition

AMS subject classifications. 49J40, 35K65

DOI. 10.1137/050624522

1. Introduction.

1.1. Background. In this paper we propose a new model for image restoration. In the version we address, an image, u, is recovered from an observed, noisy image, I, where the two are related by $I = u + noise$. The proposed model incorporates the strengths of the various types of diffusion arising from the minimization problem

$$\min \int_\Omega |Du|^p + \frac{\lambda}{2}(u - I)^2 \tag{1.1}$$

for $1 \leq p \leq 2$ ($\lambda \geq 0$, and Ω is an open, bounded subset of $\mathbb{R}^n$ with Lipschitz boundary). Specifically, we exploit the benefits of isotropic diffusion ($p = 2$), total variation (TV)-based diffusion ($p = 1$), and more general anisotropic diffusion ($1 < p < 2$).

TV minimization, $p = 1$. TV-based regularization, $p = 1$, as first proposed by Rudin, Osher, and Fatemi [31], does an excellent job of preserving edges while reconstructing images. Mathematically this is reasonable, since it is natural to study solutions of this problem in the space of functions of bounded variation, $BV(\Omega)$, allowing for discontinuities which are necessary for edge reconstruction. This phenomenon can also be explained physically, since the resulting diffusion is strictly orthogonal to the gradient of the image. The TV model has been studied extensively (see, e.g., [1, 15]) and has proved to be an invaluable tool for preserving sharp edges.

*Received by the editors February 15, 2005; accepted for publication (in revised form) December 12, 2005; published electronically May 3, 2006. The first and third authors were supported in part by NIH R01 NS42075. The second author was supported in part by NSF DMS 0505729.
http://www.siam.org/journals/siap/66-4/62452.html
†Department of Mathematics, P.O. Box 118505, University of Florida, Gainesville, FL 32611 (yun@math.ufl.edu, rao@math.ufl.edu).
‡Department of Mathematics and Computer Science, Duquesne University, Pittsburgh, PA 15282 (sel@mathcs.duq.edu).

Given the success of TV-based diffusion, various modifications have been introduced. For instance, Strong and Chan [32] proposed the adaptive total variation model

$$\min \int_\Omega \alpha(x)|\nabla u|$$

in which they introduce a control factor, $\alpha(x)$, which slows the diffusion at likely edges. This controls the *speed* of the diffusion and has demonstrated good results, as it aids in noise reduction. It is also good at reconstructing edges, since the *type* of diffusion (strictly orthogonal to the image gradient) is the same as that of the original TV model.

TV-based denoising favors solutions that are piecewise constant. This sometimes causes a *staircasing* effect in which noisy smooth regions are processed into piecewise constant regions (see Figure 5.1), a phenomenon long observed in the literature; see, e.g., [9, 15, 20, 28, 30, 34, 36]. Not only do "blocky" solutions fail to satisfy the ubiquitous "eyeball norm," but they can also develop "false edges," which can mislead a human or computer into identifying erroneous features not present in the true image.

Minimization problem (1.1) with $1 < p \leq 2$. On the other hand, one can explore different *types* of diffusion arising from (1.1). Choosing $p = 2$ results in isotropic diffusion, which solves the staircasing problem but alone is not good for image reconstruction since it has no mechanism for preserving edges. Different values of $1 < p < 2$ result in anisotropic diffusion, which is somewhere between TV-based and isotropic smoothing. This type of diffusion can be effective in reconstructing piecewise smooth regions. However, a fixed value of $1 < p < 2$ may not allow for discontinuities, thus obliterating edges. This was shown to be true in the discrete case [28].

Combination of TV-based and isotropic diffusion. Given the strengths of (1.1) for different values of p, it seems worthwhile to investigate a model which could self-adjust in order to reap the benefits of each type of diffusion. To this end, Chambolle and Lions [15] proposed minimizing the following energy functional, which combines isotropic and TV-based diffusion:

$$\min_{u \in BV(\Omega)} \frac{1}{2\beta} \int_{|\nabla u| \leq \beta} |\nabla u|^2 + \int_{|\nabla u| > \beta} |\nabla u| - \frac{\beta}{2}. \tag{1.2}$$

In this model, the diffusion is strictly perpendicular to the gradient, where $|\nabla u| > \beta$, that is, where edges are most likely present, and isotropic where $|\nabla u| \leq \beta$. This model is successful in restoring images in which homogeneous regions are separated by distinct edges; however, if the image intensities representing objects are nonuniform, or if an image is highly degraded, this model may become sensitive to the threshold, β (see Figures 5.2, 5.3, and 5.4). In this case, one might want more flexibility when choosing both the direction and speed of diffusion.

Blomgren et al. [9] proposed the minimization problem

$$\min \int_\Omega |\nabla u|^{p(|\nabla u|)} dx,$$

where $\lim_{s \to 0} p(s) = 2$, $\lim_{s \to \infty} p(s) = 1$, and p is monotonically decreasing. This model should reap the benefits of both isotropic and TV-based diffusion, as well as a combination of the two. However, it is difficult to study mathematically since the lower semicontinuity of the functional is not readily evident.

1.2. Functionals with variable exponent $1 \leq p(x) \leq 2$. The model proposed in this paper capitalizes on the strengths of (1.1) for the different values of $1 \leq p \leq 2$. It ensures TV-based diffusion ($p \equiv 1$) along edges and Gaussian smoothing ($p \equiv 2$) in homogeneous regions. Furthermore, it employs anisotropic diffusion ($1 < p < 2$) in regions which may be piecewise smooth or in which the difference between noise and edges is difficult to distinguish. We let $p = p(x)$ depend on the location, x, in the image. This way, the direction and speed of diffusion at each location depend on the local behavior. Moreover, our choice of exponent yields a model which we can show is theoretically sound.

To this end, the proposed model is as follows:

$$\min_{u \in BV \cap L^2(\Omega)} \int_\Omega \phi(x, Du) + \frac{\lambda}{2}(u - I)^2, \tag{1.3}$$

where

$$\phi(x, r) := \begin{cases} \frac{1}{q(x)}|r|^{q(x)}, & |r| \leq \beta, \\ |r| - \frac{\beta q(x) - \beta^{q(x)}}{q(x)}, & |r| > \beta, \end{cases} \tag{1.4}$$

where $\beta > 0$ is fixed and $1 < \alpha \leq q(x) \leq 2$. For instance, one can choose

$$q(x) = 1 + \frac{1}{1 + k|\nabla G_\sigma * I(x)|^2}, \tag{1.5}$$

where $G_\sigma(x) = \frac{1}{\sigma}\exp(-|x|^2/4\sigma^2)$ is the Gaussian filter and $k > 0$ and $\sigma > 0$ are fixed parameters.

The main benefit of (1.3)–(1.4) is the manner in which it accommodates the local image information. Where the gradient is sufficiently large (i.e., at likely edges), only TV-based diffusion will be used. Where the gradient is close to zero (i.e., in homogeneous regions), the model is isotropic. At all other locations, the filtering is somewhere between Gaussian and TV-based diffusion. Specifically, the type of anisotropy at these ambiguous regions varies according to the strength of the gradient. This enables the model to have a much lower dependence on the threshold (see Figures 5.2, 5.3, 5.4, and 5.5).

For several reasons, we've chosen here to prove the well-posedness of the Dirichlet boundary value problem

$$\min_{u \in BV_g \cap L^2(\Omega)} \int_\Omega \phi(x, Du) + \frac{\lambda}{2}(u - I)^2, \tag{1.6}$$

where

$$BV_g(\Omega) := \{u \in BV(\Omega) | u = g \text{ on } \partial\Omega\}, \tag{1.7}$$

and its associated flow

$$\dot{u} - \operatorname{div}(\phi_r(x, Du)) + \lambda(u - I) = 0 \quad \text{in } \Omega^T, \tag{1.8}$$

$$u(x, t) = g(x) \quad \text{on } \partial\Omega^T, \tag{1.9}$$

$$u(0) = I \quad \text{in } \Omega, \tag{1.10}$$

where

$$\Omega^T := \Omega \times [0, T] \quad \text{and} \quad \partial\Omega^T := \partial\Omega \times [0, T]. \tag{1.11}$$

First, the theory is more interesting and challenging mathematically than (1.3). Second, all of the techniques used to study (1.6) also directly solve (1.3). Finally, the Dirichlet problem (1.6) also has direct application in image processing, as it can be used for image interpolation [13, 27], also referred to as noise-free image inpainting [8, 17, 18].

For a special case of (4), where $q(x) \equiv 2$, i.e., $\phi(r) := |r| - \frac{1}{2}$ when $|r| > 1$ and $\phi(r) := \frac{1}{2}|r|^2$ when $|r| \leq 1$, the existence, uniqueness, and long-time behavior of solutions of (1.6) and its related flow were studied by Zhou [37]. Later, these results were extended to general convex linear-growth functionals, $\phi = \phi(Du)$, by Hardt and Zhou [23]. In this paper we study the more general case, where the functional has a variable exponent and where $\phi = \phi(x, Du)$. We use a different approximate functional than [37], so different estimates are required. Our analysis is also based on techniques introduced in [19]; however, our energy functional requires alternate techniques to establish lower semicontinuity and to pass to the limit from the approximate solution. More related work on linear-growth functionals and their flows can be found in [2, 3, 4, 5, 6, 7, 12, 14, 33]. We also refer the reader to the work in [15, 16, 26, 35] for an alternate variational approach for reducing staircasing by minimizing second order functionals.

The paper is organized as follows. In section 2 we establish some important properties of $\phi(x, Du)$. In section 3 we prove the existence and uniqueness of the solution of minimization problem (1.6). In section 4 we study the associated evolution problem (1.8)–(1.10). Specifically, we define the notion of a weak solution of (1.8)–(1.10), derive estimates for the solution of an approximating problem, prove existence and uniqueness of the solution of (1.8)–(1.10), and discuss the behavior of the solution as $t \to \infty$. In section 5 we provide our numerical algorithm and experimental results to illustrate the effectiveness of our model in image restoration.

2. Properties of ϕ. Recall that for $u \in BV(\Omega)$,

$$Du = \nabla u \cdot \mathcal{L}^n + D^s u$$

is a Radon measure, where ∇u is the density of the absolutely continuous part of Du with respect to the n-dimensional Lebesgue measure, $\mathcal{L}^n$, and $D^s u$ is the singular part (see [21]).

DEFINITION 2.1. *For $v \in BV(\Omega)$, define*

$$\int_\Omega \phi(x, Dv) := \int_\Omega \phi(x, \nabla v)dx + \int_\Omega |D^s v|,$$

where ϕ is defined as in (1.4). *Furthermore, denote*

$$\Phi_\lambda(v) := \int_\Omega \phi(x, Dv) + \frac{\lambda}{2}\int_\Omega |v - I|^2 dx, \tag{2.1}$$

$$\Phi_g(v) := \int_\Omega \phi(x, Dv) + \int_{\partial\Omega} |v - g| d\mathcal{H}^{n-1}, \tag{2.2}$$

and

$$\Phi_{\lambda,g}(v) := \int_\Omega \phi(x, Dv) + \frac{\lambda}{2}\int_\Omega |v - I|^2 dx + \int_{\partial\Omega} |v - g| d\mathcal{H}^{n-1}. \tag{2.3}$$

Remark 2.2. For simplicity, we assume that the threshold $\beta = 1$ in (1.4) for all of our theoretical results.

Similarly to the idea of [10, 11], we can establish lower semicontinuity of the functional Φ_g. For the convenience of the reader, we include the proof below.

LEMMA 2.3. *Using the notation in Definition* 2.1,

$$\Phi_g(u) = \widetilde{\Phi}_g(u) \tag{2.4}$$

for all $u \in BV(\Omega)$, *where*

$$\widetilde{\Phi}_g(u) := \sup_{\substack{\psi \in C^1(\overline{\Omega}, \mathbb{R}^n) \\ |\psi| \leq 1}} \int_\Omega -u \mathrm{div}\psi - \frac{q(x)-1}{q(x)} |\psi|^{\frac{q(x)}{q(x)-1}} dx + \int_{\partial\Omega} \psi \cdot n g d\mathcal{H}^{n-1}.$$

Furthermore, $\Phi_g(u)$ *is lower semicontinuous on* $L^1(\Omega)$*; that is, if* $u_j, u \in BV(\Omega)$ *satisfy* $u_j \to u$ in $L^1(\Omega)$ *as* $j \to \infty$, *then* $\Phi_g(u) \leq \liminf_{j\to\infty} \Phi_g(u_j)$.

Proof. For each $\psi \in C^1(\overline{\Omega}, \mathbb{R}^n)$, the map $u \to \int_\Omega -u \mathrm{div}\psi - \frac{q(x)-1}{q(x)} |\psi|^{\frac{q(x)}{q(x)-1}} dx + \int_{\partial\Omega} \psi \cdot n g d\mathcal{H}^{n-1}$ is continuous and affine on $L^1(\Omega)$. Therefore, $\widetilde{\Phi}_g(u)$ is convex and lower semicontinuous on $L^1(\Omega)$ and the domain of $\widetilde{\Phi}_g(u)$, $\{u \,|\, \widetilde{\Phi}_g(u) < \infty\}$, is precisely $BV(\Omega)$.

We now show that $\Phi_g(u) = \widetilde{\Phi}_g(u)$. For $u \in BV(\Omega)$, we have that for each $\psi \in C^1(\overline{\Omega}, \mathbb{R}^n)$,

$$-\int_\Omega u \mathrm{div}\psi dx = \int_\Omega \nabla u \cdot \psi dx + \int_\Omega D^s u \cdot \psi - \int_{\partial\Omega} u\psi \cdot n d\mathcal{H}^{n-1}.$$

Therefore, since the measures dx, $D^s u$, and $d\mathcal{H}^{n-1}$ are mutually singular, standard arguments show that

$$\widetilde{\Phi}_g(u) = \sup_{\substack{\psi \in C^1(\overline{\Omega}, \mathbb{R}^n) \\ |\psi| \leq 1}} \int_\Omega \nabla u \cdot \psi - \frac{q(x)-1}{q(x)} |\psi|^{\frac{q(x)}{q(x)-1}} dx + \int_\Omega |D^s u| + \int_{\partial\Omega} |u - g| d\mathcal{H}^{n-1}.$$

The proof is then complete once we establish that

$$\int \phi(x, \nabla u) dx = \sup_{\substack{\psi \in C^1(\overline{\Omega}, \mathbb{R}^n) \\ |\psi| \leq 1}} \int \nabla u \cdot \psi - \frac{q(x)-1}{q(x)} |\psi|^{\frac{q(x)}{q(x)-1}} dx. \tag{2.5}$$

Since any $\rho \in L^\infty(\Omega, \mathbb{R}^n)$ can be approximated in measure by $\psi \in C^1(\overline{\Omega}, \mathbb{R}^n)$, we have that

(2.6)

$$\sup_{\substack{\psi \in C^1(\overline{\Omega}, \mathbb{R}^n) \\ |\psi| \leq 1}} \int_\Omega \nabla u \cdot \psi - \frac{q(x)-1}{q(x)} |\psi|^{\frac{q(x)}{q(x)-1}} dx = \sup_{\substack{\rho \in L^\infty(\Omega, \mathbb{R}^n) \\ |\rho| \leq 1}} \int_\Omega \nabla u \cdot \rho - \frac{q(x)-1}{q(x)} |\rho|^{\frac{q(x)}{q(x)-1}} dx.$$

Choosing $\rho(x) = 1_{\{|\nabla u| \leq 1\}} |\nabla u|^{q(x)-1} \frac{\nabla u}{|\nabla u|} + 1_{\{|\nabla u| > 1\}} \frac{\nabla u}{|\nabla u|}$, where 1_E is the indicator function on E, we see that the right-hand side of (2.6) is

(2.7)

$$\geq \int_\Omega \frac{1}{q(x)} |\nabla u|^{q(x)} 1_{\{|\nabla u| \leq 1\}} + \left[|\nabla u| - \frac{q(x)-1}{q(x)} \right] 1_{\{|\nabla u| > 1\}} dx = \int_\Omega \phi(x, \nabla u) dx.$$

To show equality in (2.5), we proceed as follows. For any $\rho \in L^\infty(\Omega, \mathbb{R}^n)$, since $q(x) > 1$ we have that for almost all x, $\nabla u(x) \cdot \rho(x) \leq \frac{1}{q(x)}|\nabla u|^{q(x)} + \frac{q(x)-1}{q(x)}|\rho(x)|^{\frac{q(x)}{q(x)-1}}$. In particular, if $|\nabla u| \leq 1$,

$$\nabla u(x) \cdot \rho(x) - \frac{q(x)-1}{q(x)}|\rho(x)|^{\frac{q(x)}{q(x)-1}} \leq \frac{1}{q(x)}|\nabla u|^{q(x)}. \tag{2.8}$$

On the other hand, if $|\nabla u| > 1$ and $|\rho| \leq 1$, then since $q(x) > 1$ for almost all x, we have that $\nabla u \cdot \rho = |\nabla u|\frac{\nabla u}{|\nabla u|} \cdot \rho \leq |\nabla u|[\frac{1}{q(x)} + \frac{q(x)-1}{q(x)}|\rho|^{\frac{q(x)}{q(x)-1}}]$, and thus

$$\begin{aligned} \nabla u \cdot \rho - \frac{q(x)-1}{q(x)}|\rho|^{\frac{q(x)}{q(x)-1}} &\leq \frac{1}{q(x)}|\nabla u| + (|\nabla u| - 1)\frac{q(x)-1}{q(x)}|\rho|^{\frac{q(x)}{q(x)-1}} \\ &\leq |\nabla u| - \frac{q(x)-1}{q(x)}. \end{aligned} \tag{2.9}$$

Combining (2.6), (2.7), (2.8), and (2.9), we have that

$$\sup_{\substack{\psi \in C^1(\overline{\Omega}, \mathbb{R}^n) \\ |\psi| \leq 1}} \int_\Omega \nabla u \cdot \psi - \frac{q(x)-1}{q(x)}|\psi|^{\frac{q(x)}{q(x)-1}} dx = \int_\Omega \phi(x, \nabla u) dx,$$

and thus for all $u \in BV(\Omega)$, $\widetilde{\Phi}_g(u) = \Phi_g(u)$, where Φ_g is defined as in (2.2). ∎

LEMMA 2.4. *Suppose $\Omega \subset \mathbb{R}^n$ is open, bounded, and has Lipschitz boundary, and let $w \in BV \cap L^2(\Omega)$. Then for each $\delta > 0$, there exists $\tilde{w}_\delta \in C^\infty \cap H^1(\Omega)$ such that*

$$\|\tilde{w}_\delta - w\|_{L^2(\Omega)} \leq \delta \tag{2.10}$$

and

$$\Phi_{\lambda,g}(\tilde{w}_\delta) \leq \Phi_{\lambda,g}(w) + \delta. \tag{2.11}$$

Furthermore, if we also assume that $Trw = TrG$ on $\partial\Omega$ for some $G \in H^1(\Omega)$, then for each $\delta > 0$, there exists $w_\delta \in H^1(\Omega)$ satisfying

$$Trw_\delta = Trw \quad \text{on } \partial\Omega, \tag{2.12}$$

$$w_\delta \to w \quad \text{strongly in } L^2(\Omega) \quad \text{as } \delta \to 0, \tag{2.13}$$

and

$$\Phi_{\lambda,g}(w_\delta) \leq \Phi_{\lambda,g}(w) + \delta, \tag{2.14}$$

where TrG is the trace of G on $\partial\Omega$.

Proof. Fix $w \in BV \cap L^2(\Omega)$. Using Lemma 2.3 and a slight modification of the proof of Theorem 1.17 and Remark 1.18 in [22], there exists a sequence $\{w_j\}$ in $C^\infty \cap H^1(\Omega)$ such that

$$Trw_j = Trw \quad \text{on } \partial\Omega, \tag{2.15}$$

$$w_j \to w \quad \text{in } L^2(\Omega), \tag{2.16}$$

and

$$\lim_{j\to\infty} \Phi_g(w_j) = \Phi_g(w). \tag{2.17}$$

Therefore, for each $\delta > 0$ there exists a function $\tilde{w}_\delta \in C^\infty \cap H^1(\Omega)$ such that (2.10) and (2.11) hold.

Now suppose also that $Trw = TrG$ on $\partial\Omega$. Then by (2.15), $\tilde{w}_\delta - G \in W_0^{1,1}(\Omega)$, and thus there exists a function $h_\delta \in C_c^\infty(\Omega)$ such that

$$\|\tilde{w}_\delta - G - h_\delta\|_{W^{1,1}(\Omega)} + \|\tilde{w}_\delta - G - h_\delta\|_{L^2(\Omega)} \leq \delta. \tag{2.18}$$

Let $w_\delta = G + h_\delta \in H^1(\Omega)$. Then we have that

$$Trw_\delta = Trw \quad \text{on } \partial\Omega,$$

$$w_\delta \to w \quad \text{in } L^2(\Omega) \quad \text{as } \delta \to 0,$$

and

$$\Phi_{\lambda,g}(w_\delta) \leq \Phi_{\lambda,g}(w) + \delta.$$

The proof is complete. ∎

Remark 2.5. If $w \in BV \cap L^\infty(\Omega)$, then there exist $w_\delta \in H^1 \cap L^\infty(\Omega)$ such that, in addition to (2.12)–(2.14), we also have that

$$\|w_\delta\|_{L^\infty(\Omega)} \leq C(\Omega)\|w\|_{L^\infty(\Omega)}. \tag{2.19}$$

3. The minimization problem. In this section we study the existence and uniqueness of the solution to minimization problem (1.6). In general, as discussed in [23, 25, 37], there may not exist a minimizer of $\Phi_\lambda(v)$ (see (2.1)), since the limit of the minimizing sequence may not take the boundary value g. However, we can prove the existence of a unique minimizer for a weaker form of (1.6), where we consider the minimization problem using the *relaxed energy*, $\Phi_{\lambda,g}(v)$, defined in (2.3). To this end, we define a pseudosolution of (1.6) as follows.

DEFINITION 3.1. *A function $u \in BV \cap L^2(\Omega)$ is a pseudosolution of* (1.6) *if it is a solution of*

$$\min_{v \in BV \cap L^2(\Omega)} \Phi_{\lambda,g}(v), \tag{3.1}$$

where $\Phi_{\lambda,g}(v)$ is defined in (2.3).

First, we show that there exists a pseudosolution of (1.6), that is, a solution of (3.1), in Theorem 3.2. In Theorem 3.5 we provide the motivation for using the notion of pseudosolution in Definition 3.1.

THEOREM 3.2. *Suppose $I \in BV \cap L^2(\Omega)$, $g = TrG$ for some function $G \in BV(\Omega)$ with $I = g$ on $\partial\Omega$, and Ω is an open bounded subset of $\mathbb{R}^n$ with Lipschitz boundary. Then there exists a unique pseudosolution of* (1.6) *as given in Definition* 3.1.

Proof. Let $\{u_n\}$ be a minimizing sequence of (3.1) in $BV \cap L^2(\Omega)$. Since $\{u_n\}$ is bounded in $BV(\Omega)$ and $L^2(\Omega)$, using the compactness of $BV(\Omega)$ and the weak compactness of $L^2(\Omega)$, we see that there exists a subsequence $\{u_{n_k}\}$ of $\{u_n\}$ and a function $u \in BV \cap L^2(\Omega)$ satisfying

$$u_{n_k} \to u \quad \text{strongly in } L^1(\Omega), \tag{3.2}$$

$$u_{n_k} \rightharpoonup u \quad \text{weakly in } L^2(\Omega). \tag{3.3}$$

By (3.2), (3.3), Lemma 2.3, and the weak lower semicontinuity of the L^2-norm, we have that

$$\Phi_{\lambda,g}(u) \le \liminf_{k\to\infty} \Phi_{\lambda,g}(u_{n_k}) = \inf_{BV\cap L^2(\Omega)} \Phi_{\lambda,g}(v).$$

Hence, u is a solution of the minimization problem. Uniqueness follows from the strict convexity of $\Phi_{\lambda,g}(v)$ in v. □

To better understand the relationship between (1.6) and (3.1), we need the following two lemmas. For $\beta > 0$, let

$$d_\beta(x) = \min\left(\frac{d(x)}{\beta}, 1\right), \tag{3.4}$$

where $d(x)$ is the distance of the point x to the boundary of Ω.

LEMMA 3.3 (*Theorem A.1 of* [19]). *For each $v \in BV \cap L^\infty(\Omega)$, the vector measures $v\nabla d_\beta$ converge weakly to $-v\gamma d\mathcal{H}^{n-1}$ as $\beta \to 0$, where γ is the unit outward normal to $\partial\Omega$, and*

$$\lim_{\beta\to 0} \int_\Omega |v||\nabla d_\beta| = \int_{\partial\Omega} |v| d\mathcal{H}^{n-1}.$$

LEMMA 3.4. *Let $G \in W^{1,1} \cap L^\infty(\Omega)$. Then for any $v \in BV \cap L^\infty(\Omega)$, there exists a sequence $\{v_\beta\}$ in $BV \cap L^\infty(\Omega)$ such that*

$$\begin{aligned} Trv_\beta &= TrG \quad \text{in } L^1(\partial\Omega), \\ v_\beta &\to v \quad \text{in } L^2(\Omega) \quad \text{as } \beta \to 0, \end{aligned}$$

and

$$\lim_{\beta\to 0} \Phi_g(v_\beta) = \Phi_g(v).$$

Proof. For $\beta > 0$, define $d_\beta(x)$ as in (3.4). Since $d(x) \in W^{1,\infty}(\Omega)$ and $|\nabla d| = 1$, we have that

$$|\nabla d_\beta| = \frac{1}{\beta} \text{ if } d(x) < \beta \quad \text{and} \quad |\nabla d_\beta| = 0 \text{ if } d(x) \ge \beta. \tag{3.5}$$

Fix $v \in BV \cap L^\infty(\Omega)$ and let $v_\beta = d_\beta v + (1 - d_\beta)G$ for $(x,t) \in \Omega$. Then

$$\begin{gathered} v_\beta \in BV \cap L^\infty(\Omega) \text{ with } Trv_\beta = TrG \text{ in } L^1(\partial\Omega), \\ v_\beta \to v \text{ in } L^2(\Omega). \end{gathered}$$

By Lemma 2.3, $\liminf_{\beta\to 0} \Phi_g(v_\beta) \ge \Phi_g(v)$, and thus it remains only to show that

$$\lim_{\beta\to 0} \Phi_g(v_\beta) \le \Phi_g(v). \tag{3.6}$$

Writing $\Phi_g = \widetilde{\Phi}_g$ (Lemma 2.3), since $D^s v_\beta = d_\beta D^s v$ and $Trv_\beta = TrG$ in $L^1(\partial\Omega)$ we have that

$$\begin{aligned} \widetilde{\Phi}_g(v_\beta) = \sup_{\substack{\psi\in C^1(\overline{\Omega},\mathbb{R}^n) \\ |\psi|\le 1}} \Bigg[&\int_\Omega \{(v-G)\nabla d_\beta + d_\beta \nabla v + (1-d_\beta)\nabla G\}\cdot\psi \\ &- \frac{q(x)-1}{q(x)} |\psi|^{\frac{q(x)}{q(x)-1}} dx + \int_\Omega d_\beta D^s v \cdot \psi \Bigg]. \end{aligned}$$

Therefore, by (2.5)

$$\begin{aligned}\widetilde{\Phi}(v_\beta) &\le \sup_{\substack{\psi\in C^1(\overline{\Omega},\mathbb{R}^n)\\ |\psi|\le 1}} \left[\int_\Omega \nabla v\cdot\psi - \frac{q(x)-1}{q(x)}|\psi|^{\frac{q(x)}{q(x)-1}}dx\right] \\ &\quad + \int_\Omega |v-G||\nabla d_\beta|dx + \int_\Omega (1-d_\beta)(|\nabla v|+|\nabla G|)dx + \int_\Omega d_\beta|D^s v| \\ &= \int_\Omega \phi(x,\nabla v) + \int_\Omega |v-G||\nabla d_\beta|dx + \int_\Omega (1-d_\beta)(|\nabla v|+|\nabla G|)dx + \int_\Omega d_\beta|D^s v|.\end{aligned}$$

By Lemma 3.3, as $\beta\to 0$,

$$\int_\Omega |v-G||\nabla d_\beta|dx \to \int_{\partial\Omega} |v-G|d\mathcal{H}^{n-1}.$$

Furthermore, the Lebesgue dominated convergence theorem gives us that

$$\int_\Omega (1-d_\beta)(|\nabla v|+|\nabla G|)dx \to 0 \quad \text{and} \quad \int_\Omega d_\beta|D^s v| \to \int_\Omega |D^s v|.$$

Therefore, (3.6) holds and the lemma is proved. ∎

THEOREM 3.5. *We have that*

$$\inf_{v\in BV_g\cap L^2(\Omega)} \Phi_\lambda(v) = \min_{v\in BV\cap L^2(\Omega)} \Phi_{\lambda,g}(v), \tag{3.7}$$

where Φ_λ *and* $\Phi_{\lambda,g}$ *are as defined in* (2.1) *and* (2.3)*, respectively, and* BV_g *is as defined in* (1.7).

Proof. Since $BV_g(\Omega)\subset BV(\Omega)$,

$$\inf_{v\in BV\cap L^2(\Omega)} \Phi_{\lambda,g}(v) \le \inf_{v\in BV_g\cap L^2(\Omega)} \Phi_{\lambda,g}(v) = \inf_{v\in BV_g\cap L^2(\Omega)} \Phi_\lambda(v).$$

To see the reverse, let $u\in BV\cap L^2(\Omega)$ be the solution of (3.1). By Lemma 3.4, there exist $v_\beta\in BV_g\cap L^\infty(\Omega)$ such that

$$\Phi_\lambda(v_\beta) = \Phi_{\lambda,g}(v_\beta) \xrightarrow{\beta\to 0} \Phi_{\lambda,g}(u) = \min_{v\in BV\cap L^2(\Omega)} \Phi_{\lambda,g}(v),$$

and thus

$$\inf_{v\in BV_g\cap L^2(\Omega)} \Phi_\lambda(v) \le \min_{v\in BV\cap L^2(\Omega)} \Phi_{\lambda,g}(v).$$

Thus, the theorem holds. ∎

4. The flow related to minimization problem (3.1).

4.1. Motivation for the weak solution. Due to the boundary term, the lower semicontinuity of $\Phi_{\lambda,g}$ with respect to the L^2-norm is not clear. Therefore, the notion of solution using the theory of maximal monotone operators [12] cannot be directly applied. However, we can establish the existence and uniqueness of the solution in the following sense. Suppose that

$$v\in L^2(0,T;H^1(\Omega)) \tag{4.1}$$

and that u is a classical solution of (1.8)–(1.10). Multiplying (1.8) by $(v-u)$, integrating over Ω, and using the convexity of ϕ, we have that

$$\int_\Omega \dot{u}(v-u)dx + \Phi_{\lambda,g}(v) \geq \Phi_{\lambda,g}(u). \tag{4.2}$$

Integrating over $[0,s]$ for any $s \in [0,T]$ then yields

$$\int_0^s \int_\Omega \dot{u}(v-u)dxdt + \int_0^s \Phi_{\lambda,g}(v)dt \geq \int_0^s \Phi_{\lambda,g}(u)dt. \tag{4.3}$$

On the other hand, setting $v = u + \epsilon w$ in (4.3) with $w \in C_0^\infty(\Omega)$ makes it clear that

$$\int_0^s \int_\Omega \dot{u}(\epsilon w)dxdt + \int_0^s \Phi_{\lambda,g}(u+\epsilon w)dt$$

attains a minimum at $\epsilon = 0$. Therefore, if u satisfies (4.3) and $u \in L^2(0,T;BV \cap L^2(\Omega))$ with $\dot{u} \in L^2(\Omega^T)$, u is also a solution of (1.8)–(1.10) in the sense of distribution. This motivates the following definition.

DEFINITION 4.1. *We say that a function $u \in L^2(0,T;BV \cap L^2(\Omega))$ with $\dot{u} \in L^2(\Omega^T)$ is a pseudosolution of* (1.8)–(1.10) *if*

1. $u(x,0) = I(x)$ *on* Ω, *and*
2. u *satisfies* (4.3) *for all* $s \in [0,T]$ *and* $v \in L^2(0,T;BV \cap L^2(\Omega))$.

4.2. The approximate functional ϕ^ϵ. For $\epsilon > 0$, define

$$\phi^\epsilon(x,r) := \begin{cases} \frac{1}{q(x)}|r|^{q(x)}, & |r| \leq 1, \\ \frac{1}{1+\epsilon}|r|^{(1+\epsilon)} - \frac{q(x)-(1+\epsilon)}{(1+\epsilon)q(x)}, & |r| > 1. \end{cases} \tag{4.4}$$

Remark 4.2. We note the following properties, as they will be useful in later computations:

1. $\phi^\epsilon(x,r)$ is convex in r.
2. $\phi^\epsilon_r(x,r) \cdot r \geq 0$ for all $r \in \mathbb{R}^n$.
3. $\phi(x,r) \leq \phi^\epsilon(x,r)$ for all $r \in \mathbb{R}^n$.

To prove the existence and uniqueness of the pseudosolution of (1.8)–(1.10), we first study solutions of the approximate problem

$$\dot{u} - \epsilon\Delta u - \operatorname{div}(\phi^\epsilon_r(x,\nabla u)) + \lambda(u-I) = 0 \quad \text{in } \Omega \times [0,T], \tag{4.5}$$

$$u(x,t) = g(x) \quad \text{in } \partial\Omega \times [0,T], \tag{4.6}$$

$$u(x,0) = \tilde{I}(x) \quad \text{in } \Omega, \tag{4.7}$$

where $\tilde{I} \in H^1 \cap L^\infty(\Omega)$, $g \in L^\infty(\partial\Omega)$ with $g = TrG$ on $\partial\Omega$ for some $G \in H^1(\Omega)$, and $\tilde{I}|_{\partial\Omega} = g$.

LEMMA 4.3. *Suppose $\tilde{I} \in H^1(\Omega)$ with $\tilde{I}|_{\partial\Omega} = g$. Then problem* (4.5)–(4.7) *has a unique solution $u \in L^2(0,T;H^1(\Omega)) \cap C(0,T;L^2(\Omega))$ with $\dot{u} \in L^2(0,T;L^2(\Omega))$ such that*

$$\int_0^\infty \int_\Omega |\dot{u}|^2 dxdt + \sup_{t>0}\left[\int_\Omega \frac{\epsilon}{2}|\nabla u|^2 + \phi^\epsilon(x,\nabla u) + \frac{\lambda}{2}|u-\tilde{I}|^2\right] \leq \int_\Omega \frac{\epsilon}{2}|\nabla\tilde{I}|^2 + \phi^\epsilon(x,\nabla\tilde{I})dx. \tag{4.8}$$

Proof. Since (4.5) is uniformly parabolic, we can conclude this lemma by standard results for parabolic equations [24] and the corresponding energy estimate. ∎

4.3. Estimates for the solution of the approximate problem.

LEMMA 4.4. *If $\tilde{I} \in H^1 \cap L^\infty \cap BV(\Omega)$, $g \in L^\infty(\partial\Omega)$ with $\tilde{I}|_{\partial\Omega} = g$, and u is a solution of* (4.5)–(4.7)*, then*

$$\|u\|_{L^\infty(\Omega^T)} \le \max(\|\tilde{I}\|_{L^\infty(\Omega)}, \|g\|_{L^\infty(\partial\Omega)}). \tag{4.9}$$

Proof. Let $M := \max(\|\widetilde{I}\|_{L^\infty(\Omega)}, \|g\|_{L^\infty(\partial\Omega)})$. Multiply (4.5) by $(u-M)_+$, where

$$(u-M)_+ = \begin{cases} u-M & \text{if } u-M \ge 0, \\ 0 & \text{otherwise,} \end{cases}$$

and integrate over Ω to get

$$\int_\Omega \dot{u}(u-M)_+ dx + \epsilon \int_\Omega |\nabla u|^2 dx + \int_\Omega \phi_r^\epsilon(x,\nabla u)\cdot\nabla u dx + \lambda \int_\Omega (u-\tilde{I})(u-M)_+ dx = 0. \tag{4.10}$$

By property 2 of Remark 4.2, we have that $\int_\Omega \phi_r^\epsilon(x,\nabla u)\cdot\nabla u dx \ge 0$, and thus

$$\frac{1}{2}\int_\Omega \frac{d}{dt}(u-M)_+^2 dx \le 0.$$

Therefore, $\frac{1}{2}\int_\Omega (u-M)_+^2 dx$ is decreasing in t, and since

$$\frac{1}{2}\int_\Omega (u-M)_+^2 dx \ge 0 \quad \text{and} \quad \frac{1}{2}\int_\Omega (u-M)_+^2 dx|_{t=0} = 0,$$

we have that

$$\frac{1}{2}\int_\Omega (u-M)_+^2 dx = 0 \quad \text{for all t} \in [0, \mathrm{T}],$$

and thus

$$u(t) \le M = \max(\|\tilde{I}\|_{L^\infty(\Omega)}, \|g\|_{L^\infty(\partial\Omega)}) \ \mathcal{L}\text{-a.e. on } \Omega \text{ for all } t > 0.$$

Multiplying (4.5) by $(u+M)_+$, a similar argument yields that $u(t) \ge -M$ for all t. Equation (4.9) follows directly. ∎

LEMMA 4.5. *Let u be the solution of* (4.5)–(4.7)*. Then for all $v \in L^2(0,T;H^1(\Omega))$ with $v|_{\partial\Omega} = g$,*

$$\begin{aligned} \int_0^s \int_\Omega \dot{u}(v-u) + \frac{\epsilon}{2}|\nabla v|^2 + \phi^\epsilon(x,Dv) + \frac{\lambda}{2}|v-\tilde{I}|^2 dx dt \\ \ge \int_0^s \int_\Omega \frac{\epsilon}{2}|\nabla u|^2 + \phi^\epsilon(x,Du) + \frac{\lambda}{2}|u-\tilde{I}|^2 dx dt. \end{aligned} \tag{4.11}$$

Proof. Multiplying (4.5) by $v-u$, then integrating by parts and using the convexity of $\phi^\epsilon(x,r)$ in r, we see that (4.11) follows. ∎

Remark 4.6. Let u be the solution of (4.5)–(4.7). Then for any $0 < \epsilon < \alpha - 1$ (where $1 < \alpha \le q(x)$), we have the following estimate, which is a direct consequence of Lemma 4.3:

$$\int_0^\infty \int_\Omega |\dot{u}|^2 dx dt + \sup_{t>0}\left[\int_\Omega \frac{1}{1+\epsilon}|\nabla u|^{1+\epsilon} dx + \frac{\lambda}{2}\int_\Omega |u-\tilde{I}|^2 dx\right] \le C, \tag{4.12}$$

where $C > 0$ is a constant depending only on Ω and $\|\nabla\tilde{I}\|_{L^2(\Omega)}$.

4.4. Existence and uniqueness of (1.8)–(1.10). Suppose I is given as in Theorem 3.2. By Lemma 2.4 and Remark 2.5, there exists a sequence $\{I_\delta\}$ in $H^1 \cap L^\infty(\Omega)$ such that

$$TrI_\delta = TrI \quad \text{on } \partial\Omega, \tag{4.13}$$

$$\|I_\delta\|_{L^\infty(\Omega)} \le C\|I\|_{L^\infty(\Omega)}, \tag{4.14}$$

$$I_\delta \to I \quad \text{strongly in } L^2(\Omega) \quad \text{as } \delta \to 0, \tag{4.15}$$

and

$$\Phi_{\lambda,g}(I_\delta) \le \Phi_{\lambda,g}(I) + \delta. \tag{4.16}$$

THEOREM 4.7 (*existence and uniqueness*). *Suppose $I \in BV \cap L^\infty(\Omega)$, $g \in L^\infty(\partial\Omega)$, and $I|_{\partial\Omega} = g$ with $g = TrG$ for some function $G \in H^1(\Omega)$. Then there exists a unique pseudosolution u of* (1.8)–(1.10) *in the sense of Definition* 4.1.

Proof. Step 1. First, we fix $\delta > 0$ and pass to the limit $\epsilon \to 0$.

Let $\{u_\delta^\epsilon\}$ be the sequence of solutions to (4.5)–(4.7) with initial data $\tilde{I} = I_\delta$. By Lemma 4.4 and Remark 4.6, there exists a subsequence $\{u_\delta^{\epsilon_i}\}$ and a function $u_\delta \in L^\infty(\Omega^\infty)$ with $\dot{u}_\delta \in L^2(\Omega^\infty)$ such that as $\epsilon_i \to 0$,

$$u_\delta^{\epsilon_i} \rightharpoonup u_\delta \quad \text{weakly}^* \text{ in } L^\infty(\Omega^\infty), \tag{4.17}$$

$$\dot{u}_\delta^{\epsilon_i} \rightharpoonup w \quad \text{weakly in } L^2(\Omega^\infty). \tag{4.18}$$

The same argument used in the proof of Lemma 3.1 in [37] gives us that $\dot{u}_\delta = w$ and $u_\delta(0) = I_\delta$.

Moreover, for all $f \in L^2(\Omega)$,

$$\begin{aligned}\int_\Omega (u_\delta^{\epsilon_i}(\cdot,t) - I_\delta) f(x)dx &= \int_0^\infty \int_\Omega \dot{u}_\delta^{\epsilon_i}(x,s) 1_{[0,t]}(s) f(x)dxds \\ &\xrightarrow{\epsilon_i \to 0} \int_0^\infty \int_\Omega \dot{u}_\delta(x,s) 1_{[0,t]}(s) f(x)dxds \\ &= \int_\Omega (u_\delta(\cdot,t) - I_\delta) f(x)dx.\end{aligned}$$

Therefore, for each $t > 0$,

$$u_\delta^{\epsilon_i}(\cdot,t) \rightharpoonup u_\delta(\cdot,t) \quad \text{weakly in } L^2(\Omega). \tag{4.19}$$

From (4.12), for each $t > 0$, $\{u_\delta^{\epsilon_i}\}$ is a bounded sequence in $W^{1,1}(\Omega)$. Therefore, there exists a convergent subsequence $\{u_\delta^{\epsilon_{i_j}}\}$ of $\{u_\delta^{\epsilon_i}\}$ such that

$$u_\delta^{\epsilon_{i_j}}(\cdot,t) \to u_\delta(\cdot,t) \quad \text{strongly in } L^1(\Omega).$$

Note that every convergent subsequence of $\{u_\delta^{\epsilon_i}\}$ converges to the same limit $u_\delta(\cdot,t)$ due to (4.19). Then, for each $t > 0$,

$$u_\delta^{\epsilon_i}(\cdot,t) \to u_\delta(\cdot,t) \quad \text{strongly in } L^1(\Omega). \tag{4.20}$$

From (4.17) and (4.20), we have that for each $t > 0$,

$$u_\delta^{\epsilon_i}(\cdot,t) \to u_\delta(\cdot,t) \quad \text{strongly in } L^2(\Omega). \tag{4.21}$$

We also have from (4.12) and (4.20) that $u_\delta \in L^\infty(0,\infty, BV \cap L^\infty(\Omega))$ with $\dot{u}_\delta \in L^2(\Omega^\infty)$. Furthermore, by Lemma 4.5, for all $v \in L^2(0,T;H^1(\Omega))$ with $v|_{\partial\Omega} = g$,

$$\begin{aligned}(4.22)\qquad &\int_0^s \int_\Omega \dot{u}_\delta^{\epsilon_i}(v - u_\delta^{\epsilon_i}) + \frac{\epsilon_i}{2}|\nabla v|^2 + \phi^{\epsilon_i}(x,\nabla v) + \frac{\lambda}{2}|v - I_\delta|^2 dxdt \\ &\quad \geq \int_0^s \int_\Omega \frac{\epsilon_i}{2}|\nabla u_\delta^{\epsilon_i}|^2 + \phi^{\epsilon_i}(x, Du_\delta^{\epsilon_i}) + \frac{\lambda}{2}|u_\delta^{\epsilon_i} - I_\delta|^2 dxdt.\end{aligned}$$

Using (4.18) and (4.21), we can let $\epsilon_i \to 0$ in (4.22) to get

$$\begin{aligned}(4.23)\qquad &\int_0^s \int_\Omega \dot{u}_\delta(v - u_\delta) + \phi(x,\nabla v) + \frac{\lambda}{2}|v - I_\delta|^2 dxdt \\ &\quad \geq \lim_{\epsilon_i \to 0} \int_0^s \int_\Omega \phi^{\epsilon_i}(x, Du_\delta^{\epsilon_i}) dxdt + \frac{\lambda}{2}\int_0^s \int_\Omega |u_\delta - I_\delta|^2 dxdt.\end{aligned}$$

By Lemma 2.3, weak lower semicontinuity (w.l.s.c.), and property 3 of Remark 4.2,

$$(4.24)\qquad \lim_{\epsilon_i \to 0} \int_0^s \int_\Omega \phi^{\epsilon_i}(x, Du_\delta^{\epsilon_i}) dxdt \geq \int_0^s \int_\Omega \phi(x, Du_\delta) dxdt + \int_0^s \int_{\partial\Omega} |u_\delta - g| d\mathcal{H}^{n-1} dt.$$

The combination of (4.23) and (4.24) gives us

$$\begin{aligned}&\int_0^s \int_\Omega \dot{u}_\delta(v - u_\delta) + \phi(x,\nabla v) + \frac{\lambda}{2}|v - I_\delta|^2 dxdt + \int_0^s \int_{\partial\Omega} |v - g| d\mathcal{H}^{n-1} dt \\ (4.25)\qquad &\geq \int_0^s \int_\Omega \phi(x,\nabla u_\delta) + \frac{\lambda}{2}|u_\delta - I_\delta|^2 dxdt + \int_0^s \int_{\partial\Omega} |u_\delta - g| d\mathcal{H}^{n-1} dt\end{aligned}$$

for all $v \in L^2(0,\infty;H^1(\Omega))$ with $v = g$ on $\partial\Omega^\infty$. By approximation, (4.25) still holds for $v \in L^2(0,\infty;BV \cap L^\infty(\Omega))$ with $v = g$ on $\partial\Omega^\infty$. To see that (4.25) holds for all $v \in L^2(0,\infty;BV \cap L^\infty(\Omega))$ (in particular, without $v = g$ on $\partial\Omega^\infty$), replace v with $v_\beta = d_\beta v + (1 - d_\beta)G$ in (4.25) and (by Lemma 3.4) let $\beta \to 0$. By approximation, we can conclude that (4.25) holds for all $v \in L^2(0,\infty;BV \cap L^2(\Omega))$.

Step 2. Now it remains only to pass to the limit as $\delta \to 0$ in (4.25) to complete the proof. First note that (4.8) holds for $u_\delta^{\epsilon_i}$ with $\tilde{I} = I_\delta$. Fix $\delta > 0$. By w.l.s.c. and (4.20), the same argument used to deduce (4.24) also gives us that

$$\lim_{\epsilon_i \to 0} \int_\Omega \phi^{\epsilon_i}(x, Du_\delta^{\epsilon_i}) dx \geq \int_\Omega \phi(x, Du_\delta) dx + \int_{\partial\Omega} |u_\delta - g| d\mathcal{H}^{n-1}.$$

Therefore, we can pass to the limit as $\epsilon_i \to 0$ in (4.8) and get

$$\begin{aligned}(4.26)\qquad &\int_0^\infty \int_\Omega |\dot{u}_\delta|^2 dxdt + \sup_{t>0}\left[\int_\Omega \phi(x, Du_\delta) dxdt + \frac{\lambda}{2}|u_\delta - I_\delta|^2 + \int_{\partial\Omega} |u_\delta - g| d\mathcal{H}^{n-1} dt\right] \\ &\qquad \leq \int_\Omega \phi(x, \nabla I_\delta) dx.\end{aligned}$$

Then $\{u_\delta\}$ is uniformly bounded in $W^{1,1}(\Omega)$ for each $t > 0$. Note also that in Lemma 4.4 the bound is independent of both ϵ and δ. Therefore, $\{u_\delta\}$ is also uniformly bounded in $L^\infty(\Omega^\infty)$. From (4.26), we also have that $\{\dot{u}_\delta\}$ is uniformly bounded in $L^2(\Omega^\infty)$.

By the same argument used to obtain (4.17), (4.18), and (4.21), there exists a subsequence $\{u_{\delta_j}\}$ of $\{u_\delta\}$ and a function $u \in L^\infty(0,\infty, BV \cap L^\infty(\Omega))$ with $\dot{u} \in L^2(\Omega^\infty)$ such that as $\delta_j \to 0$,

$$u_{\delta_j} \rightharpoonup u \quad \text{weakly}^* \text{ in } L^\infty(\Omega^\infty), \tag{4.27}$$

$$\dot{u}_{\delta_j} \rightharpoonup \dot{u} \quad \text{weakly in } L^2(\Omega^\infty), \tag{4.28}$$

$$u_{\delta_j}(\cdot,t) \to u(\cdot,t) \quad \text{strongly in } L^2(\Omega) \text{ and uniformly in } t. \tag{4.29}$$

Let $\delta = \delta_j$ in (4.25). Using w.l.s.c. and (4.27)–(4.29), we can let $\delta_j \to 0$ in (4.25) to conclude that for all $v \in L^2(0,\infty, BV \cap L^2(\Omega))$,

$$\int_0^s \int_\Omega \dot{u}(v-u)dxdt + \int_0^s \Phi_{\lambda,g}(v)dt \geq \int_0^s \Phi_{\lambda,g}(u)dt.$$

Existence is proved.

Step 3 (*uniqueness*). Suppose that u_1, u_2 are both weak solutions of (1.8)–(1.10). As in [23, 37], we can obtain two inequalities: the first by setting $u = u_1$ and $v = u_2$ in (4.3) and the second by setting $u = u_2$ and $v = u_1$. Adding these two inequalities gives us that for all $s > 0$,

$$\int_0^s \int_\Omega \frac{1}{2}\frac{d}{dt}|u_1 - u_2|^2 dxdt \leq 0.$$

Therefore, $u_1 = u_2$ in Ω^∞. ∎

4.5. Behavior as $t \to \infty$.

THEOREM 4.8. *As $t \to \infty$, the weak solution, $u(x,t)$, of* (1.8)–(1.10) *converges strongly in $L^2(\Omega)$ to a minimizer $\tilde{u}$ of the function $\Phi_{\lambda,g}$, i.e., the pseudosolution, $\tilde{u}$, of* (3.1).

Proof. Since u satisfies (4.3), for any $s > 0$ we can substitute $v(x) \in BV \cap L^2(\Omega)$ into (4.3) as follows:

$$\begin{aligned}\int_\Omega (u(x,s) - I(x))v(x)dx - \frac{1}{2}\int_\Omega (u^2(x,s) - I^2(x))dx + s\int_\Omega \phi(x,\nabla v) \\ + s\frac{\lambda}{2}\int_\Omega |v-I|^2 dx + s\int_{\partial\Omega} |v-g| d\mathcal{H}^{n-1} \\ \geq \int_0^s \int_\Omega \phi(x,\nabla u)dt + \frac{\lambda}{2}\int_0^s \int_\Omega |u-I|^2 dxdt + \int_0^s \int_{\partial\Omega} |u-g| d\mathcal{H}^{n-1} dt.\end{aligned} \tag{4.30}$$

Proceeding as in [19], let

$$w(x,s) = \frac{1}{s}\int_0^s u(x,t)dt.$$

Since $u \in L^\infty(0,\infty; BV \cap L^\infty(\Omega))$, for each $s > 0$ we have that $w(\cdot,s) \in BV \cap L^\infty(\Omega)$ with $\{w(\cdot,s)\}$ uniformly bounded in $BV(\Omega)$ and $L^\infty(\Omega)$. Therefore, there exists a subsequence $\{w(\cdot,s_i)\}$ of $\{w(\cdot,s)\}$ which converges strongly in $L^1(\Omega)$ and weakly in $BV(\Omega)$ and $L^\infty(\Omega)$ to a function $\tilde{u} \in BV \cap L^\infty(\Omega)$ as $s_i \to \infty$. Since $\{w(\cdot,s)\}$ is uniformly bounded in $L^\infty(\Omega)$, $\{w(\cdot,s_i)\}$ also converges strongly in $L^2(\Omega)$ to $\tilde{u}$.

Dividing (4.30) by s and taking the limit along $s_i \to \infty$ gives us that

$$\int_\Omega \phi(x,\nabla v) + \frac{\lambda}{2}\int_\Omega |v-I|^2 dx + \int_{\partial\Omega} |v-g| d\mathcal{H}^{n-1} \geq \int_\Omega \phi(x,\nabla \tilde{u}) + \frac{\lambda}{2}\int_\Omega |\tilde{u}-I|^2 dx + \int_{\partial\Omega} |\tilde{u}-g| d\mathcal{H}^{n-1}$$

for all $v \in BV \cap L^2(\Omega)$; i.e., $\tilde{u}$ is a pseudosolution of (3.1). ∎

5. Numerical methods and experimental results. We solve the minimization problem (1.3) numerically using the flow of its associated Euler–Lagrange equation,

$$\frac{\partial u}{\partial t} - \operatorname{div}(\phi_r(x,Du)) + \lambda(u-I) = 0 \quad \text{in } \Omega\times[0,T], \tag{5.1}$$

$$\frac{\partial u}{\partial n}(x,t) = 0 \quad \text{on } \partial\Omega\times[0,T], \tag{5.2}$$

$$u(0) = I \quad \text{in } \Omega. \tag{5.3}$$

To approximate (5.1), we use an explicit finite difference scheme. The degenerate diffusion term,

$$\operatorname{div}(\phi_r(x,\nabla u)) = |\nabla u|^{p(x)-2}\left[(p(x)-1)\Delta u + (2-p(x))|\nabla u|\operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right) + \nabla p\cdot\nabla u \log|\nabla u|\right]$$

with

$$p(x) = \begin{cases} q(x) \equiv 1 + \frac{1}{1+k|\nabla G_\sigma * I(x)|^2}, & |\nabla u| < \beta, \\ 1, & |\nabla u| \geq \beta \end{cases} \tag{5.4}$$

for $k,\sigma > 0$, and G_σ the Gaussian filter, is approximated as follows:

- The coefficient, $|\nabla u|^{p(x)-2}$, is approximated using central differences;
- the isotropic diffusion term, Δu, is approximated using central differences;
- the curvature term, $|\nabla u|\operatorname{div}\big(\frac{\nabla u}{|\nabla u|}\big)$, is approximated using the minmod scheme for $\operatorname{div}\big(\frac{\nabla u}{|\nabla u|}\big)$ (see [31]) and central differences for $|\nabla u|$;
- if $|\nabla u| \neq 0$, the hyperbolic term $\nabla p\cdot\nabla u \log|\nabla u|$ is computed using an upwind scheme for $\nabla p\cdot\nabla u$ (see [29]) and central differences for $\log|\nabla u|$. Otherwise, the hyperbolic term is set to zero.

We found that the behavior of (1.3) is an innate behavior of the model, and variants on this numerical scheme also yield very good results.

We compared our model with the flow of the Euler–Lagrange equation associated with (1.2) (modified only by a fidelity term),

$$\frac{\partial u}{\partial t} = (p-1)\Delta u + (2-p)\operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right) - \lambda(u-I), \quad \text{where } p = \begin{cases} 2, & |\nabla u| < \beta, \\ 1, & |\nabla u| \geq \beta. \end{cases} \tag{5.5}$$

An explicit finite difference scheme was used, where central differences were used to implement Δu and the minmod scheme [31] was used to implement $\operatorname{div}(\frac{\nabla u}{|\nabla u|})$.

All of the images ranged in intensity from 0 to 255. The parameters $\lambda = .05$, $\sigma = .5$, $k = .0075$ and time step $= .05$ consistently yielded optimal results for all of the models we tested here. We compared various thresholds, β, to test the sensitivity of both models (1.2) and (1.3) to this parameter. All of the "edge maps" in the figures to follow were computed using the function

$$\textit{edge map of } u\text{:} \quad \frac{1}{1 + k|\nabla G_\sigma * u(x)|^2} \tag{5.6}$$

with $k = .0075$ (the same value of k is also used to compute the exponent $p(x)$ in (5.4)). We found that this value also gave the clearest edge map for each model. The number of iterations was chosen large enough so that the standard deviation between subsequent images was at most .005.

In Figure 5.1 we illustrate the proposed model's ability to reconstruct piecewise smooth functions while avoiding the staircasing effect. The first row from the top contains a piecewise smooth function plotted as both an image and a surface, and also contains its edge map (5.6). The surface is viewed from two different orientations: the first view displays the upper left corner of the image at the origin, and the second view displays the same surface rotated 180°. The second row contains the same series of images for the image degraded by Gaussian noise with mean zero. The third, fourth, and fifth rows contain reconstructions using isotropic diffusion only ($p \equiv 2$), TV-based diffusion only ($p \equiv 1$), and the proposed model, respectively. Isotropic diffusion reconstructs smooth regions, but edges are severely blurred. TV-based diffusion reconstructs sharp edges, but the staircasing effect is clearly present. This in turn creates false edges, which could lead to an incorrect segmentation of the image. The proposed model reconstructs sharp edges as effectively as TV-based diffusion *and* recovers smooth regions as effectively as pure isotropic diffusion (in particular, without staircasing).

Figure 5.2 contains a reconstruction of another piecewise smooth image with additive Gaussian noise. Our goal is to once again reconstruct the smooth regions while preserving their boundaries and without introducing false edges. Furthermore, we also wanted to compare the sensitivity of models (1.2) and (1.3) to the threshold, β. The top row shows the original and noisy images with their edge maps (5.6). The bottom three rows contain reconstructions using models (1.2) and (1.3). The first column from the left contains reconstructions using (1.2) with thresholds $\beta = 30, 50,$ and 70, respectively, and the second column contains their corresponding edge maps (5.6). The third column contains reconstructions using TV-based diffusion only ((1.2) or (1.3) with $\beta = 0$) and the proposed model (1.3) with thresholds $\beta = 30$ and 100, respectively. The last column contains their corresponding edge maps (5.6). TV-based diffusion only ($\beta = 0$) shows clear evidence of staircasing, while the proposed model is relatively insensitive to a broad range of thresholds, β. Although the exact behavior of the diffusion changes slightly at likely edges between $\beta = 30$ and $\beta = 100$, the effect on the resulting image is minimal. On the other hand, model (1.2) demonstrates a large change in behavior at both noise and edges across the range of thresholds $\beta = 30, 50,$ and 70. Similar experiments on the noisy radar images in Figures 5.3 and 5.4 yielded very similar results. Note that in Figure 5.3, even fine details, such as the lettering at the bottom of the image, are preserved using the proposed model. In Figure 5.4, the proposed model preserves the boundaries of the land mines as effectively as the TV model without enhancing the background noise.

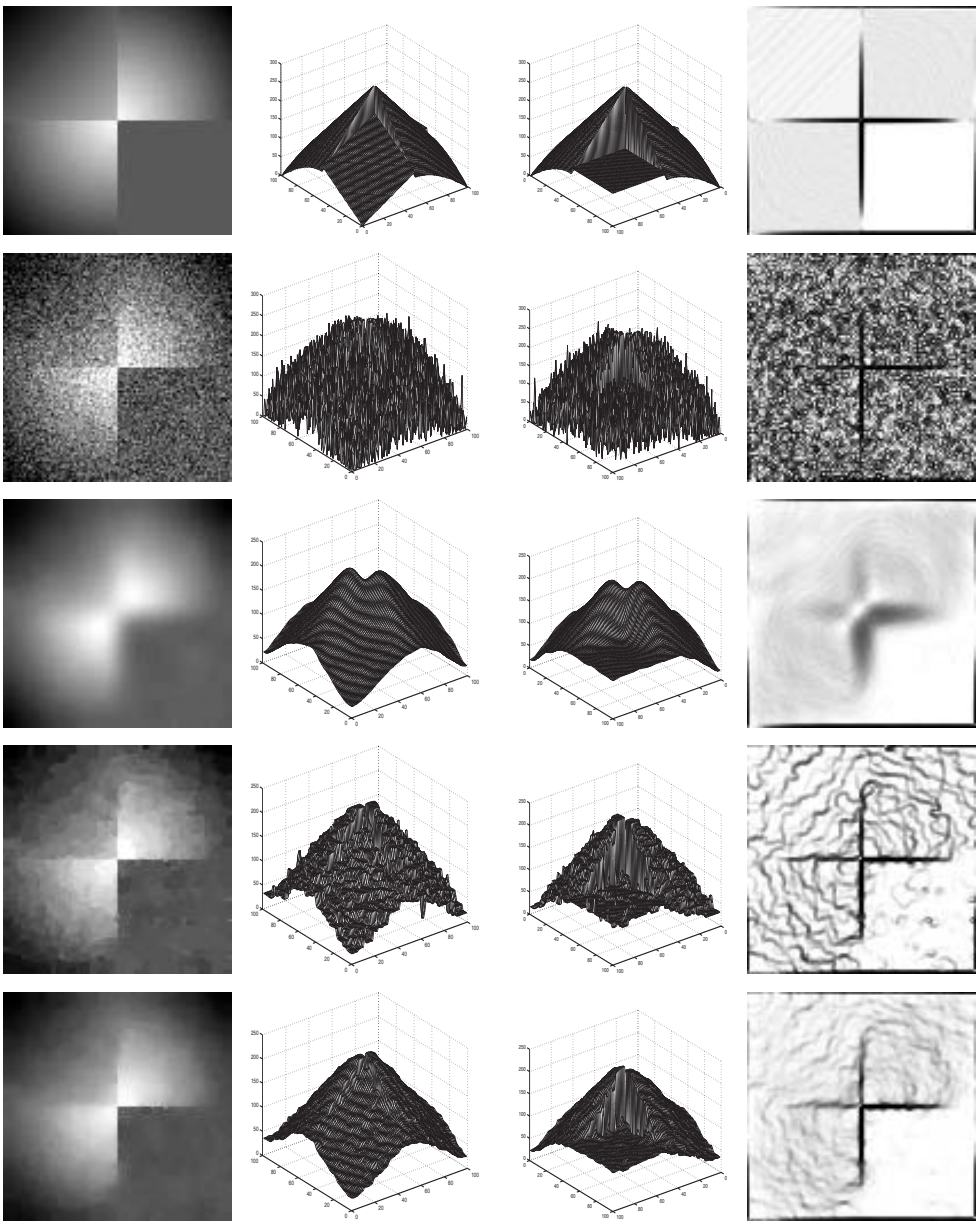


FIG. 5.1. *Top row*: *true image* (100×100), *surface plot of the image, surface rotated* $180°$, *edge map* (5.6). *Second row*: *image + noise. Third row*: *reconstruction using isotropic diffusion only* (200 *iterations*). *Fourth row*: *reconstruction using TV-based diffusion only* (2000 *iterations*). *Fifth row*: *reconstruction using the proposed model* (1000 *iterations*, $\beta = 30$, $k = .0075$).

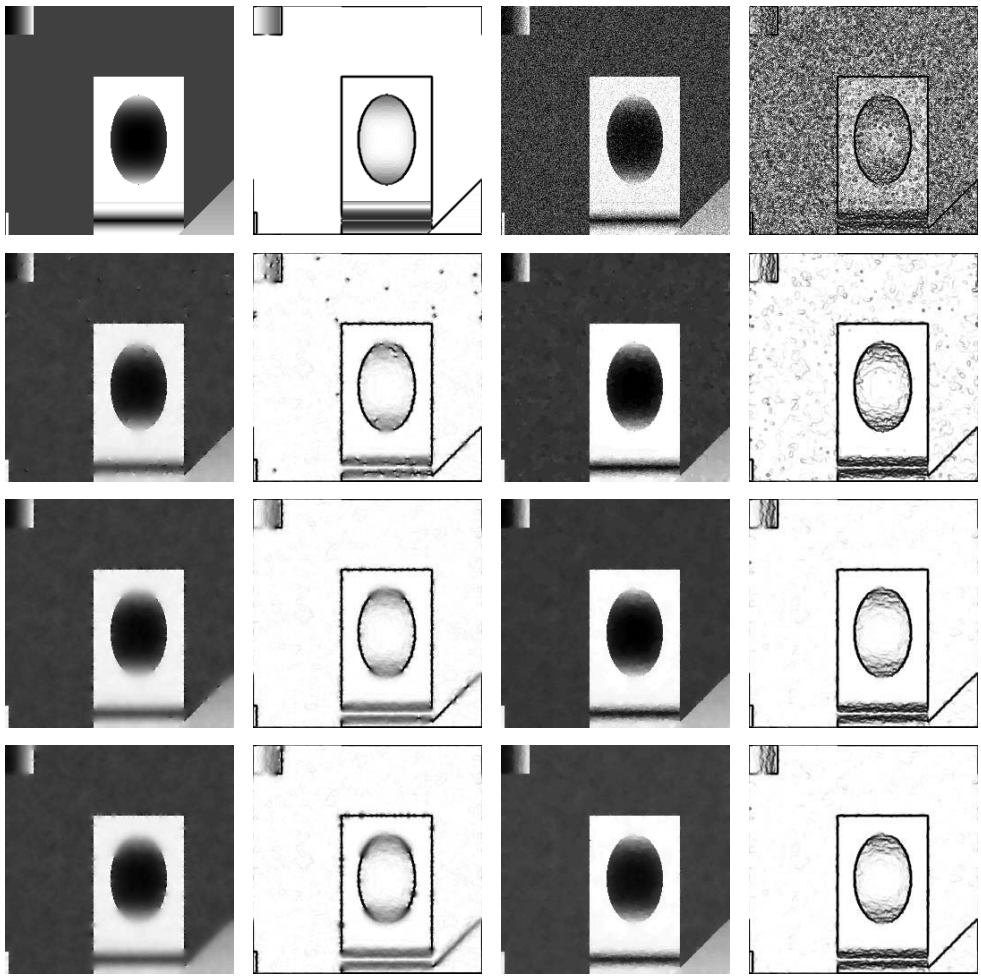


FIG. 5.2. *Top row*: *original piecewise smooth image* (256×256) *and edge map* (5.6)*, image + noise and edge map* (5.6)*. Bottom three rows*: *First column from left: reconstructions using* (1.2) *with thresholds* $\beta = 30, 50, 70$*, respectively* (1000 *iterations*)*. Second column*: *corresponding edge maps* (5.6)*. Third column*: *reconstruction using TV-based diffusion only* (2000 *iterations*) *and the proposed model with thresholds* $\beta = 30, 100$*, respectively* (1000 *iterations*)*. Fourth column*: *corresponding edge maps* (5.6) (*all images:* $k = .0075$*,* $\lambda = .05$*,* $\sigma = .5$*, time step* $= .05$).

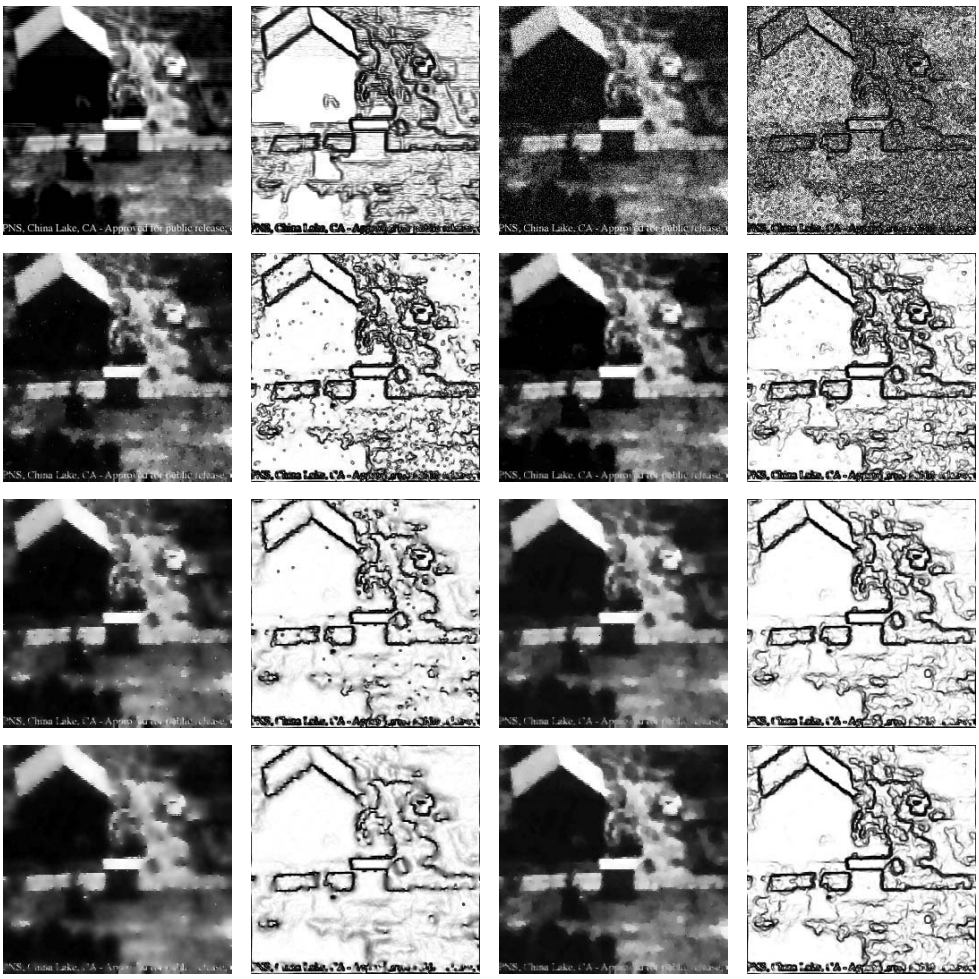


FIG. 5.3. *Top row*: *radar image* (256×256) *and edge map* (5.6); *radar image with Gaussian noise and edge map* (5.6) *with* $k = .0075$. *Bottom three rows*: *First column: reconstructions using* (1.2) *with thresholds* $\beta = 10, 20, 30$, *respectively* (1000 *iterations*). *Second column*: *corresponding edge maps* (5.6). *Third column*: *reconstruction using TV-based diffusion only* (4000 *iterations*) *and the proposed model with thresholds* $\beta = 30, 100$, *respectively* (1000 *iterations*). *Fourth column*: *corresponding edge maps* (5.6) (*all images:* $k = .0075$, $\lambda = .05$, $\sigma = .5$, *time step* $= .05$).

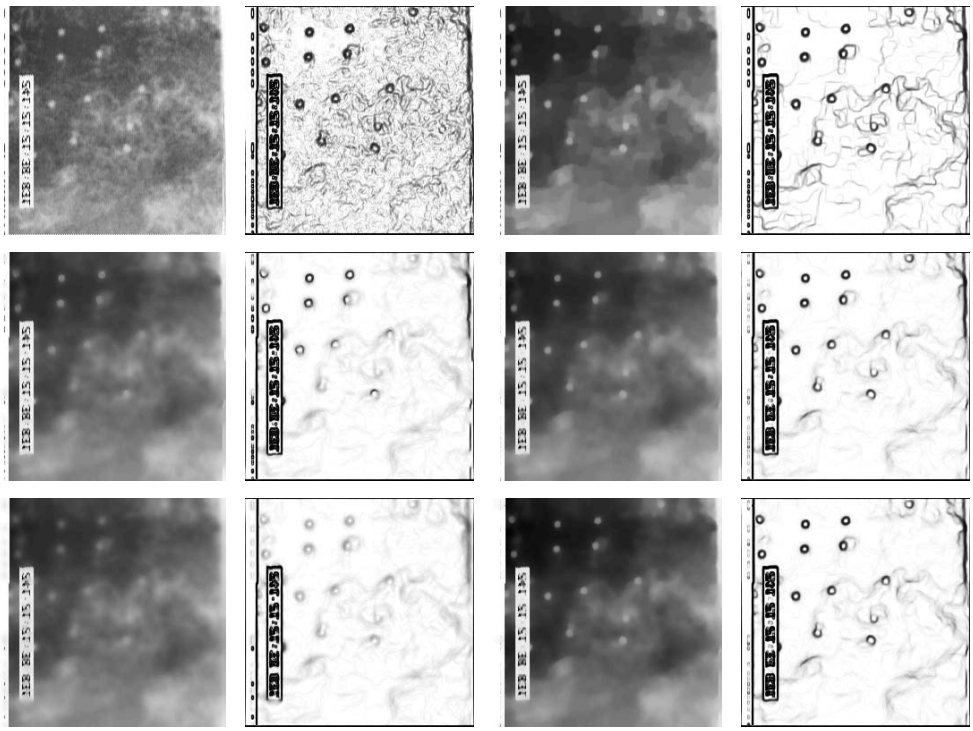


FIG. 5.4. *First column from left*: *radar image of land mines* (256×256); *reconstructions using* (1.2) *with thresholds* $\beta = 10, 20$, *respectively* (750 *iterations*). *Second column*: *corresponding edge maps* (5.6). *Third column*: *reconstruction using TV-based diffusion only* (1000 *iterations*); *reconstructions using the proposed model with thresholds* $\beta = 30, 100$, *respectively* (750 *iterations*). *Fourth column*: *corresponding edge maps* (5.6) (*all images:* $k = .0075$, $\lambda = .05$, $\sigma = .5$, *time step* $= .05$).

Figure 5.5 provides another successful reconstruction of a piecewise smooth image with additive Gaussian noise. TV-based diffusion alone creates false edges, while the proposed model preserved accurate object boundaries while minimizing the creation of false ones. Tests with $\beta = 30$ and 100 again demonstrate the proposed model's insensitivity to the threshold, β. Figure 5.6 displays a similar experiment with an MRI of a human heart. The original image is successfully denoised using the proposed model (top row). We then added more noise and, as in the previous experiments, found that TV alone created false edges, while the proposed model generated much fewer false artifacts.

Figure 5.7 contains several more examples in which the noise in each of the images was acquired directly through acquisition, storage, or transmission. The first row contains a diffusion tensor image (DTI) of a human brain; the second contains a magnetic resonance image (MRI) of a human chest cavity; the third contains a transmission electron microscope (TEM) image of aluminum. In all of these images, the goal is to detect "true" object boundaries without creating any false edges. The proposed model is successful in all of these cases.

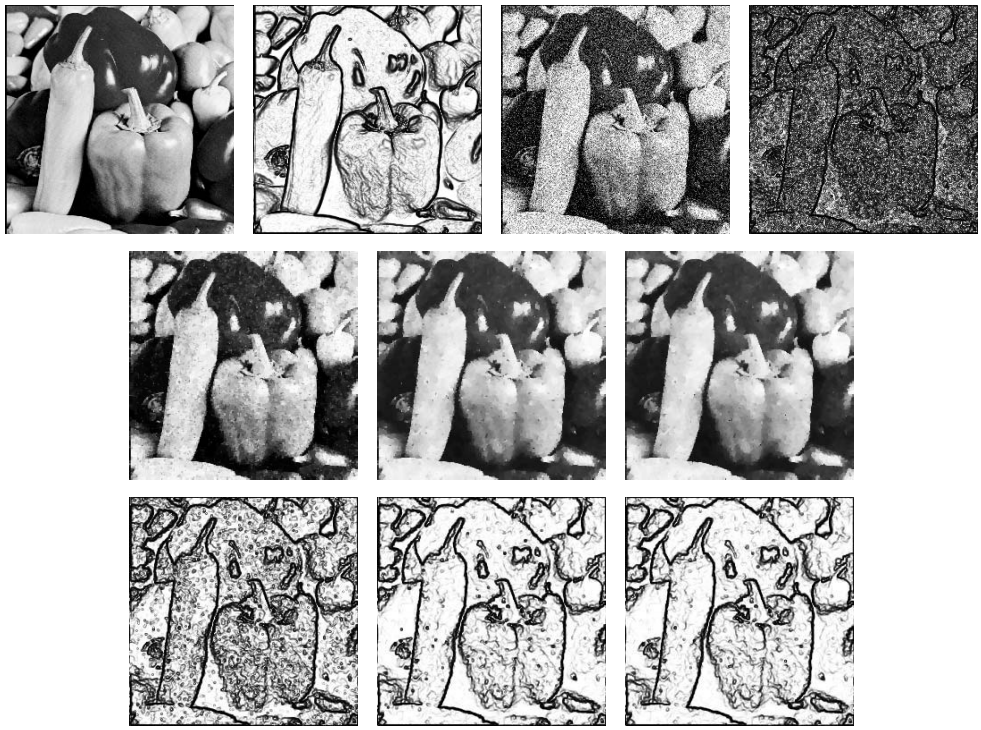


FIG. 5.5. *Top row*: *true image* (256×256) *and edge map* (5.6); *true image + noise and edge map* (5.6). *Second row*: *reconstructions using TV-based diffusion only* (2000 *iterations*) *and the proposed model with thresholds* $\beta = 30, 100$ (1000 *iterations,* $k = .0075$). *Third row*: *corresponding edge maps* (5.6) (*all images:* $k = .0075$, $\lambda = .05$, $\sigma = .5$, *time step* $= .05$).

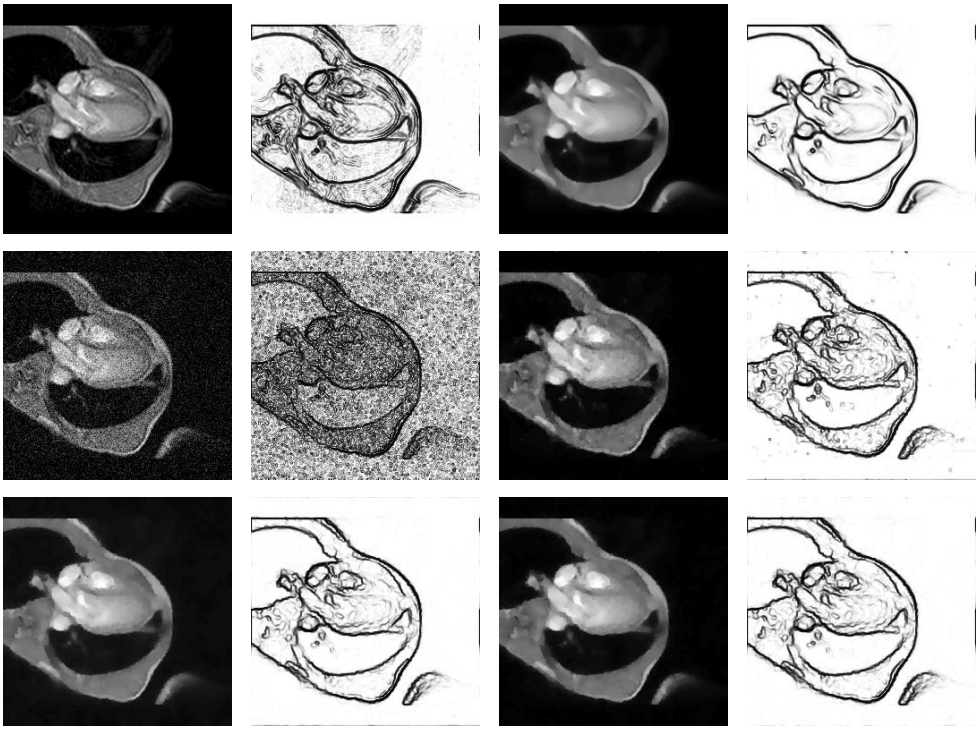


FIG. 5.6. *Top row*: *true image: human heart MRI* (256×256), *corresponding edge map* (5.6), *reconstruction using the proposed model* (*threshold* $\beta = 30$), *edge map* (5.6). *Second row: human heart MRI + noise, edge map* (5.6), *reconstruction using TV-based diffusion only, edge map* (5.6). *Third row*: *reconstruction using the proposed model* (*threshold* $\beta = 30$), *edge map* (5.6), *reconstruction using the proposed model* (*threshold* $\beta = 100$), *edge map* (5.6) (*all images*: $k = .0075$, $\lambda = .05$, $\sigma = .5$, *time step* $= .05$).

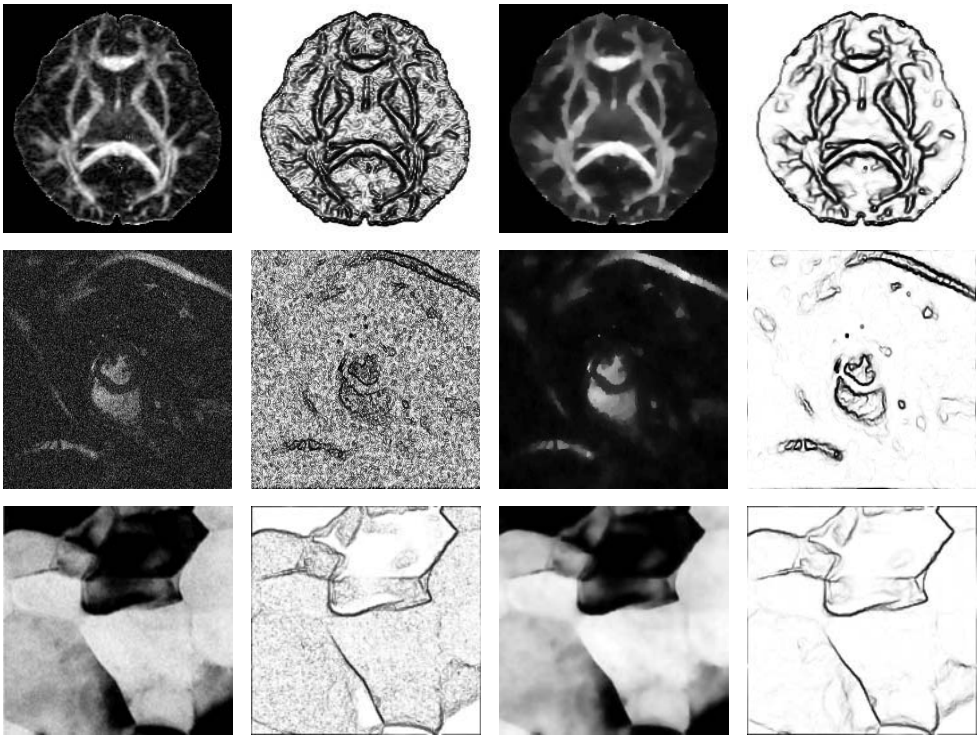


FIG. 5.7. *First column from left*: *true images*: *human brain DTI* (201 × 171), *human chest cavity MRI* (209 × 256), *TEM image of aluminum* (512 × 512). *Second column*: *corresponding edge maps* (5.6). *Third column*: *reconstructions using the proposed model* (420, 1000, 500 *iterations, respectively*). *Fourth column*: *corresponding edge maps* (5.6) *with* $k = .0075$ (*all images*: *threshold* = 30, $k = .0075$, $\lambda = .05$, $\sigma = .5$, *time step* = .05).

Acknowledgments. The authors would like to thank Bernard Mair for the radar images in Figures 5.3 and 5.4; Sebastien Barre for the MRI in Figure 5.6; and Yijun Liu for the DTI, Mark Griswold for the MRI, and Katayun Barmak for the TEM image in Figure 5.7. We would also like to thank the reviewers for their helpful comments.

REFERENCES

[1] R. ACAR AND C. R. VOGEL, *Analysis of bounded variation penalty methods for ill-posed problems*, Inverse Problems, 10 (1994), pp. 1217–1229.

[2] F. ANDREU, C. BALLESTER, V. CASELLES, AND J. M. MAZÓN, *Minimizing total variation flow*, C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 867–872.

[3] F. ANDREU, C. BALLESTER, V. CASELLES, AND J. M. MAZÓN, *The Dirichlet problem for the total variation flow*, J. Funct. Anal., 180 (2001), pp. 347–403.

[4] F. ANDREU, J. M. MAZÓN, J. S. MOLL, AND V. CASELLES, *The minimizing total variation flow with measure initial conditions*, Commun. Contemp. Math., 6 (2004), pp. 431–494.

[5] F. ANDREU-VAILLO, V. CASELLES, AND J. M. MAZÓN, *Parabolic Quasilinear Equations Minimizing Linear Growth Functionals*, Progr. Math. 223, Birkhäuser Verlag, Basel, 2004.

[6] G. AUBERT AND L. VESE, *A variational method in image recovery*, SIAM J. Numer. Anal., 34 (1997), pp. 1948–1979.

[7] G. BELLETTINI, V. CASELLES, AND M. NOVAGA, *The total variation flow in* $\mathbb{R}^N$, J. Differential Equations, 184 (2002), pp. 475–525.

[8] M. BERTALMÍO, G. SAPIRO, V. CASELLES, AND C. BALLESTER, *Image inpainting,* in Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), ACM Press/Addison Wesley, New York, 2000, pp. 417–424.

[9] P. Blomgren, T. F. Chan, P. Mulet, and C. Wong, *Total variation image restoration: Numerical methods and extensions*, in Proceedings of the IEEE International Conference on Image Processing, Vol. III, IEEE, Los Alamitos, CA, 1997, pp. 384–387.
[10] G. Bouchitté and G. Dal Maso, *Integral representation and relaxation of convex local functionals on* BV(Ω), Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 20 (1993), pp. 483–533.
[11] G. Bouchitté and M. Valadier, *Integral representation of convex functionals on a space of measures*, J. Funct. Anal., 80 (1988), pp. 398–420.
[12] H. Brézis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North-Holland Mathematics Studies 5, Notas de Matemática (50), North-Holland, Amsterdam, 1973.
[13] V. Caselles, J.-M. Morel, and C. Sbert, *An axiomatic approach to image interpolation*, IEEE Trans. Image Process., 7 (1998), pp. 376–386.
[14] V. Caselles, and L. Rudin, *Multiscale total variation*, in Proceedings of the Second European Conference on Image Processing (Palma, Spain, September, 1995), pp. 376–386.
[15] A. Chambolle and P.-L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188.
[16] T. Chan, A. Marquina, and P. Mulet, *High-order total variation-based image restoration*, SIAM J. Sci. Comput., 22 (2000), pp. 503–516.
[17] T. F. Chan, S. H. Kang, and J. Shen, *Euler's elastica and curvature-based inpainting*, SIAM J. Appl. Math., 63 (2002), pp. 564–592.
[18] T. F. Chan and J. Shen, *Mathematical models for local nontexture inpaintings*, SIAM J. Appl. Math., 62 (2002), pp. 1019–1043.
[19] Y. Chen and M. Rao, *Minimization problems and associated flows related to weighted p energy and total variation*, SIAM J. Math. Anal., 34 (2003), pp. 1084–1104.
[20] D. C. Dobson and F. Santosa, *Recovery of blocky images from noisy and blurred data*, SIAM J. Appl. Math., 56 (1996), pp. 1181–1198.
[21] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.
[22] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Monogr. Math. 80, Birkhäuser Verlag, Basel, 1984.
[23] R. Hardt and X. Zhou, *An evolution problem for linear growth functionals*, Comm. Partial Differential Equations, 19 (1994), pp. 1879–1907.
[24] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1967.
[25] A. Lichnewsky and R. Temam, *Pseudosolutions of the time-dependent minimal surface problem*, J. Differential Equations, 30 (1978), pp. 340–364.
[26] M. Lysaker, A. Lundervold, and X.-C. Tai, *Noise removal using fourth-order partial differential equation with applications to medical magnetic resonance images in space and time*, IEEE Trans. Image Process., 12 (2003), pp. 1579–1590.
[27] S. Masnou and J. M. Morel, *Level lines based disocclusion*, in Proceedings of the IEEE International Conference on Image Processing (Chicago, IL, October 4–7, 1998), Vol. III, IEEE, Los Alamitos, CA, pp. 259–263.
[28] M. Nikolova, *Weakly constrained minimization: Application to the estimation of images and signals involving constant regions*, J. Math. Imaging Vision, 21 (2004), pp. 155–175.
[29] S. Osher and J. A. Sethian, *Propagating with curvature-dependent speed: Algoritms based on Hamilton–Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.
[30] W. Ring, *Structural properties of solutions to total variation regularization problems*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 799–810.
[31] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[32] D. M. Strong and T. F. Chan, *Spatially and Scale Adaptive Total Variation Based Regularization and Anisotropic Diffusion in Image Processing*, Technical Report CAM96-46, University of California, Los Angeles, CA, 1996. Available online at http://www.math.ucla.edu/applied/cam/index.html.
[33] L. Vese, *A study in the BV space of a denoising-deblurring variational problem*, Appl. Math. Optim., 44 (2001), pp. 131–161.
[34] R. T. Whitaker and S. M. Pizer, *A mulit-scale approach to nonuniform diffusion*, Comput. Vision Graph. Image Process. Image Understand., 57 (1993), pp. 99–110.
[35] Y.-L. You and M. Kaveh, *Fourth-order partial differential equations for noise removal*, IEEE Trans. Image Process., 9 (2000), pp. 1723–1730.
[36] Y.-L. You, W. Xu, A. Tannenbaum, and M. Kaveh, *Behavioral analysis of anisotropic diffusion in image processing*, IEEE Trans. Image Process., 5 (1996), pp. 1539–1553.
[37] X. Zhou, *An evolution problem for plastic antiplanar shear*, Appl. Math. Optim., 25 (1992), pp. 263–285.

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1407–1423

THE IDENTIFICATION OF A TIME DEPENDENT SORPTION PARAMETER FROM SOIL COLUMN EXPERIMENTS*

K. RENEE FISTER†, MAEVE L. MCCARTHY†, AND SETH F. OPPENHEIMER‡

Abstract. Soil column studies are used frequently in seeking to understand the behavior of a particular contaminant in a saturated homogeneous soil of a given type. The concentration of the contaminant is modeled by a parabolic partial differential equation. We seek to identify the sorption partitioning coefficient as a function of time from limited boundary data. We discuss an output least squares formulation of the problem with Tikhonov regularization. We explicitly characterize a source condition that determines the rate of convergence of the method. Numerical examples are presented.

Key words. parameter identification, inverse problems, Tikhonov regularization, parabolic partial differential equation, sorption partitioning

AMS subject classifications. 35K57, 35R30, 47A52

DOI. 10.1137/050626302

1. Introduction.

The purpose of this paper is to develop theoretical and numerical approaches for approximating an unknown time dependent parameter in a parabolic partial differential equation, given limited boundary data. The equation we shall be working with is a model of a soil column study. These column studies are used frequently in seeking to understand the behavior of a particular contaminant in a saturated homogeneous soil of a given type. The parameter we are seeking to approximate is the sorption partitioning coefficient. This parameter is a measure of the proportion of contaminant that is bound to the soil. In isothermal situations when there are no other contaminants present, the partitioning coefficient is usually taken to be constant. However, if there is another contaminant, e.g., sea salt, or if the temperature is changing, the partitioning coefficient may change as well. Therefore, strictly speaking, the partitioning coefficient is a function of some physical factor other than time. However, if we understand the controlling physical factor as a function of time, we may treat the partitioning coefficient also as a function of time and deduce the true physical functional relationship after the partitioning coefficient has been found as a function of time.

This approach will allow us considerable savings in time and resources when determining how a partitioning coefficient varies with different physical factors. These savings will result from identifying the partitioning coefficient's dependence on the relevant physical factor by means of a single column experiment rather then a large number of separate batch tests.

Section 2 shall be devoted to a discussion of the model, the simplifying assumptions that are applicable, and various necessary facts about the forward problem.

*Received by the editors March 8, 2005; accepted for publication (in revised form) December 28, 2005; published electronically May 3, 2006.
http://www.siam.org/journals/siap/66-4/62630.html

†Department of Mathematics and Statistics, Murray State University, 6C Faculty Hall, Murray, KY 42071 (renee.fister@murraystate.edu, maeve.mccarthy@murraystate.edu). The first author was supported in part by National Science Foundation grant DMS 0414011. The second author was supported in part by National Science Foundation grant DMS 0209562.

‡Department of Mathematics and Statistics, Mississippi State University, P. O. Drawer MA, Mississippi State, MS 39762 (seth@math.msstate.edu). This author was supported in part by National Institutes of Health grant DHHS/NIH P20 RR17661. This is Center for Environmental Health Sciences Publication 109.

Section 3 establishes identifiability of the sorption partitioning coefficient from the available experimental data and applies an output least squares method with Tikhonov regularization to the parameter identification problem. The method is tested in section 4.

2. Discussion of the model. The general approach to modeling convection-diffusion-sorption models may be found in the work of Leij and Dane [12] and Domenico and Schwartz [6]. Sorption of mixtures was treated by Oppenheimer [15]. A thorough treatment of the modeling of the sorption process may be found in Oppenheimer, Kingery, and Han [16]. Column studies are discussed in detail by Adrian, Ozkan, and Alshawabkeh [1].

We model the one dimensional flow of water with a dissolved contaminant through a soil column. The contaminant can be dissolved in the water or bound, that is sorbed, to the soil. We will assume cylindrical symmetry in the column to reduce the problem to one spatial dimension. The column will be assumed to have length L and is modeled as the interval $[0, L]$. The distance from the inflow end of the column will be given by z. The time since the start of the experiment will be given by τ. At spatial point z and time τ, the solution concentration (in mass of contaminant per unit volume of water) is denoted $c(z, \tau)$, and the sorbed concentration (in mass of contaminant per unit mass of soil) is denoted $q(z, \tau)$.

We will assume the equilibrium relationship $q = f(c)$. The function f is called the equilibrium isotherm.

We will assume the following physical values: $\rho_V(z)$ is the void density per unit volume at spatial coordinate z, $\rho_S(z)$ is the soil mass density per unit volume, A is the cross-sectional area of the column, v is the fluid velocity, and D is the diffusivity constant that appears in Fick's law for diffusion. Fick's law assumes that the rate of diffusion of the contaminant in the fluid is given by $-D\partial c/\partial z$.

Choosing z to be any interior point of the column, τ any positive time, and Δz and $\Delta\tau$ small positive numbers, the change in the amount of contaminant stored in the section of the column $[z, z + \Delta z]$ between τ and $\tau + \Delta\tau$ is given by

$$\int_z^{z+\Delta z} [A\rho_V(\eta)c(\eta, \tau + \Delta\tau) + A\rho_S(\eta)q(\eta, \tau + \Delta t) - (A\rho_V(\eta)c(\eta, \tau) + A\rho_S(\eta)q(\eta, \tau))]d\eta.$$

This will equal the total inward flux less the outward flux at z and $z + \Delta z$ over the time from τ to $\tau + \Delta\tau$,

$$\int_\tau^{\tau+\Delta\tau} \left[vA\rho_V(z)\, c(z, s) - DA\rho_V(z)\, \frac{\partial c}{\partial z}(z, s) - \left(vA\rho_V(z + \Delta z)\, c(z + \Delta z, s) - DA\rho_V(z + \Delta z)\, \frac{\partial c}{\partial z}(z + \Delta z, s) \right) \right] ds.$$

Setting the two expressions equal, dividing by $\Delta z\Delta\tau$, and letting Δz and $\Delta\tau$ tend to zero yields

$$A\left[\rho_V(z)c(z, \tau) + \rho_S(z)q(z, \tau)\right]_\tau = -vA\left[\rho_V(z)c(z, \tau)\right]_z + DA\left[\rho_V(z)c_z(z, \tau)\right]_z.$$

If we assume that the mass density $\rho_S = M/(LA)$ and the pore volume density $\rho_V = V/(LA)$ are constant, where M is the total mass of soil in the cylinder and V

is the total void space in the cylinder, we obtain

$$[Vc + Mq]_\tau = -vVc_z + DVc_{zz}. \tag{2.1}$$

We have already described the expected equilibrium relationship between the sorbed concentration, q, and the solution concentration, which is given by $q = f(c)$. Since the solution concentration is changing in time we must either assume that this equilibrium relationship holds even as c changes in time or we must specify how q changes as c changes. For completeness we will describe the nonequilibrium modeling approach before we make our final assumptions. The standard model [16, 12] for sorption when c is known at each time t is

$$\frac{\partial q}{\partial \tau} = rF(c, q), \tag{2.2}$$

where r is a sorption rate constant and F satisfies the following requirements: If $q < f(c)$, then F is positive; if $q = f(c)$, then $F = 0$; and if $q > f(c)$, then F is negative. Some typical examples of the isotherm f are the Henry or linear isotherm

$$f(c) = \xi c,$$

the Langmuir isotherm

$$f(c) = \frac{\xi c}{1 + \beta c},$$

and the Freundlich isotherm

$$f(c) = \xi c^\gamma.$$

An example of F is a simple reversible sink

$$F(c, q) = f(c) - q,$$

where the rates of sorption and desorption are the same and where, regardless of whether the process is sorbing or disorbing, the same fixed c value will yield the same equilibrium point. Another example of F is a simple irreversible sink

$$(f(c) - q)^+,$$

where there is hysteresis occurring, and while a contaminant can be sorbed, it cannot be desorbed. There are a wide variety of such models, and the reader is referred to [16].

When the local kinetics (2.2) are combined with the conservation-of-mass equation (2.1) previously derived, we obtain

$$\begin{aligned} V\frac{\partial c}{\partial \tau} + M\frac{\partial q}{\partial \tau} &= -vV\frac{\partial c}{\partial z} + DV\frac{\partial^2 c}{\partial z^2}, \\ \frac{\partial q}{\partial \tau} &= rF(c, q). \end{aligned} \tag{2.3}$$

The equilibrium partitioning assumption is that r is much larger than D and v. Dividing the second system by r and defining $\varepsilon = 1/r$, we may consider this a singular perturbation problem:

$$\begin{aligned} V\frac{\partial c}{\partial \tau} + M\frac{\partial q}{\partial \tau} &= -vV\frac{\partial c}{\partial z} + DV\frac{\partial^2 c}{\partial z^2}, \\ \varepsilon\frac{\partial q}{\partial \tau} &= F(c, q). \end{aligned} \tag{2.4}$$

Since we will consider only cases where we are close to equilibrium, we need consider only the outer solution to the unperturbed problem

$$F(c,q)=0 \qquad \text{or} \qquad q=f(c).$$

Thus, we can replace system (2.4) with

$$\frac{\partial}{\partial\tau}\left(Vc+Mf(c)\right)=-vV\frac{\partial c}{\partial z}+DV\frac{\partial^2 c}{\partial z^2}.$$

In this paper we will accept the equilibrium partitioning assumption and use a Henry isotherm. The linear partitioning assumption is usually valid when concentrations are low. The common set of boundary conditions that we will be using is

$$c(0,\tau)=0, \qquad \frac{\partial c}{\partial z}(L,\tau)=0.$$

The first boundary condition is used to model the case where the inflow of water contains no contaminant, and the second boundary condition models the fact that there is no diffusion across the end of the column, only convection.

We will perform the standard change of variables [12] with respect to time and length by introducing new variables

$$t=v\tau/L \qquad \text{and} \qquad x=z/L.$$

Using these new variables, we obtain the form of the model we wish to study:

$$\begin{aligned}\frac{\partial}{\partial t}(\beta c)&=-\frac{\partial c}{\partial x}+K\frac{\partial^2 c}{\partial x^2},\\ c(0,t)&=0,\\ \frac{\partial c}{\partial x}(1,t)&=0,\\ c(x,0)&=c_0(x),\end{aligned} \tag{2.5}$$

where

$$\beta=1+\frac{\xi M}{V} \tag{2.6}$$

and

$$K=\frac{D}{vL}$$

is the nondimensionalized diffusion coefficient. The pseudotime variable measures pore volumes; that is, $t=1$ is the time it takes the flow to move from the top of the column to the bottom of the column.

In column studies, measurements of the exit solution concentration are taken. Therefore, the extra information available is a sequence of N time measurements taken at the end of the column $x=1$,

$$c(1,t_1),\ldots,c(1,t_N).$$

It is worthwhile to briefly discuss how the forward model came to be. The same model, with constant β, was successfully used to model column studies of fresh water

sediments. However, the constant β model failed when dealing with salt water sediments, and it was hypothesized by Myers [13] that β was changing with the saline concentration. Since it was expected that the saline concentration would equalize much more quickly than the concentration of the contaminant being studied, we chose to approximate the salt concentration as being spatially uniform and, thus, β as spatially uniform. In the technical report on this approach [14], we used a decreasing exponential ansatz for the saline concentration, assuming that the concentration would be dominated by the decay in the first eigenfunction. This yielded model fits that were considered reasonable by the engineers on the project. Henceforth, we assume that ξ, and hence β, is changing with time.

The initial contaminant concentration is taken to be spatially constant, $c(x,0) = c_0 > 0$, because the samples have time to equilibrate before the experiment begins. It is worth noting that this initial contaminant concentration, while physically accurate, does not meet the boundary condition. Indeed, in the explicit solutions generated for the technical report [14], there is a Gibb's phenomenon. Fortunately the problem is governed by a parabolic evolution operator, and solutions satisfy the boundary conditions for all positive times.

2.1. Contaminant mass constraint. We will now compute the value $\beta(0)$, which is related to the equilibrium coefficient ξ; see (2.6). We will assume that we know the total mass of soil in the column M, the total volume of water in the column V, the cross-sectional area of the column A, the diffusivity constant D, and the fluid velocity v. We will also assume that we know that the initial solution concentration is a constant, c_0. The mass flow of contaminant out of the tube at time t will be given by

$$Ac\,(1,t)\,V/A.$$

Thus, if we let the process continue until almost all of the contaminant has been flushed from the column at time t_N, we have that the total mass of contaminant present in the column at time $t = 0$ will be approximately

$$\int_0^{t_N} c\,(1,t)\,V\,dt. \tag{2.7}$$

We also know that the total mass of contaminant at time $t = 0$ will be given by

$$\beta\,(0)\,c_0 V. \tag{2.8}$$

Equating the expressions in (2.7) and (2.8), we obtain

$$\beta\,(0) \approx \frac{1}{c_0}\int_0^{t_N} c\,(1,t)\,dt. \tag{2.9}$$

We need to add a caveat at this point. The physical system allows contaminant to leave the cylinder only through the boundary at $x = 1$. However, examining the original system and integrating with respect to x yields

$$\begin{aligned}\frac{\partial}{\partial t}\int_0^1 (\beta c)\,dx &= c\,(0,t) - c\,(1,t) + K\frac{\partial c}{\partial x}\,(1,t) - K\frac{\partial c}{\partial x}\,(0,t)\\ &= -c\,(1,t) - K\frac{\partial c}{\partial x}\,(0,t)\,.\end{aligned}$$

Integrating the previous equation with respect to t from 0 to t_N and assuming that the $c\,(x, t_N) \approx 0$ yields

$$\beta\,(0)\,c_0 \approx \int_0^{t_N} c\,(1,t)\,dt + \int_0^{t_N} K\frac{\partial c}{\partial x}\,(0,t)\,dt.$$

Finally, solving for $\beta\,(0)$, we have

$$\beta\,(0) = \frac{1}{c_0}\int_0^{t_N} c\,(1,t)\,dt + \frac{1}{c_0}\int_0^{t_N} K\frac{\partial c}{\partial x}\,(0,t)\,dt.$$

Thus, the model allows for contaminant to leave the cylinder at $x = 0$, which will give some discrepancy. We require that $D \ll v$, which implies that $K \approx 0$, in order to minimize this error.

3. Identification of the sorption coefficient. The forward problem is

$$\begin{aligned} (\beta c)_t &= -c_x + Kc_{xx}, && 0 < x < 1,\ 0 < t < T = t_N, \\ c(0,t) &= 0, \qquad c_x(1,t) = 0, && 0 < t < T, \\ c\,(x,0) &= c_0\,(x)\,, && 0 < x < 1. \end{aligned} \tag{3.1}$$

Our goal is to estimate the parameter β from noisy measurements of $c(1,t)$. Although identification problems for parabolic equations have been addressed both theoretically [11, 2, 3, 5, 18] and numerically [4, 10], the general framework is to consider the problem

$$\begin{aligned} u_t &= L(a)[u] && \text{in } \Omega \times [0,T], \\ u(x,0) &= u_0(x) && \text{on } \Omega, \\ G(a)[u] &= 0 && \text{on } \partial\Omega \times [0,T], \end{aligned} \tag{3.2}$$

subject to additional information $B[u] = 0$ on $\Omega \times [0,T]$ or $\partial\Omega \times [0,T]$. The spatial operators L, G, B may be linear or quasi-linear. The unknown coefficient a may be part of L or G and may depend on x, t, or u. The goal is to recover a from information about u, the solution of (3.2). Although our forward problem (3.1) can be transformed into the form (3.2) by setting $u = \beta c$, the boundary data available from our experiment is $c(1,t)$. Thus boundary data for the transformed problem $u(1,t)$ would require knowledge of β, the parameter we seek. Similar issues arise with other transformation approaches. As in [3], trace-type functionals can then be used to establish existence of a solution. This approach can also be implemented numerically [10]. However, the dependence of the operators on nonlocal information can lead to numerical instabilities. We wish to develop an algorithm that uses our available data directly and avoids the use of nonlocal information. We begin by establishing identifiability of the parameter β from the available data $c(1,t)$. We apply output least squares with Tikhonov regularization to this problem. We investigate the rate of convergence and determine an appropriate source condition.

3.1. Identifiability. Recall the contaminant mass constraint (2.9),

$$\beta\,(0) \approx \frac{1}{c_0}\int_0^{t_N} c\,(1,t)\,dt,$$

discussed in section 2.1. As a consequence of this and the fact that our data is $c(1,t)$, it is reasonable to assume that $\beta(0)$ is fixed and to let

$$\mathcal{B} = \left\{\beta \in H^1(0,T) \mid 0 < m < \beta(t) < \mathcal{M},\ \beta(0) = b\right\}.$$

Our existence result follows from the application of standard results; see [20, 9].

THEOREM 3.1. *If $\beta \in \mathcal{B}$, then $c_\beta \in W = L^2((0,T);H^1(0,1))$ and $c_\beta(1,\cdot) \in L^2(0,T)$.*

In order to establish the identifiability of β we must establish the injectivity of the parameter-to-output map

$$\beta \to \gamma c_\beta,$$

where γ denotes the trace operator

$$\gamma : L^2((0,T);H^1(0,1)) \to L^2(0,T), \qquad \gamma c = c(1,t).$$

THEOREM 3.2. *Let $c_1(x,t)$ and $c_2(x,t) \in W$ be solutions of the direct problem (3.1) corresponding to $\beta_1(t)$ and $\beta_2(t) \in \mathcal{B}$. If $\gamma c_1 = \gamma c_2$, then $\beta_1(t) = \beta_2(t)$ for all $t \in [0,T]$.*

Proof. Use β_1, c_1 and β_2, c_2 in (3.1) and subtract to find

$$(\beta_1 c_1 - \beta_2 c_2)_t = -(c_1 - c_2)_x + K(c_1 - c_2)_{xx}.$$

Let $\phi = \beta_1 c_1 - \beta_2 c_2$ and rearrange terms

$$\phi_t = -\frac{1}{\beta_1}\phi_x + \frac{K}{\beta_1}\phi_{xx} + \frac{(\beta_2 - \beta_1)}{\beta_1}(\mathcal{A}c_2),$$

where $\mathcal{A}c = -c_x + Kc_{xx}$. Multiply by ϕ and integrate with respect to x:

$$\begin{aligned}\int_0^1 \phi_t\phi\,dx &= -\frac{1}{\beta_1}\int_0^1 \phi_x\phi\,dx + \frac{K}{\beta_1}\int_0^1 \phi_{xx}\phi\,dx + \frac{(\beta_2-\beta_1)}{\beta_1}\int_0^1 (\mathcal{A}c_2)\,\phi\,dx \\ &= -\frac{1}{2\beta_1}\int_0^1 \left(\phi^2\right)_x dx - \frac{K}{\beta_1}\int_0^1 (\phi_x)^2\,dx + \frac{(\beta_2-\beta_1)}{\beta_1}\int_0^1 (\mathcal{A}c_2)\,\phi\,dx.\end{aligned}$$

Since $0 < m < \beta_i(t) < \mathcal{M}$, $i = 1,2$, and $\phi(0,t) = 0$, it follows that

$$\frac{1}{2}\frac{d}{dt}\left(\|\phi\|^2_{L^2(0,1)}\right) \le \frac{2\mathcal{M}}{m}\|\mathcal{A}c_2\|_{L^2(0,1)}\|\phi\|_{L^2(0,1)}.$$

By results in [14] or [17], there exists a constant C_1 such that

$$\|\mathcal{A}c_2\|_{L^2(0,1)} \le \frac{C_1\beta_2(0)}{t}\|c_0\|_{L^2(0,1)}.$$

Hence

$$\frac{d}{dt}\left(\|\phi\|_{L^2(0,1)}\right) \le \frac{2\mathcal{M}C_1\beta_2(0)}{mt}\|c_0\|_{L^2(0,1)}.$$

Integrating this over (t_1,t_2) yields

$$\begin{aligned}\|\phi\|_{L^2(0,1)}(t_2) &\le \|\phi\|_{L^2(0,1)}(t_1) + \left(2\mathcal{M}C_1\beta_2(0)\|c_0\|_{L^2(0,1)}/m\right)\ln|t_2/t_1| \\ &\le \|\phi\|_{L^2(0,1)}(t_1) + \left(2\mathcal{M}C_1\beta_2(0)\|c_0\|_{L^2(0,1)}/m\right)\ln|1+\varepsilon|,\end{aligned}$$

where $\varepsilon = (t_2 - t_1)/t_1$. Since t_2 can be chosen to be arbitrarily close to t_1, and since $\ln(1+z) < z$ for all $z > 0$, it follows that

$$\|\phi\|_{L^2(0,1)}(t_2) \le \|\phi\|_{L^2(0,1)}(t_1) + \left(2\mathcal{M}C_1\beta_2(0)\|c_0\|_{L^2(0,1)}/m\right)\varepsilon$$

for all $\varepsilon > 0$. Therefore $\|\phi\|_{L^2(0,1)}(t_2) \leq \|\phi\|_{L^2(0,1)}(t_1)$ or

$$\|\beta_1 c_1 - \beta_2 c_2\|_{L^2(0,1)}(t_2) \leq \|\beta_1 c_1 - \beta_2 c_2\|_{L^2(0,1)}(t_1).$$

Letting t_1 approach 0, we have

$$\|\beta_1 c_1 - \beta_2 c_2\|_{L^2(0,1)}(t) \leq \|\beta_1 c_1 - \beta_2 c_2\|_{L^2(0,1)}(0) = |\beta_1(0) - \beta_2(0)| \, \|c_0\|_{L^2(0,1)} = 0$$

for small t. This implies that $\beta_1(t)c_1(x,t) - \beta_2(t)c_2(x,t) = 0$ almost everywhere on $[0,1]$. It follows from $c_1(1,t) = c_2(1,t)$ that $\beta_1(t) = \beta_2(t)$ for small t. Repeated application of this argument extends the result to $[0,T]$. ∎

3.2. Output least squares and Tikhonov regularization. We define

$$G(\beta) \equiv \gamma c_\beta$$

with

$$G : \mathcal{B} \to L^2(0,T).$$

In the presence of perfect data z, we would solve the nonlinear ill-posed problem

$$G(\beta^0) = z, \tag{3.3}$$

where c_{β^0} is the solution of the direct problem with $\beta = \beta^0$. To do this using Tikhonov regularization would involve approximating the solution by minimizing

$$\min_{\beta \in \mathcal{B}} \|G(\beta) - z\|^2_{L^2(0,T)} + \alpha \|\beta - \hat{\beta}\|^2_{L^2(0,T)},$$

where $\alpha > 0$ is a small parameter and $\hat{\beta}$ is an a priori guess of the true solution β^0. In real applications, measurement errors mean that exact data is not available. Noisy data is assumed to have an error level δ,

$$\left\|z^\delta - z\right\|_{L^2(0,T)} \leq \delta.$$

We assume attainability of a true solution; i.e., if $z \in L^2(0,T)$, there exists $\beta^0 \in \mathcal{B}$ such that

$$G(\beta^0) = z. \tag{3.4}$$

We seek the minimizer $\beta^\delta_\alpha \in \mathcal{B}$ of

$$J_\alpha(\beta) = \left\|G(\beta) - z^\delta\right\|^2_{L^2(0,T)} + \alpha \|\beta - \hat{\beta}\|^2_{L^2(0,T)} \tag{3.5}$$

for appropriate choices of $\hat{\beta} \in \mathcal{B}$ and α. We begin by establishing the weak-closedness of the map $\beta \to \gamma c_\beta$. This will lead to the existence of a minimizer β^δ_α. Continuous dependence on the data z^δ for fixed α, and the convergence of β^δ_α toward the true parameter β^0 as the noise level δ and the regularization parameter α go to zero, also follow.

THEOREM 3.3. *If $\beta_n \rightharpoonup \beta_* \in \mathcal{B}$ in $H^1(0,T)$, then $c_{\beta_n} \rightharpoonup c_{\beta_*}$ in W and $\gamma c_{\beta^n} \rightharpoonup \gamma c^*_\beta$ in $L^2(0,T)$.*

Proof. Since we have existence of a unique solution to (3.1) from Theorem 3.1, we define $c_n = c(\beta_n)$. We make a change of variables, $w = e^{-\lambda t}c$, where λ is to be chosen. The state equation and initial and boundary conditions become

$$\begin{aligned}(\beta w)_t + \lambda\beta w - Kw_{xx} + w_x &= 0,\\ w(0,t) = 0, w_x(1,t) = 0, \quad t &\in [0,T],\\ w(x,0) &= c_0(x).\end{aligned}$$

Let $\langle\cdot,\cdot\rangle$ denote the duality between $(H^1(0,1))^*$ and $H^1(0,1)$. We use the weak definition of the transformed equation and integrate in time to obtain

(3.6)
$$0 = \int_0^t \langle(\beta_n w_n)_t, \beta_n w_n\rangle dt + \int_0^t\int_0^1 \left[\lambda(\beta_n w_n)^2 + K\beta_n(w_n)_x^2 + \beta_n(w_n)_x w_n\right] dxdt.$$

Upon simplification, use of $0 < m < \beta_n(t) < \mathcal{M}$, and Cauchy's inequality, we have

$$\begin{aligned}&\frac{m^2}{2}\int_0^1 [w_n(x,t)]^2\,dx + \lambda m^2\int_0^t\int_0^1 (w_n)^2 dxdt + Km\int_0^t\int_0^1 (w_n)_x^2 dxdt\\ &\le \frac{\mathcal{M}^2}{2}\int_0^1 [c_0(x)]^2 dx + \frac{\mathcal{M}^2}{2Km}\int_0^t\int_0^1 (w_n)^2 dxdt + \frac{Km}{2}\int_0^t\int_0^1 (w_n)_x^2 dxdt.\end{aligned}$$

After dividing by m^2, collecting terms, and choosing $\lambda > \frac{(2Km^3+1)\mathcal{M}^2}{2Km^3}$, we obtain

$$\begin{aligned}&\int_0^1 [w_n(x,t)]^2 dxdt + \frac{K}{m}\int_0^t\int_0^1 (w_n)_x^2 dxdt + 2\mathcal{M}^2\int_0^t\int_0^1 (w_n)^2 dxdt\\ &\le \frac{\mathcal{M}^2}{m^2}\int_0^1 [c_0(x)]^2 dx.\end{aligned} \tag{3.7}$$

We can conclude that $\|w_n\|_{L^2((0,T),H^1(0,1))}$, and hence $\|c_n\|$ is uniformly bounded independent of n. Using this bound and the state equation, we also have uniform bounds on $\|(\beta_n c_n)_t\|$. We can extract a subsequence such that

$$\begin{aligned}c_n &\rightharpoonup c_* \quad \text{in} \quad L^2((0,T);H^1(0,1)),\\ (c_n)_t &\rightharpoonup (c_*)_t \quad \text{in} \quad L^2((0,T);(H^1(0,1))^*),\\ \beta_n &\to \beta_* \quad \text{in} \quad L^2(0,T),\end{aligned}$$

where c_* and β_* are the relevant weak limits. In order to show that $c^* = c(\beta_*)$, we must establish that c_* is the state solution associated with β_*. We consider the weak form of the partial differential equation satisfied by c_n,

$$\int_0^T \langle(\beta^n c^n)_t, \phi\rangle dt + K\int_0^T\int_0^1 c_x^n\phi_x dxdt + \int_0^T\int_0^1 c_x^n\phi dxdt = 0, \tag{3.8}$$

where $\phi \in L^2((0,T);H^1(0,1))$.

Since $\beta_n \to \beta_*$ in $L^2(0,T)$ we know that $\beta_n \rightharpoonup \beta_*$ in $H^1(0,T)$. Therefore $(\beta_n)' \rightharpoonup (\beta_*)'$ in $L^2(0,T)$, where $'$ is used to indicate the derivative here because β is a function of one variable, t. We examine the first term in the weak definition of state solution,

$$\begin{aligned}&\int_0^T\int_0^1 \left[(\beta_n c_n)_t - (\beta_* c_*)_t\right]\phi dxdt\\ &= \int_0^T\int_0^1 \left[(\beta_n)'\left[c_n - c_*\right]\phi + (\beta_n - \beta_*)' c_*\phi\right.\\ &\qquad \left. + \beta_n((c_n)_t - (c_*)_t)\phi + (\beta_n - \beta_*)(c_*)_t\phi\right] dxdt.\end{aligned} \tag{3.9}$$

We note that, from a comparison result in [20], we have that $c_n \to c_*$ in $L^2((0,1)\times(0,T))$. By $(\beta_n)' \rightharpoonup (\beta_*)'$ in $L^2(0,T)$ and $c_n \to c_*$ in $L^2((0,1)\times(0,T))$, the first term of (3.9) converges to zero as $n \to \infty$. The second term converges to zero since $(\beta_n)' \rightharpoonup (\beta_*)'$ in $L^2(0,T)$. The third and fourth terms converge to zero because of the strong convergence of the β_n sequence. As we pass to the limit in the weak definition of the solution, we obtain that $c_* = c(\beta_*)$. ∎

Existence of a minimizer β_α^δ now follows from the lower semicontinuity of the $L^2(0,T)$ norm. Continuous dependence on the data z^δ for fixed α and the convergence of β_α^δ toward the true parameter β^0 follow from standard results [19].

COROLLARY 3.4. *For any data $z^\delta \in L^2(0,T)$, a minimizer β_α^δ of (3.5) exists.*

COROLLARY 3.5. *For fixed α, the minimizers depend continuously on the data z^δ. If $\alpha(\delta)$ satisfies*

$$\alpha(\delta) \to 0, \qquad \delta^2/\alpha(\delta) \to 0 \qquad \text{as } \delta \to 0,$$

then

$$\lim_{\delta\to 0} \left\|\beta_\alpha^\delta - \beta^0\right\|_{L^2(0,T)} = 0.$$

3.3. Convergence rates. Although we have established convergence of the minimizer β_α^δ to the true parameter β^0, the rate of convergence may be arbitrarily slow. We apply the theory of Engl, Hanke, and Neubauer [7] and Engl, Kunisch, and Neubauer [8] to determine a source condition that will guarantee a certain rate of convergence. Recall that we seek to solve the nonlinear problem (3.3), $G(\beta) = z$, where $G(\beta) \equiv \gamma c_\beta$. The true solution is β^0, and $\hat\beta$ is an a priori guess. Let $L(\beta)$ be the differential operator

$$L(\beta)u \equiv (\beta u)_t + u_x - K u_{xx}$$

on the domain $D(L) = \{u \in W \,|\, u(0,t) = u_x(1,t) = 0\}$. The soil problem (3.1) satisfies

$$L(\beta)c = 0, \qquad c(x,0) = c_0(x).$$

We establish next an estimate of the rate of convergence of our algorithm. Even when our regularization parameter α is comparable to our noise level δ, convergence requires assumptions involving $c(1,t)$ and $\beta^0 - \hat\beta$.

THEOREM 3.6. *Let $By \equiv -y'' + y$ with*

$$B : D(B) = \left\{y \in H^2(0,T) \mid y'(0) = y'(T) = 0\right\} \to L^2(0,T).$$

If

$$\frac{B\left(\beta^0(t) - \hat\beta(t)\right)}{c(1,t)} \in H^{-1}(0,T) \tag{3.10}$$

and if

$$\left\| \int_t^T \frac{B\left(\beta^0(t) - \hat\beta(t')\right)}{c(1,t')}\, dt' \right\|_{L^2(0,T)} \quad \text{is sufficiently small,} \tag{3.11}$$

then for the choice $\alpha \sim \delta$ *we obtain*

$$\left\|\beta_\alpha^\delta - \beta^0\right\|_{H^1(0,T)} = O(\sqrt{\delta}) \qquad \textit{and} \qquad \left\|G(\beta_\alpha^\delta) - z^\delta\right\|_{L^2(0,T)} = O(\delta).$$

Proof. We define

$$F(\beta) \equiv c(x,t).$$

If $\Psi_h(x,t) = F'(\beta)h(t)$ is the first Fréchet derivative in the direction h, then $\Psi_h \in D(L)$ and

$$L(\beta)\Psi_h = -(hc)_t, \qquad \Psi_h(x,0) = 0.$$

Similarly, if $\Phi_h(x,t) = F''(\beta)\left(h(t),h(t)\right)$ is the second Fréchet derivative in the direction h, then $\Phi_h \in D(L)$ and

$$L(\beta)\Phi_h = -2(h\Psi_h)_t, \qquad \Phi_h(x,0) = 0.$$

By continuity of the trace operator, we have

$$G'(\beta)h(t) = \Psi_h(1,t) = \gamma\Psi_h, \qquad G''(\beta)\left(h(t),h(t)\right) = \Phi_h(1,t) = \gamma\Phi_h.$$

Thus, G is twice Fréchet differentiable.

Define $p \in W$ to be the solution of

$$L(\beta)^* p = h, \qquad p(x,T) = 0,$$

where the adjoint L^* is taken with respect to $L^2(0,T)$. Notice that

$$\begin{aligned}
\langle g, G'(\beta)^* h\rangle_{H^1(0,T)} &= \langle G'(\beta)g, h\rangle_{L^2(0,T)} \\
&= \langle \gamma\Psi_g, h\rangle_{L^2(0,T)} = \gamma\langle \Psi_g, h\rangle_{L^2(0,T)} \\
&= \gamma\langle -L^{-1}(gc)_t, h\rangle_{L^2(0,T)} = -\gamma\langle (gc)_t, \left(L^{-1}\right)^* h\rangle_{L^2(0,T)} \\
&= -\gamma\langle (gc)_t, \left(L^*\right)^{-1} h\rangle_{L^2(0,T)} = -\gamma\langle (gc)_t, p\rangle_{L^2(0,T)} \\
&= \gamma\langle g, cp_t\rangle_{L^2(0,T)},
\end{aligned}$$

provided that $g(0) = 0$. Since

$$\langle g, v\rangle_{L^2(0,T)} = \left\langle \left(B^{-1}\right)^* g, v\right\rangle_{H^1(0,T)}$$

for every $v \in H^1(0,T)$, it follows that

$$\gamma\left\langle g, (cp_t)\right\rangle_{L^2(0,T)} = \gamma\left\langle \left(B^{-1}\right)^* g, cp_t\right\rangle_{H^1(0,T)} = \gamma\left\langle g, B^{-1}\left(cp_t\right)\right\rangle_{H^1(0,T)}.$$

This means that

$$G'(\beta)^* h = \gamma B^{-1}\left(cp_t\right)$$

with $L(\beta)^* p = h$, $p(x,T) = 0$. Since (3.10) holds, there exists $w \in L^2(0,T)$ such that

$$\beta^0 - \hat{\beta} = G'(\beta_0)^* w \tag{3.12}$$

or $B(\beta^0 - \hat{\beta}) = c(1,t)p_t(1,t)$ with

$$L(\beta^0)c = 0, c(x,0) = c_0(x), \qquad L(\beta^0)^* p = w(t), p(x,T) = 0.$$

Let $h := \beta_\alpha^\delta - \beta^0$ and $\beta^s := \beta^0 + sh$. Then

$$\begin{aligned}
&\left| 2\left\langle w, \int_0^1 G''(\beta^s)(h,h)(1-s)\,ds \right\rangle_{L^2(0,T)} \right| \\
&\quad = \left| 2\int_0^T w(t) \int_0^1 G''(\beta^s(t))(h(t),h(t))(1-s)\,ds\,dt \right| \\
&\quad \le \sup_{0\le s\le 1} \left| \langle w, G''(\beta^s)(h,h)\rangle_{L^2(0,T)} \right| \\
&\quad = \sup_{0\le s\le 1} \left| \left\langle \gamma L(\beta^0)^* p,\ , G''(\beta^s)(h,h)\right\rangle_{L^2(0,T)} \right| \\
&\quad = \|\gamma p\|_{L^2(0,T)} \left\| \gamma L(\beta^0)\Psi_h \right\|_{L^2(0,T)}.
\end{aligned}$$

As in [8], it can be shown that there exist constants C_1 and C_2 such that

$$\|\Phi_h\|_W \le C_1 \|h\|_{L^2(0,T)} \|\Psi_h\|_W, \qquad \|\Psi_h\|_W \le C_2 \|h\|_{L^2(0,T)} \|c\|_W.$$

By the trace theorem [9, p. 258] and boundedness of the operator $L(\beta^0)$, there exists C_3 such that

$$\left\| \gamma L(\beta^0)\Psi_h \right\|_{L^2(0,T)} \le C_3 \|h\|^2_{H^1(0,T)} \|c\|_W.$$

As the quantity appearing in (3.11) is $\|\gamma p\|_{L^2(0,T)}$, we deduce that

$$2\left\langle w, \int_0^1 G''(\beta^s)(h,h)(1-s)\,ds \right\rangle_{L^2(0,T)} \le \rho \|h\|^2_{H^1} \tag{3.13}$$

with $\rho < 1$. Since G is twice Fréchet differentiable and both (3.12) and (3.13) hold, application of the theory of Engl, Kunisch, and Neubauer [8, Theorem 2.4] yields the desired convergence result. ∎

Discussion of source condition. The condition (3.10) requires that the difference between the a priori guess $\hat{\beta}$ and the true solution β^0 must be in $D(B) \subset H^2(0,T)$. In practical applications, this regularity assumption is very restrictive.

It was established in Oppenheimer [14] for Hölder continuous β that there exist constants $C \ge 0$ and $\theta > 0$ such that

$$\|c(\cdot,t)\|_{L^2(0,1)} \le C\frac{\beta(0)}{\beta(t)} e^{-\theta t} \|c_0\|_{L^2(0,1)}.$$

Since $\beta \in \mathcal{B}$ is bounded, there exist constants $C \ge 0$ and $\theta > 0$ such that

$$\|c(\cdot,t)\|_{L^2(0,1)} \le Ce^{-\theta t} \|c_0\|_{L^2(0,1)}.$$

The requirement (3.11) means that the difference between our a priori guess $\hat{\beta}$ and the true parameter β^0 must be small and very smooth when our measurement $c(1,t)$ is small. This is both a local and global restriction. Since $c(\cdot,t)$ decays exponentially, this is possible only for sufficiently small T.

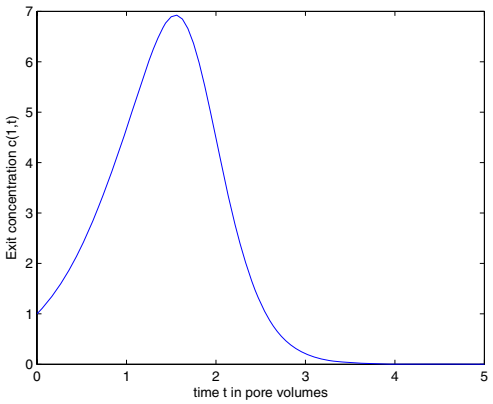


FIG. 4.1. *Exit concentration.*

4. Numerical results. In order to demonstrate the effectiveness of Tikhonov regularization for this application, we consider two examples. Recall that the solution of our forward problem (3.1) decays over time, i.e., $c(x,t) \to 0$ as $t \to \infty$. As a result, we do not expect to be able to use all of the available data or to recover $\beta(t)$ over the whole time interval.

All computations were carried out in MATLAB. The Tikhonov functional

$$J_\alpha(\beta) = \int_0^T \left(c(1,t) - z^\delta\right)^2 dt + \alpha \int_0^T \left(\beta(t) - \hat{\beta}(t)\right)^2 dt$$

was minimized using a Gauss–Newton method. Here z^δ and $\hat{\beta}$ represent noisy data and an a priori guess of the parameter. During the computation of $J_\alpha(\beta)$, exit concentrations $c(1,t)$ associated with a particular β were computed using an implicit finite-difference algorithm. The integrals were computed using a trapezoidal rule. Exit concentration data was generated using a method of lines algorithm with high accuracy. The a priori guess was chosen to be $\hat{\beta} = \beta(0)$, which was estimated using the approximation $\beta(0) \approx (\int_0^{t_N} c(1,t)\,dt)/c_0$. Strategies for the discussion of regularization parameters are discussed in [21]. For the purposes of this discussion, we will choose the regularization parameter to be $\alpha = 0$ in the absence of noise.

Example 1. Let $c_0(x) = 1$ and $K = 0.07$. Consider the sorption coefficient

$$\beta(t) = 1 + 10e^{-2t}.$$

The exit concentration $c(1,t)$ associated with this β is shown in Figure 4.1. Since the data decays over time, we restrict the recovery of β to the time interval $[0, 2.5]$. An initial guess of $\beta = 1$ was used. The parameter β and its recovery $\beta_{nonoise}$ from noiseless data are shown in Figure 4.2. Notice that quality of the recovery degrades after $t = 2$. This is due to the fact that the exit concentrations become very small and begin to amplify numerical error in the algorithm.

Example 2. Let $c_0(x) = 1$ and $K = 0.07$. Consider the sorption coefficient

$$\beta(t) = 2 + \cos(10t).$$

The exit concentration $c(1,t)$ associated with this β is shown in Figure 4.3. Once again the data decays over time, and we restrict the recovery of β to the time interval $[0, 2]$.

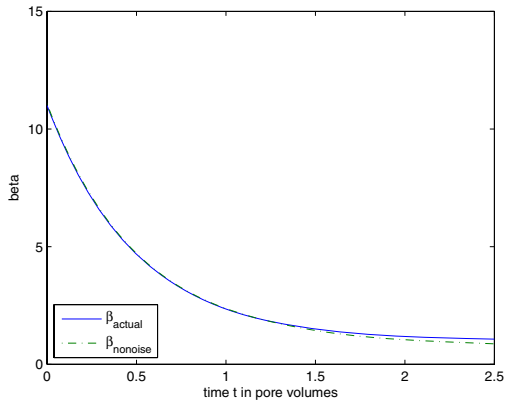


FIG. 4.2. *$\beta_{nonoise}$ recovered from exact data.*

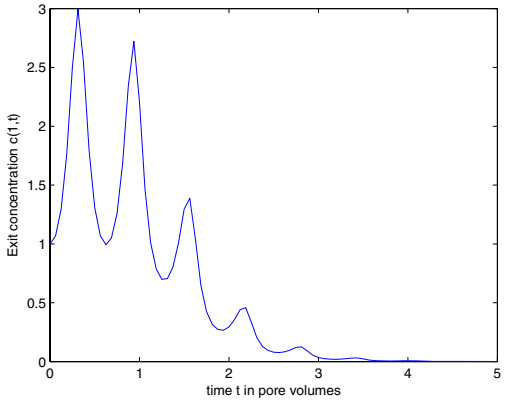


FIG. 4.3. *Exit concentration.*

An initial guess of $\beta = 2$ was used. The parameter β and its recovery $\beta_{nonoise}$ from noiseless data are shown in Figure 4.4. Notice that quality of the recovery degrades after $t = 1.25$.

Example 3*: Noisy data.* Since data for this problem is measured experimentally, it will contain a certain amount of noise. The nature of these experiments suggests that the noise level may be as much as 20%. Noise is introduced into the data from Example 1 via a normally distributed random number generator. Figure 4.5 shows the noisy exit concentration data.

In engineering applications, it is not practical to expect to have sufficient information about the unknown parameter β^0 in order to choose an a priori guess $\hat{\beta}$ so that $\beta^0 - \hat{\beta} \in D(B) \subset H^2(0,T)$. Hence Theorem 3.6 does not apply. Instead, a regularization parameter of $\alpha = 10^{-3}$ was chosen heuristically by an L-curve method, [21]. Figure 4.6 shows the recovery of β_{noise} with and without regularization. Clearly $\alpha = 10^{-3}$ produces better results than $\alpha = 0$, although the recovery is not as good as the noiseless case in Example 1.

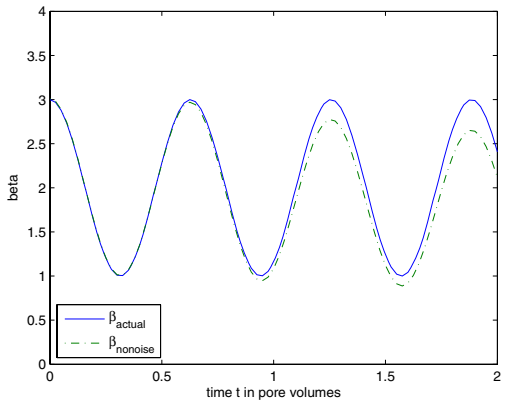


FIG. 4.4. *$\beta_{nonoise}$ recovered from exact data.*

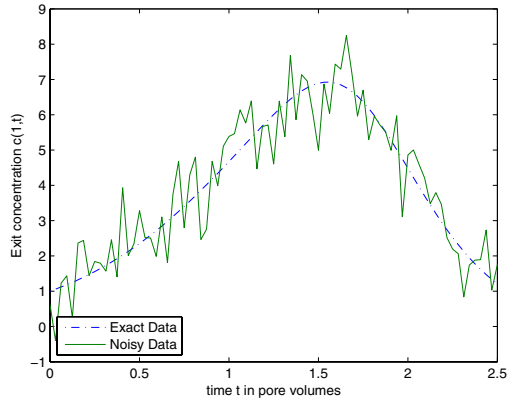


FIG. 4.5. *Exit concentration data with* 20% *noise.*

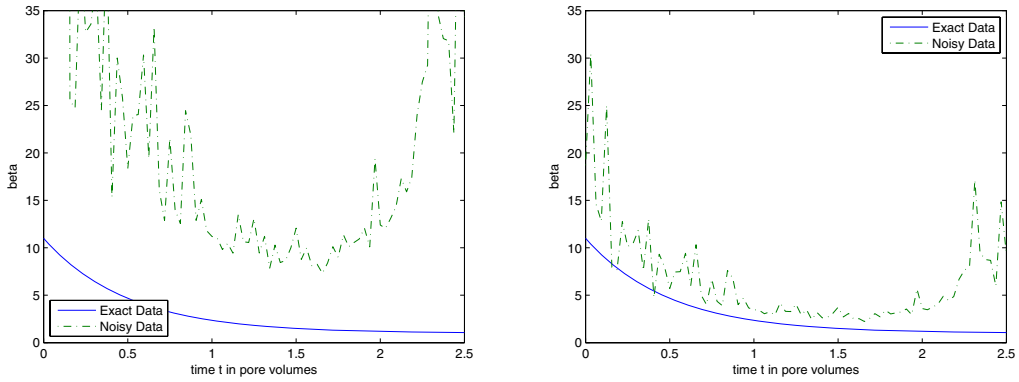


FIG. 4.6. *β_{noise} recovered from noisy data without regularization and with a regularization parameter of $\alpha = 10^{-3}$.*

5. Conclusions. In this paper, we have identified the sorption partitioning coefficient as a function of time from limited boundary data. A numerical approach for approximating this time dependent parameter in a parabolic partial differential equation has been analyzed. This work has brought insight into how a partitioning coefficient varies with different physical factors such as temperature fluctuations and contaminant introduction in the soil column. In the column studies, the boundary data is represented by the measurements of contaminant concentrations as the solution exits the soil column. The identifiability of the soil sorption parameter, β, is determined from these noisy exit concentration measurements. In order to establish the identifiability of the parameter β, we proved the injectivity of the parameter to the output map. We then discussed an output least squares formulation of the problem with Tikhonov regularization. Using this format, we found a minimizer to our approximate problem and were able to prove that this minimizer converges to the true parameter as the noise level and the regularization parameter approach zero. Although we proved convergence, the rate of convergence may be arbitrarily slow. Therefore, we established a source condition that guarantees a given rate of convergence. However, there is a trade-off here. The condition requires that the difference between our a priori guess and the true parameter must be small and relatively smooth when the boundary measurements are small. We found that this in possible only over a small time interval because the contaminant concentration decays over time. Within the numerical examples, this is seen after $t = 2$ in Example 1 and after $t = 1.6$ in Example 2. However, with the noisy data, the quality of the identification significantly improves with the inclusion of a regularization parameter. Consequently, the implementation of Tikhonov regularization provides a more tractable result.

REFERENCES

[1] D. D. Adrian, S. Ozkan, and A. N. Alshawabkeh, *Tracer transport in a soil column for sine wave loading*, in Physical and Chemical Processes of Water and Solute Transport/Retention in Soil, H. M. Selim and D. L. Sparks, eds., Special Publication 56, Soil Science Society of America, Madison, WI, 2001, pp. 169–188.

[2] J. R. Cannon, *Determination of an unknown coefficient in a parabolic differential equation*, Duke Math J., 30 (1963), pp. 313–324.

[3] J. R. Cannon, P. DuChateau, and K. Steube, *Unknown ingredient inverse problems and trace-type functional differential equations*, in Inverse Problems in Partial Differential Equations, D. Colton, R. Ewing, and W. Rundell, eds., SIAM, Philadelphia, 1990, pp. 187–202.

[4] J. R. Cannon, Y. Lin, and S. Xu, *Numerical procedures for the determination of an unknown coefficient in semi-linear parabolic differential equations*, Inverse Problems, 10 (1994), pp. 227–243.

[5] J. R. Cannon and W. Rundell, *Recovering a time-dependent coefficient in a parabolic differential equation*, J. Math. Anal. Appl., 160 (1991), pp. 572–582.

[6] P. A. Domenico and F. W. Schwartz, *Physical and Chemical Hydrogeology*, 2nd ed., John Wiley & Sons, New York, 1998.

[7] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Norwell, MA, 2000.

[8] H. W. Engl, K. Kunisch, and A. Neubauer, *Convergence rates for Tikhonov regularisation of non-linear ill-posed problems*, Inverse Problems, 5 (1989), pp. 523–540.

[9] L. C. Evans, *Partial Differential Equations*, American Mathematical Society, Providence, RI, 1998.

[10] A. G. Fatullayev, *Numerical procedure for the simultaneous determination of unknown coefficients in a parabolic equation*, Appl. Math. Comput., 164 (2005), pp. 697–705.

[11] V. Isakov, *Inverse Problems for Partial Differential Equations*, Springer, New York, 1998.

[12] F. J. Leij and J. H. Dane, *A Review of Physical and Chemical Processes Pertaining to Solute Transport*, Agronomy and Soils Departmental Series 132, Alabama Agricultural Experiment Station, Auburn University, Auburn, AL, 1989.

[13] T. E. Myers, *Metals Release from Freshwater and Estuarine Sediments in Thin-Disk Leaching Columns*, Ph.D. thesis, Civil and Environmental Engineering Department, Louisiana State University and Agricultural and Mechanical College, Baton Rouge, LA, 1999.

[14] S. F. Oppenheimer, *Parameter Identification for an Advection-Dispersion Equation with Salinity Controlled Partitioning*, Technical report DACW 39-92-M-6776, U.S. Army Engineers, Waterways Experiment Station, Vicksburg, MS, 1992.

[15] S. F. Oppenheimer, *The sorption of mixtures under linear equilibrium partitioning and chemical transformation*, Math. Methods Appl. Sci., 18 (1995), pp. 803–823.

[16] S. F. Oppenheimer, W. L. Kingery, and F. X. Han, *Phase plane analysis and dynamical systems approaches to the study of metal sorption in soils*, in Heavy Metals Release in Soils, H. M. Selim and D. L. Sparks, eds., Lewis Publishers, New York, 2001, pp. 109–130.

[17] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1983.

[18] M. S. Pilant and W. Rundell, *Undetermined coefficient problems for quasilinear parabolic equations*, in Inverse Problems in Partial Differential Equations, D. Colton, R. Ewing, and W. Rundell, eds., SIAM, Philadelphia, 1990, pp. 165–185.

[19] T. I. Seidman and C. R. Vogel, *Well-posedness and convergence of some regularization methods for non-linear ill-posed problems*, Inverse Problems, 5 (1989), pp. 227–238.

[20] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.

[21] C. R. Vogel, *Computational Methods for Inverse Problems*, Frontiers in Appl. Math. 23, SIAM, Philadelphia, 2002.

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1424–1453

INTRODUCING A POPULATION INTO A STEADY COMMUNITY: THE CRITICAL CASE, THE CENTER MANIFOLD, AND THE DIRECTION OF BIFURCATION*

BARBARA BOLDIN†

Abstract. In this paper we study deterministic, finite dimensional, continuous, as well as discrete time population invasion models. The ability of a newly introduced population, either a new species or a reproductively isolated subpopulation of one of the already present species, to settle in the community relies upon the basic reproduction ratio of the invader, $\mathcal{R}_0$. When $\mathcal{R}_0$ exceeds 1, the invading population meets with success, and when $\mathcal{R}_0$ is below 1, the invasion fails. The aim of this paper is to investigate the possible effects of an invasion when the parameters of a model are varied so that $\mathcal{R}_0$ of the invading population passes the value 1. We argue that population invasion models, regardless of the biology that underlies them, take a specific form that significantly simplifies the center manifold analysis. We make a uniform study of ecological, adaptive dynamics and disease transmission models and derive a simple formula for the direction of bifurcation from a steady state in which only the resident populations are present. Furthermore, we observe that among those bifurcation parameters that satisfy a certain condition, we acquire the same direction of bifurcation. The obtained mathematical results are used to gain insight into the biology of invasions. The theory is illustrated by several examples.

Key words. population model, physiologically structured population, *i*-state, *p*-state, reproductively isolated, population, species, center manifold, transcritical bifurcation, direction of bifurcation, basic reproduction ratio, finitely many states at birth, next generation matrix, invasibility

AMS subject classifications. 37C10, 37G10, 92D25, 92D30, 92D40

DOI. 10.1137/050629082

1. Introduction.

One of the basic questions of population biology is the following. Suppose that a population is introduced into a steady community in which it has not been present before. Under what conditions will this newly introduced population be able to settle in the community, and when will the dynamics lead to its extinction?

The literature that deals with this question is vast and diverse. We could roughly group the biological settings into the following categories:

1. In *ecology* one is studying an introduction of (i) a population of predators that forages on a resident prey community (i.e., the so-called predator-prey models; see [7], [11], [20], [26], [27], [30], [31], [32], [34] for some examples), or (ii) a population that competes for resources with the resident community (for examples of competition models, see [7], [20], [26], [31], [33]).

While it is common that the newly introduced population is also a new species, i.e., one that is not present in the resident community, there are also examples in which one is interested in the ability of what we shall call a reproductively isolated subpopulation of one species to be able to settle among individuals of another subpopulation of the same species. We shall define the precise meaning of the term *reproductive isolation* in Appendix B. The reader may at this point have in mind, for example, studying interactions (say, competition for shared resources) among different year classes of

*Received by the editors April 12, 2005; accepted for publication (in revised form) January 12, 2006; published electronically May 12, 2006.
http://www.siam.org/journals/siap/66-4/62908.html
†Department of Mathematics, University of Utrecht, P. O. Box 80010, 3508 TA Utrecht, The Netherlands (boldin@math.uu.nl).

semelparous species [2], [8], [9], [10] or different morphs in size-structured populations [3], [4].

2. In a branch of the theory of evolution called *adaptive dynamics* one is investigating the ability of a rare mutant phenotype to invade the environment set by the resident community (see [12], [21], [25] and the references therein).

3. *Epidemiology of infectious diseases* is concerned with introductions of infectious pathogens into susceptible populations (see, for example, [13], [18], [19], [22], [34], [35]).

When the resident community is at a stable equilibrium and we describe the process of invasion by a deterministic model we can, regardless of the biological background, answer the invasibility question in terms of the *basic reproduction ratio* [13], [15] of the invading population, $\mathcal{R}_0$, as follows: if $\mathcal{R}_0 < 1$, the invading population will go extinct, and if $\mathcal{R}_0 > 1$, it will settle in the community. The transition hence occurs when $\mathcal{R}_0 = 1$.

Now, what happens when $\mathcal{R}_0$ passes the value 1? The answer can be formulated mathematically or biologically. In mathematical terms one says that a *transcritical bifurcation* of a steady state and an *exchange of stability* take place [5], [6], [24], [36]. From a strictly mathematical point of view there is but one generic type of transcritical bifurcation. But when it comes to seeing the results from a biologist's point of view one must realize that a steady state is meaningful only when all its components are nonnegative, in particular those corresponding to the invading population. In many models the latter requirement is fulfilled only when $\mathcal{R}_0 > 1$. The bifurcation is then called *supercritical* or *forward* or, also, *soft* or *smooth*, since the size of the invading population remains small when $\mathcal{R}_0 - 1$ is positive but small. In some models, however, the positivity requirement is fulfilled only for $\mathcal{R}_0 < 1$, and one then speaks of a *subcritical* or *backward* bifurcation.

While in the case of a supercritical bifurcation the invasion fails when $\mathcal{R}_0$ of the invading population falls below 1, the invader can meet with success even if $\mathcal{R}_0 < 1$ (when introduced in sufficiently large quantities) when the bifurcation is backward. Moreover, even when the invader is introduced in small quantities, a small perturbation of $\mathcal{R}_0$ to a value greater than 1 can in a subcritical case lead to a rather large invader population size. This phenomenon is sometimes called *catastrophic transition* (see Figure 2.2).

Clearly then, it is important to be able to tell which of the two cases applies in any given situation, and in this paper we provide the reader with a simple criterion to distinguish between the two scenarios.

Of course, the invasibility question is equally meaningful when the resident community resides in a dynamic attractor. This situation is, however, outside the scope of this paper.

Throughout this paper we consider communities whose members differ in a finite number of characteristics. These characteristics are in the context of population models often called *i-states* [13], [15], with i standing for individual. Ideally, they should capture precisely the features that are relevant for the description of the process one is studying, and are hence to be considered for each problem separately.

In a general setting we assume that the community is divided into $m + n$ subpopulations, of which m subpopulations constitute the invading population and the remaining n make up the resident community. We denote by

$$\mathcal{Y} = \{(y_1, \dots, y_m) \ ; \ y_j \geq 0 \ \text{ for } j = 1, \dots, m\} = \mathbb{R}^m_+$$

the population state space (p-state space) of the invading population (i.e., for each $j \in \{1, \dots, m\}$ we denote by y_j the number (or density) of individuals in the jth subpopulation) and by

$$\mathcal{Z} = \{(z_1, \dots, z_n) \ ; \ z_j \geq 0 \ \text{ for } j = 1, \dots, n\} = \mathbb{R}^n_+$$

the community state space of the resident community. The c-state space of the joint community will be written as $\mathcal{Y} \times \mathcal{Z}$.

Now let $(y(t), z(t))$ denote the community state at time t, where time is measured from some conveniently chosen point. The dynamics of $(y(t), z(t))$ in time often depends not only on the present community state, but also on a number of parameters, such as per capita death rates, birth rates, etc., and, quite commonly, population models involve more than one parameter.

The aim of this paper is to study the ability of a newly introduced population to invade the existing community in the case when its basic reproduction ratio is near 1 and to derive a formula for the direction of bifurcation from a steady state in which the invading population is not present. We shall therefore concentrate on one distinguished parameter which we call the bifurcation parameter.

With this in mind we already at this point include only one (real) parameter μ and assume that the process we study is either a continuous time process described by a parametrized system of ordinary differential equations,

$$\begin{aligned} &\dot{y} = g(y, z, \mu), \\ &\dot{z} = h(y, z, \mu), \qquad\qquad y \in \mathcal{Y}, z \in \mathcal{Z}, \mu \in \mathbb{R}, \end{aligned} \tag{1.1a}$$

or a discrete time process described by a parametrized map,

$$\begin{aligned} &y \mapsto g(y, z, \mu), \\ &z \mapsto h(y, z, \mu), \qquad\qquad y \in \mathcal{Y}, z \in \mathcal{Z}, \mu \in \mathbb{R}. \end{aligned} \tag{1.1b}$$

If we consider a steady state of (1.1a) (or (1.1b)) in which the invading population is not present (these steady states lie on the boundary of the c-state space) and study the effect of perturbations corresponding to an introduction (in small quantities) of the missing population, we find that such an equilibrium is locally asymptotically stable when $\mathcal{R}_0$ of the invading population is below 1, and unstable when $\mathcal{R}_0$ exceeds 1. Moreover, stability can in these two cases be inferred from the linearization of (1.1) around the steady state.

The Perron–Frobenius theory of nonnegative matrices [1], [29], which applies for problems in population dynamics, leads us to the observation that the critical case, i.e., the case when $\mathcal{R}_0 = 1$, corresponds to the situation when

(i) the linearization of (1.1a) around the steady state yields a zero eigenvalue,
(ii) the linearization of (1.1b) around the steady state yields an eigenvalue 1.

In other words, when $\mathcal{R}_0 = 1$ we are dealing with nonhyperbolic steady states, and it is well known [24], [36] that the stability of nonhyperbolic equilibria cannot be determined by linearization alone.

Several papers (e.g., [11], [14], [18], [19], [22], [28], [30], [32], [35]) deal with this situation in the context of population models, most of them (with the exception of [14]) treating special cases or restricting their analysis to models describing the spread of infectious diseases.

In the present paper we study the critical case for general (not restricted to any particular biological background) finite dimensional population models. We will

argue that an introduction of either one new species or a reproductively isolated subpopulation of one of the existing species yields a property of (1.1) that significantly simplifies the center manifold analysis. More precisely, (1.1a) will be shown to be of the form

$$\begin{aligned} \dot{y} &= G(y,z,\mu)y, \\ \dot{z} &= h(y,z,\mu), \qquad y \in \mathcal{Y}, z \in \mathcal{Z}, \mu \in \mathbb{R}. \end{aligned} \tag{1.2}$$

A similar decomposition can be obtained for parametrized maps in (1.1b).

This will lead us to the observation that an introduction of a population whose basic reproduction ratio is close to 1 corresponds to a transcritical bifurcation of a steady state of (1.1) in which only the resident populations are present. In order to obtain the direction of bifurcation from such a steady state only the first derivatives of G and h are needed. This reduction of the order of the derivatives needed (in general, second order derivatives are needed) is, of course, most useful when one is dealing with large systems.

We will also see that among those bifurcation parameters for which

$$\begin{cases} \mu < 0 \iff \mathcal{R}_0 < 1, \\ \mu = 0 \iff \mathcal{R}_0 = 1 \end{cases}$$

holds on some neighborhood of $\mu = 0$ and the crossing of the point $\mathcal{R}_0 = 1$ occurs at a nonzero "speed," we obtain the same direction of bifurcation.

Moreover, we will show how G in (1.2) can be obtained by only considering the basic modeling ingredients, such as birth, growth, and survival rates—an approach that might be of interest to more biologically inclined readers.

The paper is structured as follows. In section 2 we study continuous time population invasion models described by (1.2). Section 3 is devoted to justifying the use of this particular form of models. We argue that this form is characteristic of population invasion models. It appears in all biological scenarios mentioned at the beginning of this Introduction and hence allows us to make a uniform study of ecological, adaptive dynamics and disease transmission models. We also show how G is obtained from basic modeling ingredients. Population models in discrete time are the theme of section 4. Section 5 provides some interpretation of the assumptions made in previous sections and draws attention to the link between continuous and discrete time population models. In section 6 we give some examples to illustrate the theory of the preceding sections. And lastly, in appendices at the end of the paper we collect some basic definitions and results regarding physiologically structured population models and put the notions of a *population, species,* and *reproductively isolated subpopulation* into a more mathematical setting.

2. Population invasion models in continuous time. We begin our study of continuous time population models by recalling the decomposition of the community state space

$$\mathcal{Y} \times \mathcal{Z} = \mathbb{R}_+^m \times \mathbb{R}_+^n,$$

where $\mathcal{Y} = \mathbb{R}_+^m$ denotes the population state space of the invading population and $\mathcal{Z} = \mathbb{R}_+^n$ the community state space of the resident community. The processes we study in this section are continuous time processes described by

$$\begin{aligned} \dot{y} &= G(y,z,\mu)y, \\ \dot{z} &= h(y,z,\mu), \qquad y \in \mathcal{Y}, z \in \mathcal{Z}, \mu \in \mathbb{R}, \end{aligned} \tag{2.1}$$

where we shall furthermore assume that $G \in M_{m\times m}(C^1(\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}, \mathbb{R}))$ and $h \in C^1(\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}, \mathbb{R}^n)$.

The form (2.1) is characteristic of continuous time population invasion models and one with which an experienced modeler may already be familiar. Those not familiar with it who feel perplexed at this point are referred to section 3, where we shall, in both mathematical and biological terms, explain why and how this form is obtained.

By writing

$$x = \begin{bmatrix} y \\ z \end{bmatrix}, \qquad f = \begin{bmatrix} Gy \\ h \end{bmatrix}, \tag{2.2}$$

we shall write (2.1) as $\dot{x} = f(x)$ and also use (1.1a) whenever this notation will be more convenient.

Consider now an equilibrium of (2.1) of the form $e = (0, z_0)$ for some $z_0 \in \mathcal{Z}$, i.e., a steady state in which only the resident populations are present. In general (2.1) can have more than one steady state of this form, and these steady states may also depend on μ. We therefore write $e(\mu) = (0, z_0(\mu))$ with $z_0(\mu) \in \mathcal{Z}$.

To study the linearized stability of $e(\mu)$ we write

$$Df\big((0, z_0(\mu), \mu)\big) = \begin{bmatrix} G(e(\mu), \mu) & 0 \\ h_y(e(\mu), \mu) & h_z(e(\mu), \mu) \end{bmatrix},$$

where

$$h_y(y, z, \mu) = \frac{\partial h(y, z, \mu)}{\partial y} \quad \text{and} \quad h_z(y, z, \mu) = \frac{\partial h(y, z, \mu)}{\partial z}. \tag{2.3}$$

Hence,

$$\sigma\big(Df(e(\mu), \mu)\big) = \sigma\big(G(e(\mu), \mu)\big) \cup \sigma\big(h_z(e(\mu), \mu)\big).$$

The next assumption that we shall make is that the equilibrium $e(\mu) = (0, z_0(\mu))$ is internally asymptotically stable; that is, it is asymptotically stable under perturbations within the invariant subspace $\{0\}^m \times \mathcal{Z}$. In other words, as long as no new population is introduced, the steady state of the resident community, $z_0(\mu)$, is locally asymptotically stable. We shall make the slightly stronger assumption that the stability can be inferred from the linearization. In mathematical terms this means that we assume

A_1. If $\lambda \in \sigma(h_z(e(\mu)))$, then $Re(\lambda) < 0$.

The spectrum of $G(e(\mu), \mu)$ thus completely determines the linearized stability of the steady state $e(\mu)$.

For the existence and uniqueness assertions that follow we need only "internal hyperbolicity," i.e., that $Re(\lambda) \neq 0$ for any $\lambda \in \sigma\big(h_z(e(\mu), \mu)\big)$. Assumption A_1 will allow us to make more detailed stability assertions, which are known as *the principle of the exchange of stability* [5], [6], [24].

Now we would like to know whether the invading population, after being introduced into the community, is able to settle in that community. As mentioned in the Introduction, the answer is *no* when the basic reproduction ratio of the newly introduced population is below 1 and *yes* when $\mathcal{R}_0$ of the invading population exceeds 1.

The basic reproduction ratio is, by definition, the spectral radius of the next generation matrix.[1] All the modeling ingredients needed to write down the next

[1] See Appendix A for more on the next generation matrix and $\mathcal{R}_0$.

generation matrix in the context of the model (see section 5) given by (2.1) are contained in G (remember that $\mathcal{R}_0$ of the invading population is the one we need), and it is known (see [13], [35] for the proof) that $\mathcal{R}_0$ of the invading population relates to the spectral bound of $G(e(\mu),\mu)$ in the following way:

$$s\big(G(e(\mu),\mu)\big) < 0 \iff \mathcal{R}_0 < 1,$$
$$s\big(G(e(\mu),\mu)\big) = 0 \iff \mathcal{R}_0 = 1,$$

where $s(.)$ denotes the spectral bound

$$s(A) = \max\{Re(\lambda); \lambda \in \sigma(A)\}.$$

Since the next generation matrix is a nonnegative matrix we can apply the Perron–Frobenius theory [1] to conclude that $\mathcal{R}_0$ is an eigenvalue with a corresponding nonnegative eigenvector. The dominant eigenvalue is often called the *transversal eigenvalue*, and if it exceeds 1 we say that the newly introduced population is able to *invade successfully*. If $\mathcal{R}_0$ is below 1, the invasion of the newly introduced population is doomed to fail.

The interesting situation to consider is hence the situation when the parameter μ is such that $s(G(e(\mu),\mu)) = 0$, the case where linearization around the steady state does not yet answer the question of invasibility.

In many models the computation of the basic reproduction ratio $\mathcal{R}_0$ and certainly the spectral bound $s(G(e(\mu),\mu))$ yields complicated functions of parameters that we may not be able to express explicitly. We therefore choose a bifurcation parameter μ with the following properties:

$$\mathrm{A}_2.\quad \begin{cases} \mu < 0 \iff s(G(e(\mu),\mu)) < 0 \iff \mathcal{R}_0 < 1, \\ \mu = 0 \iff s(G(e(\mu),\mu)) = 0 \iff \mathcal{R}_0 = 1, \\ \mu > 0 \iff s(G(e(\mu),\mu)) > 0 \iff \mathcal{R}_0 > 1. \end{cases}$$

The results that follow are based on local information only. It therefore suffices that A_2 holds on some neighborhood of $\mu = 0$.

Assumption A_2 means that the function $\mu \mapsto s(G(e(\mu),\mu))$ crosses the origin. We shall furthermore assume that this crossing occurs at a nonzero speed, i.e.,

A_3. $\frac{d}{d\mu} s(G(e(\mu),\mu))\big|_{\mu=0} > 0$.

We now denote by e an equilibrium that corresponds to $\mathcal{R}_0 = 1$, i.e., $e = e(0)$; denote $e' = e'(0)$; and also shorten the notation by defining

$$H_y = h_y(e,0), \quad H_z = h_z(e,0), \quad G_0 = G(e,0). \tag{2.4}$$

Denoting by E^c the center subspace of G_0, we shall furthermore assume the following.

A_4. $\dim E^c = 1$.

We have already given the interpretation behind the first three assumptions. We shall return to this last assumption in section 5 and explain in more detail which biological requirements are sufficient in order for A_4 to hold. Let us remark only that, in systems that arise from modeling population dynamics, the matrix G_0 will be a matrix with nonnegative off-diagonal entries, and hence the Perron–Frobenius theory guarantees that A_4 is satisfied when G_0 is irreducible.

Before stating the main result we make the following observation, which will be useful later on.

LEMMA 2.1. *Let $\mu \mapsto G(e(\mu),\mu) \in C^1(\mathbb{R},\mathbb{R}^{m\times m})$, assume* A_2 *and* A_4*, and let w and v denote, respectively, the left and the right eigenvector of G_0 corresponding to eigenvalue zero, normalized so that $v\cdot w = 1$. Then*

$$\frac{d}{d\mu} s(G(e(\mu),\mu))\Big|_{\mu=0} = w\cdot\Big(D_x G(e,0)e' + D_\mu G(e,0)\Big)v. \tag{2.5}$$

Proof. According to the implicit function theorem there exists a neighborhood of $\mu = 0$, say U, on which a branch of eigenvalues of $G(e(\mu),\mu)$ is defined. That is,

$$G(e(\mu),\mu)v(\mu) = \lambda(\mu)v(\mu) \tag{2.6}$$

for $\mu \in U$, and since $\mu \mapsto G(e(\mu),\mu) \in C^1(\mathbb{R},\mathbb{R}^{m\times m})$ we have $\mu \mapsto \lambda(\mu) \in C^1(U,\mathbb{R})$. Moreover, $\mu \mapsto v(\mu) \in C^1(U,\mathbb{R}^m)$. Differentiation of (2.6) with respect to μ yields

$$\left(\frac{\partial G}{\partial e}e'(\mu) + \frac{\partial G}{\partial \mu}\right)v(\mu) + Gv'(\mu) = \lambda'(\mu)v(\mu) + \lambda(\mu)v'(\mu). \tag{2.7}$$

Since zero is also the spectral bound of G_0 and since the spectral bound $s(G(e(\mu),\mu))$ is a continuous function of μ we have that $\lambda(\mu) = s(G(e(\mu),\mu))$ in some neighborhood of $\mu = 0$. By taking $\mu = 0$ in (2.7) and taking into account that $\lambda(0) = 0$, we obtain

$$\frac{d}{d\mu} s(G(e(\mu),\mu))\Big|_{\mu=0} v = \Big(D_x G(e,0)e' + D_\mu G(e,0)\Big)v,$$

which brings us, after premultiplication by w on both sides, to (2.5). ∎

We can now prove the following result.

THEOREM 2.2. *Consider a population model described by* (2.1)*, and let $e(\mu) = (0, z_0(\mu))$ be a steady state of* (2.1)*. Assume that* A_1*,* A_2*,* A_3*, and* A_4 *hold. Furthermore, assume that $\mu \mapsto e(\mu) \in C^1(\mathbb{R},\mathbb{R}^{n+m})$, and denote by e the steady state that corresponds to $\mathcal{R}_0 = 1$, i.e., $e = e(0)$ and by $e' = e'(0)$. Let G_0, H_y, and H_z be as in* (2.4)*, and let w and v denote, respectively, the left and the right eigenvector of G_0 corresponding to eigenvalue zero, normalized so that $v\cdot w = 1$. Let*

$$\begin{aligned} M = &\sum_{i,j,k=1,\dots,m} w_i\left(\frac{\partial G_{ij}(e,0)}{\partial y_k} + \frac{\partial G_{ik}(e,0)}{\partial y_j}\right)v_j v_k \\ &- 2\sum_{\substack{i,j=1,\dots,m\\ k=1,\dots,n}} w_i \frac{\partial G_{ij}(e,0)}{\partial z_k} v_j (H_z^{-1}H_y v)_k. \end{aligned} \tag{2.8}$$

There exists a $\delta > 0$ such that

(i) *if $M < 0$, there is a branch $\mu \mapsto (y(\mu), z(\mu))$, defined for $\mu \in (0,\delta)$, of positive, locally asymptotically stable steady states of* (2.1)*;*

(ii) *if $M > 0$, there is a branch $\mu \mapsto (y(\mu), z(\mu))$, defined for $\mu \in (-\delta, 0)$, of positive, unstable steady states of* (2.1)*.*

In other words, there exists, in a neighborhood of $\mu = 0$, a branch of nontrivial, positive (and hence biologically meaningful) steady states of (2.1)*, and M tells us about its initial slope. The former case, case* (i)*, is often referred to as a* supercritical *bifurcation and the latter, case* (ii)*, as a* subcritical *or* backward *bifurcation.*

At this point the following remarks regarding the terminology are in order.

Remark 1. As already mentioned in the Introduction, the resulting bifurcations are the so-called *transcritical bifurcations*. They correspond to an intersection of two

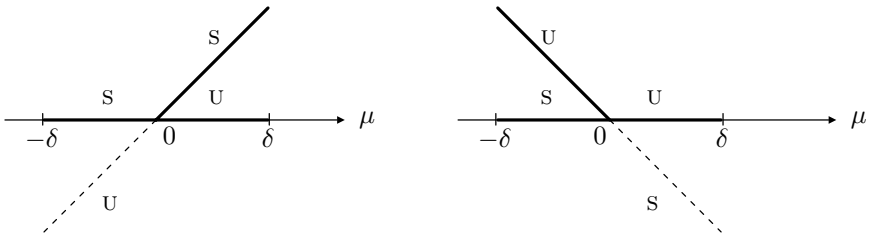


FIG. 2.1. *Supercritical (on the left) and subcritical (on the right) bifurcation. The branch of nonnegative steady states is denoted by a solid line. A dashed line represents steady states with negative components. Stability of equilibria is indicated by* **s** *(for stable) and* **u** *(for unstable).*

branches of equilibria, the trivial and the nontrivial, at $\mu = 0$, where the branches exchange stability (see Figure 2.1). In contrast with the purely mathematical point of view where these two transitions are qualitatively the same, we need to distinguish between the two in the biological context, since in that case only the nonnegative equilibria are of any relevance.

Remark 2. Note that only the first order derivatives of G and h are needed to determine the direction of bifurcation from e. Moreover, the expression M for the direction of bifurcation is independent of the bifurcation parameter except for the restrictions A_2 and A_3. In other words, provided that A_2 and A_3 are satisfied, we obtain the same direction of bifurcation for any choice of the bifurcation parameter.

The principle of the exchange of stability guarantees that the biologically meaningful, nontrivial bifurcating branch consists of stable equilibria in the supercritical case and of unstable equilibria in the subcritical case. The stable manifold of an unstable equilibrium then serves as a separatrix between the domains of attraction of the "residents only" steady state and some other attractor (frequently the same branch bent forward in a saddle node bifurcation).

Suppose now that an invader is introduced in small quantities into the resident community. In both the supercritical and the subcritical cases, this invasion will fail if the basic reproduction ratio of the invader is below 1.

In the supercritical case, the invader will be successful when its basic reproduction ratio exceeds 1, but its population size will be small when $\mathcal{R}_0 - 1$ is small. Because of this smooth transition, one sometimes calls this bifurcation *soft* or *smooth*.

In the subcritical case, on the other hand, a small introduction of the invading population for $\mathcal{R}_0 - 1$ small but positive leads to a large invader population size. Accordingly one also calls this bifurcation *hard* or *catastrophic*. Moreover, the invader can meet with success, despite $R_0 < 1$, if it is introduced in sufficiently large quantities. Catastrophic transition is illustrated in Figure 2.2, where unstable equilibria are denoted by a dashed line, stable by a solid line.

To restate Theorem 2.2 in biological terms we could say the following. When a new population, an invader, is successfully introduced into the community we observe one of the following:

(i) a smooth change to a positive but small invader population size or
(ii) a sudden, catastrophic transition to a rather large invader population size.

When all the assumptions of Theorem 2.2 are met, the sign of M in (2.8) determines which of the two scenarios we will observe in a concrete situation.

We now prove Theorem 2.2.

Proof. We have $G_0 v = 0, w^T G_0 = 0$, and $v \cdot w = 1$. By A_1, the matrix H_z is

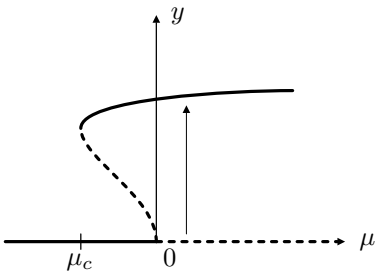


FIG. 2.2. *Catastrophic transition.*

invertible. The left and the right zero eigenvectors of $Df(e)$, denoted by W and V, are then of the form

$$W = \begin{bmatrix} w \\ 0 \end{bmatrix}, \qquad V = \begin{bmatrix} v \\ -H_z^{-1} H_y v \end{bmatrix}. \tag{2.9}$$

Moreover, $V \cdot W = 1$. By A_1 and A_4, the dimension of the center linear subspace equals 1, and the subspace is spanned by V.

We take the (generalized) right eigenvectors of $Df(e)$ for the basis of $\mathbb{R}^{m+n}$. It is known that the right (generalized) eigenvectors of $Df(e)$ that correspond to nonzero eigenvalues are orthogonal to W.

The center manifold theory [24], [36] states that the center manifold of the equilibrium e, denoted by $\mathcal{M}^c(e)$, can be (locally) parametrized by μ and a real variable u as

$$\mathcal{M}^c(e) = \{(x, \mu);\ x = e(\mu) + uV + \Phi(u, \mu)\},$$

where $\Phi(.)$ is defined on some neighborhood of the origin. Moreover, $\Phi(0,0) = D\Phi(0,0) = 0$ and $W \cdot \Phi(u, \mu) = 0$ for every u and μ.

The center manifold is also invariant under (2.1); that is,

$$\dot{x} = \dot{u}V + \dot{\Phi}(u, \mu) = f(x, \mu) = f(e(\mu) + uV + \Phi(u, \mu), \mu).$$

Since $W \cdot \frac{d}{dt}(\Phi(u, \mu)) = \frac{d}{dt}(W \cdot \Phi(u, \mu)) = 0$ and $V \cdot W = 1$ the inner product with W yields

$$\dot{u} = W \cdot f(e(\mu) + uV + \Phi(u, \mu), \mu) = w \cdot g(e(\mu) + uV + \Phi(u, \mu), \mu),$$

where we have used (2.9) in the last equality. Using the Taylor series expansion around $(e, 0)$, we can continue as follows:

$$\begin{aligned} \dot{u} \ = \ & w \cdot g(e,0) + w \cdot D_x g(e,0)(e(\mu) - e + uV + \Phi(u,\mu)) + w \cdot D_\mu g(e,0)\mu \\ & + \tfrac{1}{2} w \cdot D_{\mu\mu} g(e,0)\mu^2 + \tfrac{1}{2} w \cdot D_{xx} g(e,0)(e(\mu) - e + uV + \Phi(u,\mu))^2 \\ & + w \cdot D_{\mu x} g(e,0)(e(\mu) - e + uV + \Phi(u,\mu))\mu + \mathcal{O}(3), \end{aligned}$$

where $\mathcal{O}(3)$ contains the terms of third and higher order in u and μ.

Now, since e is an equilibrium of (2.1) the first term equals zero. So does the second because $w^T G_0 = 0$ and $D_z g(e, 0) = 0$. Since $g = Gy$ the third and the fourth terms also equal zero. Hence

$$\begin{aligned} \dot{u} = & \tfrac{1}{2} w \cdot D_{xx} g(e,0)(e(\mu) - e + uV + \Phi(u,\mu))^2 \\ & + w \cdot D_{\mu x} g(e,0)(e(\mu) - e + uV + \Phi(u,\mu))\mu + \mathcal{O}(3). \end{aligned}$$

By writing $e(\mu) = e + e'(0)\mu + \mathcal{O}(2)$ and taking into account that Φ has no constant and linear terms in u and μ, we can continue with

$$\begin{aligned}\dot{u} &= w \cdot \Big(\tfrac{1}{2} D_{xx} g(e,0) e'^2 + D_{\mu x} g(e,0) e'\Big)\mu^2 \\ &\quad + w \cdot \Big(D_{xx} g(e,0) e' V + D_{\mu x} g(e,0) V\Big)\mu u + \tfrac{1}{2} w \cdot D_{xx} g(e,0) u^2 V^2 + \mathcal{O}(3) \\ &= w \cdot \Big(D_{xx} g(e,0) e' V + D_{\mu x} g(e,0) V\Big)\mu u + \tfrac{1}{2} w \cdot D_{xx} g(e,0) u^2 V^2 + \mathcal{O}(3),\end{aligned}$$

where we have in the last equality taken into account that the first m components of $e(\mu)$ equal zero and the fact that $g = Gy$ implies $D_{zz} g(e,0) = D_{z\mu} g(e,0) = 0$. Moreover, this special form of g then gives us

$$\dot{u} = \mu u w \cdot \Big(D_x G(e,0) e' + D_\mu G(e,0)\Big) v + \tfrac{1}{2} w \cdot D_x G(e,0) u^2 V^2 + \mathcal{O}(3),$$

which, by denoting

$$N = w \cdot \Big(D_x G(e,0) e' + D_\mu G(e,0)\Big) v,$$

using (2.8), (2.9) and the fact that the first m components of e' equal zero, becomes

$$\dot{u} = \mu N u + \tfrac{1}{2} M u^2 + \mathcal{O}(3). \tag{2.10}$$

Note that, according to the Lemma 2.1, $N = \frac{d}{d\mu} s(G(e(\mu), \mu))\big|_{\mu=0}$, and so by assumption A_3, $N \neq 0$.

Now, the center manifold theory also states that the stability of the steady state under the initial system is determined by its stability under the restriction of the system to the center manifold. This restriction is now given in (2.10).

For u and μ close to zero we can neglect the higher order terms that are collected in $\mathcal{O}(3)$. The nontrivial steady state solutions of (2.10) that are near the origin are then close to the line $u = -2\mu N M^{-1}$, assuming, of course, that $M \neq 0$. By assumption, N is nonzero.

Our assumptions were that the steady state e is locally stable for $\mu < 0$ and unstable when $\mu > 0$. This steady state corresponds to $u = 0$. The local stability analysis shows that the nontrivial steady states are locally stable when $\mu > 0$ and unstable when $\mu < 0$. We shall see in the following section that we can choose the eigenvectors v and w so that all their components are nonnegative. Hence, the steady states of (2.1) that correspond to nontrivial equilibria of (2.10) can be biologically meaningful only when either $M < 0$ and $\mu > 0$ or $M > 0$ and $\mu < 0$.

Of course, when M is zero, higher order terms of the Taylor expansion need to be taken into account in order to obtain some information about the nontrivial equilibria of (2.10). □

The determination of the direction of bifurcation simplifies in a number of cases. For example, as mentioned before we can choose the eigenvectors v and w so that all their components are nonnegative. The sign of M can hence sometimes be determined without explicitly calculating the eigenvectors.

In the remarks that follow we describe a couple of situations in which further simplifications can be made.

Remark 3. One situation in which the expression for the direction of bifurcation can be further simplified is when (2.1) describes the spread of an infectious

disease. Introduction of an infectious agent to the community of hosts results in a redistribution of hosts to new compartments, such as, for example, latent or infectious individuals. A quite common assumption is that the population of hosts has reached an invariant attracting affine set (the reader can find two such examples in section 6), which means that we can eliminate one of the variables. In the case when the population of susceptible hosts is homogeneous (i.e., $n = 1$) we can, by choosing to eliminate the variable corresponding to the susceptible subpopulation (z), redefine G (which is now a function of y only and will be denoted by $\hat{G}$) and arrive at

$$M = \sum_{i,j,k=1,\dots,m} w_i \left(\frac{\partial \hat{G}_{ij}(e,0)}{\partial y_k} + \frac{\partial \hat{G}_{ik}(e,0)}{\partial y_j} \right) v_j v_k.$$

Remark 4. Another circumstance that allows for simplification of (2.8) is when the newly introduced population is homogeneous, i.e., $m = 1$. We can then choose $v = w = 1$, and the expression for the direction of bifurcation becomes

$$\tfrac{1}{2} M = \frac{\partial G(e,0)}{\partial y} - \sum_{k=1,\dots,n} \frac{\partial G(e,0)}{\partial z_k} (H_z^{-1} H_y)_k.$$

The reader can find two examples in this spirit in section 6.

3. On the characteristic form of population invasion models. The derivation of G. The purpose of this section in twofold. We first fulfill the promise made in section 6 and show that assuming that population invasion models in continuous time have the form (2.1) did not confine our study to a certain subclass of population invasion models. We will see that population invasion models, regardless of the biology that underlies them, indeed have a distinctive form of which the continuous time version is given in (2.1).

Once this part is established we shall provide the reader with a way of obtaining G by only considering the basic modeling ingredients, such as birth, survival, and reproduction rates.

So let us suppose that the process of invasion is described by the more general system (1.1) with $g \in C^2(\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}, \mathbb{R}^m)$ and $h \in C^1(\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}, \mathbb{R}^n)$.

Now, in ecology and adaptive dynamics we consider invasions of either one new species or a reproductively isolated subpopulation of one of the already present species, and we have by the very definition of reproductive isolation (see Appendix B for this definition) that $g(0, z, \mu) = 0$ for every $z \in \mathcal{Z}, \mu \in \mathbb{R}$ and also $h(y, 0, \mu) = 0$ for every $y \in \mathcal{Y}, \mu \in \mathbb{R}$.

On the other hand, when (1.1a) (or (1.1b)) describes the spread of an infectious disease into a population of susceptible hosts, a slight modification of terminology is needed. Namely, when an infectious agent is introduced into a population of susceptible hosts we indeed introduce another species, the pathogen. However, from that point on we are, on a population level, interested in how this agent spreads among the population of hosts. In this case, therefore, $\mathcal{Y}$ captures the subpopulations of hosts (i.e., members of the resident community) that carry the agent (i.e., the invading species). Since susceptible individuals don't have infected offspring we have that $g(0, z, \mu) = 0$ for every $z \in \mathcal{Z}, \mu \in \mathbb{R}$. But since infected individuals (that belong to $\mathcal{Y}$) may become susceptible again (i.e., enter $\mathcal{Z}$) once they get rid of the infection or they might have susceptible offspring, the subspace of the invading community, $\mathcal{Y} \times \{0\}^n$, may not be invariant under (1.1).

In any case we can say the following: since individuals in $\mathcal{Z}$ don't have offspring in $\mathcal{Y}$ the subspace of the resident community, $\{0\}^m \times \mathcal{Z}$, remains invariant under (1.1). In other words,

$$g(0, z, \mu) = 0 \quad \text{for every } z \in \mathcal{Z}, \mu \in \mathbb{R}. \tag{3.1}$$

Hence the following known result, due to Hadamard, can be used.

LEMMA 3.1. *Let $f = (f_1, \dots, f_k)^T \in C^r(\mathbb{R}^m \times \mathbb{R}^n, \mathbb{R}^k)$ for some $r \in \mathbb{N}$ be such that $f(0, y) = 0$ for every $y \in \mathbb{R}^n$. There exists $F \in C^{r-1}(\mathbb{R}^m \times \mathbb{R}^n, \mathbb{R}^{k \times m})$ such that*

$$f(x, y) = F(x, y)x, \qquad x = (x_1, \dots, x_m)^T.$$

The proof of this result can be found in [17].

The property (3.1) therefore yields a matrix $G \in M_{m \times m}(C^1(\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}, \mathbb{R}))$ such that g in (1.1) can be written as

$$g(y, z, \mu) = G(y, z, \mu)y, \qquad y = (y_1, \dots, y_m)^T. \tag{3.2}$$

Note that this decomposition is in general not unique, as the following simple example shows.

Example 1. Take $y = (y_1, y_2)$ and $g = y_1 y_2$. Then $G = [y_2, 0]$ and $G = [0, y_1]$ are two possible decompositions.

This nonuniqueness will, however, not affect our results—for our purposes, any correct decomposition will do. We shall nevertheless now point out one way, more interpretation motivated (and hence perhaps mainly of use to more biologically inclined readers), of obtaining G in (3.2).

The key to this decomposition is the so-called *environmental condition* [15], [16]. The defining property of an environmental condition is that individuals are independent of one another (and hence the equations are linear) when this condition is prescribed as a function of time. We then view (1.1) as a linear system together with feedback equations that tell us how, in turn, the environmental condition is influenced by the population size and composition. In general, the environment is set by all subpopulations involved. The environmental condition will hence be a function of $x = (y, z)$. Readers that are not familiar with the notion of an environmental condition and find this general definition a bit unclear are encouraged to take a look at Appendix A, where we explain the notion of an environmental condition by way of a simple example.

In order to arrive directly at the desired decomposition of g in (1.1) we first separate the reproduction in $\mathcal{Y}$ from all other processes.

Since individuals in $\mathcal{Z}$ don't have offspring in $\mathcal{Y}$ the invading population completely determines the reproduction in $\mathcal{Y}$. To describe it we define the following matrix:

$$P_{ij}(x) := \textit{the rate with which individuals with birth state } i \textit{ are born to an individual with state } j\textit{, given a constant environmental condition } x \in \mathcal{X}.$$

What remains is to describe other processes, namely maturation and survival.

We consider the dynamics of an individual's state after birth as a Markov process on the set of i-states, where the probabilities of changing a state are again determined by the environmental condition $x \in \mathcal{Y} \times \mathcal{Z}$. We define the matrix Q by

$$Q_{ij}(x) = \begin{cases} \textit{the rate of leaving state } j \textit{ to go to state } i, & i \neq j, \\ -(\textit{the rate of leaving state } j), & i = j, \end{cases}$$

given an environmental condition x.

Hence, the off-diagonal elements of Q describe the changes of states as long as the individual remains alive and does not move to $\mathcal{Z}$, and the diagonal elements denote the rate of leaving the state, either by leaving to another state in $\mathcal{Y}$, to $\mathcal{Z}$, or by death.

By taking into account all the processes, we can now write the matrix G as

$$G = P + Q. \tag{3.3}$$

This decomposition has, apart from offering biological interpretation, another advantage. Since the off-diagonal elements of G in (3.3) are nonnegative we can apply the theory of nonnegative matrices [1] to see that we can indeed choose the (left and right) eigenvector of G_0 in (2.4) corresponding to eigenvalue zero to be nonnegative.

If, for example, the off-diagonal elements of G_0 are strictly positive, the eigenvectors can be chosen to be strictly positive. In many cases this observation makes it a lot easier to determine the sign of M in (2.8).

Note that one could make a similar "per capita" description for the resident populations. However, for our purposes, this description is irrelevant since we are interested in only the c-states of the resident community that are not close to zero. It may help, though, when one wants to find $z_0(\mu)$ (see [14]).

4. Population invasion models in discrete time. Sometimes the nature of the problem, the available data, or some other reason makes it more convenient to formulate a discrete time population model. Since the linearization theorem of Hartman and Grobman (see [24], [36]) and the center manifold theory apply for discrete time dynamical systems generated by maps as well as for flows generated by vector fields we can reformulate the problem and the results so that they hold for population invasion models in discrete time.

In the same way as before we decompose the population state space

$$\mathcal{Y} \times \mathcal{Z} = \mathbb{R}^m_+ \times \mathbb{R}^n_+$$

so that $\mathcal{Y}$ denotes the population state space of the newly introduced population and $\mathcal{Z}$ the community state space of the resident community.

We shall now study processes described by a parametrized map

$$\begin{aligned} &y \mapsto G(y,z,\mu)y, \\ &z \mapsto h(y,z,\mu), \qquad\qquad y \in \mathcal{Y}, z \in \mathcal{Z}, \mu \in \mathbb{R}, \end{aligned} \tag{4.1}$$

where we shall furthermore assume that $G \in M_{m\times m}(C^1(\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}, \mathbb{R}))$ and $h \in C^1(\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}, \mathbb{R}^n)$.

We shall use the notation in (1.1b) or in (2.2) to write the map (4.1) as $x \mapsto f(x)$ whenever this notation will be more convenient.

Suppose that we have a steady state, that is, a fixed point of (4.1) of the form $e(\mu) = (0, z_0(\mu))$ for some $z_0(\mu) \in \mathcal{Z}$. The associated linear map is then given by

$$Df\big(e(\mu)\big) = \begin{bmatrix} G(e(\mu),\mu) & 0 \\ h_y(e(\mu),\mu) & h_z(e(\mu),\mu) \end{bmatrix}.$$

Hence

$$\sigma\big(Df(e(\mu),\mu)\big) = \sigma\big(G(e(\mu),\mu)\big) \cup \sigma\big(h_z(e(\mu),\mu)\big).$$

We shall again assume that the steady state $e(\mu)$ is internally asymptotically stable, i.e., that it is asymptotically stable under perturbations within the invariant subspace $\{0\}^m \times \mathcal{Z}$ and that this can be inferred from the linearization. In the discrete time setting this means that we assume

B_1. If $\lambda \in \sigma\big(h_z(e(\mu),\mu)\big)$, then $|\lambda| < 1$.

The spectrum of $G(e(\mu),\mu)$ hence determines the linearized stability of $e(\mu)$. Again, the theory of nonnegative matrices tells us that the spectral radius of G is an eigenvalue. The interesting case to consider is therefore when the parameter μ is such that $G(e(\mu),\mu)$ has an eigenvalue one, a situation where the linearization alone does not tell us whether the newly introduced population is able to settle in the community.

We now again take for a bifurcation parameter some μ such that the fixed points of (4.1) of the form $e(\mu) = (0, z_0(\mu))$ are linearly stable for $\mu < 0$ and unstable when $\mu > 0$. Thus,

$$B_2. \quad \begin{cases} \mu < 0 \iff r(G(e(\mu),\mu)) \iff \mathcal{R}_0 < 1, \\ \mu = 0 \iff r(G(e(\mu),\mu)) \iff \mathcal{R}_0 = 1, \\ \mu > 0 \iff r(G(e(\mu),\mu)) \iff \mathcal{R}_0 > 1, \end{cases}$$

where $r(.)$ denotes the spectral radius. Let us note that here $\mathcal{R}_0$ refers to the basic reproduction ratio in the context of the model, and we refer to [7], [29] for the justification of the equivalence between $r(.)$ and $\mathcal{R}_0$.

Since the results that follow rely upon local information only, it suffices that B_2 holds in some neighborhood of $\mu = 0$.

Assumption B_2 tells us that the function $\mu \mapsto r(G(e(\mu),\mu))$ crosses the point $(\mu, r(.)) = (0,1)$. We shall again assume that this crossing occurs at nonzero speed, i.e.,

$$B_3. \quad \frac{d}{d\mu} r(G(e(\mu),\mu))\Big|_{\mu=0} > 0.$$

Let e denote the equilibrium that corresponds to $\mathcal{R}_0 = 1$, i.e., $e = e(\mu = 0)$, and let $e' = e(0)$. We shall also use the notation introduced in (2.4). Denoting by E^c the center subspace of G_0, we shall furthermore assume that

$$B_4. \quad \dim E^c = 1$$

and refer the reader to the next section for the biological interpretation of this assumption.

We can now prove the discrete time analogue of Theorem 2.2.

THEOREM 4.1. *Consider a population model described by* (4.1)*, and let* $e(\mu) = (0, z_0(\mu))$ *be a steady state of* (4.1)*. Assume that* B_1*,* B_2*,* B_3*, and* B_4 *hold. Furthermore assume that* $\mu \mapsto e(\mu) \in C^1(\mathbb{R}, \mathbb{R}^{n+m})$*, and denote by* e *the steady state that corresponds to* $\mathcal{R}_0 = 1$*, i.e.,* $e = e(0)$ *and by* $e' = e'(0)$*. Let* G_0, H_y*, and* H_z *be as in* (2.4)*, and let* w *and* v *denote, respectively, the left and the right eigenvectors of* G_0 *corresponding to eigenvalue one, normalized so that* $v \cdot w = 1$*. Denote*

$$\begin{aligned} M = &\sum_{i,j,k=1,\dots,m} w_i \left(\frac{\partial G_{ij}(e,0)}{\partial y_k} + \frac{\partial G_{ik}(e,0)}{\partial y_j} \right) v_j v_k \\ &- 2 \sum_{\substack{i,j=1,\dots,m \\ k=1,\dots,n}} w_i \frac{\partial G_{ij}(e,0)}{\partial z_k} v_j ((I - H_z)^{-1} H_y v)_k. \end{aligned} \tag{4.2}$$

There exists a $\delta > 0$ *such that*

(i) *if $M < 0$, there is a branch $\mu \mapsto (y(\mu), z(\mu))$, defined for $\mu \in (0, \delta)$, of positive, locally asymptotically stable steady states of* (4.1). *In other words, the bifurcation is supercritical.*

(ii) *if $M > 0$, there is a branch $\mu \mapsto (y(\mu), z(\mu))$, defined for $\mu \in (-\delta, 0)$, of positive, unstable steady states of* (4.1). *That is, the bifurcation is subcritical.*

Remark 5. We again see that only the first order derivatives of G and h are needed to determine the direction of bifurcation from e. Moreover, the expression for the direction of bifurcation is independent of the bifurcation parameter except for the restrictions B_2 and B_3. In other words, provided that all the assumptions of Theorem 4.1 are satisfied, we obtain the same direction of bifurcation for any bifurcation parameter.

Remark 6. For some further remarks on the terminology and on the interpretation of the results of Theorem 4.1 in biological terms we refer the reader to the remarks made after Theorem 2.2.

Proof. We have $G_0 v = v, w^T G_0 = w^T$, and $w \cdot v = 1$. By B_1 the matrix $I - H_z$ is invertible. We can then calculate the left and the right eigenvectors of $Df(e)$ corresponding to eigenvalue one, denote them by W and V, and find that

$$W = \begin{bmatrix} w \\ 0 \end{bmatrix}, \qquad V = \begin{bmatrix} v \\ (I - H_z)^{-1} H_y v \end{bmatrix}. \tag{4.3}$$

Moreover, $W \cdot V = 1$.

By B_1 and B_4, the dimension of the center linear subspace equals 1, and the subspace is spanned by V. We take for the basis of $\mathbb{R}^{m+n}$ (generalized) eigenvectors of $Df(e)$. The eigenvectors of $Df(e)$ that correspond to eigenvalues different from 1 are orthogonal to W.

The center manifold theory states that there exists a center manifold of the equilibrium e, denoted by $\mathcal{M}^c(e)$, that can be locally parametrized by μ and a real variable u as

$$\mathcal{M}^c(e) = \{(x, \mu);\ x = e(\mu) + uV + \Phi(u, \mu)\},$$

where Φ is defined on some neighborhood of the origin.

Moreover, $\Phi(0,0) = D\Phi(0,0) = 0$ and $W \cdot \Phi(u, \mu) = 0$ for every u and μ. Since the center manifold is also invariant under (4.1) we have

$$\begin{aligned} x(k+1) &= e(\mu) + u(k+1)V + \Phi(u(k+1), \mu) \\ &= f(x(k), \mu) \\ &= f(e(\mu) + u(k)V + \Phi(u(k), \mu), \mu). \end{aligned}$$

We calculate the inner product with W, take into account that $W \cdot e(.) = 0$, $W \cdot V = 1$, and $W \cdot \Phi(.) = 0$, and obtain

$$u(k+1) = w \cdot g(e(\mu) + u(k)V + \Phi(u(k), \mu), \mu).$$

Written differently, the restriction of (4.1) to the center manifold is given by a map

$$u \mapsto w \cdot g(e(\mu) + uV + \Phi(u, \mu), \mu).$$

Using the Taylor series, we can now write

$$\begin{aligned} u \mapsto\ & w \cdot g(e,0) + w \cdot D_x g(e,0)(e(\mu) - e + uV + \Phi(u,\mu)) + w \cdot D_\mu g(e,0)\mu \\ &+ \tfrac{1}{2} w \cdot D_{\mu\mu} g(e,0)\mu^2 + \tfrac{1}{2} w \cdot D_{xx} g(e,0)(e(\mu) - e + uV + \Phi(u,\mu))^2 \\ &+ w \cdot D_{\mu x} g(e,0)(e(\mu) - e + uV + \Phi(u,\mu))\mu + \mathcal{O}(3), \end{aligned}$$

where $\mathcal{O}(3)$ denotes third and higher order terms in u and μ.

Now, the first term equals zero since e is a fixed point of f, and therefore $g(e,0) = 0$. The second term equals u since $w^T G_0 = w^T$, $W \cdot e(\mu) = W \cdot e = 0$, $W \cdot V = 1$, and $W \cdot \Phi(.) = 0$. Since $g = Gy$, the third and fourth terms are also equal to zero. Furthermore, by writing $e(\mu) = e(0) + e'(0)\mu + \mathcal{O}(2)$, taking into account that Φ has no constant and no linear terms in u and μ, and noting that $W^T = (w, 0)^T$, we are left with

$$\begin{aligned} u \mapsto u &+ w \cdot \Big(\tfrac{1}{2} D_{xx} g(e,0) e'^2 + D_{\mu x} g(e,0) e'\Big)\mu^2 \\ &+ w \cdot \Big(D_{xx} g(e,0) e' V + D_{\mu x} g(e,0) V\Big)\mu u \\ &+ \tfrac{1}{2} w \cdot D_{xx} g(e,0) u^2 V^2 + \mathcal{O}(3), \end{aligned}$$

which, by writing

$$N = w \cdot \Big(D_x G(e,0) e' + D_\mu G(e,0)\Big) v,$$

taking into account that $g = Gy$, (4.2), and the fact that the first m components of e equal zero, becomes

$$u \mapsto u + \tfrac{1}{2} M u^2 + \mu N u + \mathcal{O}(3). \tag{4.4}$$

Similar reasoning as in Lemma 2.1 establishes that assumptions B_2 and B_4 lead to

$$\frac{d}{d\mu} r(G(e(\mu), \mu))\big|_{\mu=0} = w \cdot \Big(D_x G(e,0) e' + D_\mu G(e,0)\Big) v, \tag{4.5}$$

and so by (4.5) and assumption B_4, $N \neq 0$.

Now, for u and μ close to zero we can neglect the higher order terms that are collected in $\mathcal{O}(3)$ and look for fixed points of $u \mapsto u + \frac{1}{2} M u^2 + \mu N u$. Nonzero fixed points are then near the line given by $u = -2\mu N M^{-1}$.

Our assumptions were that the steady state e is locally stable for $\mu < 0$ and unstable when $\mu > 0$. This steady state corresponds to $u = 0$. The local stability analysis yields that the nontrivial steady states are locally stable when $\mu > 0$ and unstable when $\mu < 0$. As we have seen in section 3, we can choose the eigenvectors v and w so that all their components are nonnegative. Hence, the steady states of (4.1) that correspond to nontrivial equilibria of (4.4) can be biologically meaningful only when either $M < 0$ and $\mu > 0$ or $M > 0$ and $\mu < 0$. If M is zero, then higher order terms of the Taylor expansion need to be taken into account in order to obtain some information about the nontrivial equilibria of (4.4). □

All the situations mentioned at the end of section 3 that lead to a simplified formula for the direction of bifurcation occur, of course, also in the discrete time setting. Modifying the obtained formulas for M to apply for discrete time models is a rather straightforward matter, and we therefore leave it to the reader.

5. On the basic reproduction ratio in the context of a model. The case $\mathcal{R}_0 = 1$. The aim of this section is to offer some interpretation of the assumptions made in previous sections. In order to do this we shall state some known results and refer the interested reader to the literature for their proofs. Two basic notions, one of the next generation matrix and the other of $\mathcal{R}_0$, are also defined in Appendix A.

The basic reproduction ratio $\mathcal{R}_0$ is defined as the expected number of offspring an "average" individual has in all of its life and is mathematically expressed as the spectral radius of the next generation operator (see [13], [15], [35]).

The key to the calculation of $\mathcal{R}_0$ of the newly introduced population in the context of the model is to decompose g in a way that separates reproduction in $\mathcal{Y}$ from other transitions (such as, for example, new infections from progressions of the disease to another stage), as was already done in section 3.

We then write

$$g(y,z) = \big(P(y,z) + Q(y,z)\big)y,$$

where P and Q are as in section 3.

Now, if we denote by $e = (0, z_0)$ a steady state of the system and define $\mathcal{P} = P(e)$ and $\mathcal{Q} = Q(e)$, then $\mathcal{P}$ is a nonnegative matrix and $\mathcal{Q}$ is a nonsingular M-matrix [1], [35]. Hence, $\mathcal{Q}$ is invertible and $-\mathcal{Q}^{-1}$ is nonnegative. Moreover, the elements of $\mathcal{Q}^{-1}$ have the following interpretation: the element $-\mathcal{Q}^{-1}_{jk}$ equals the time that an individual that was born with i-state k is expected to spend in state j [1], [13], [35]. In other words, the matrix $-\mathcal{Q}^{-1}$ describes an individual's i-state dynamics. The matrix $\mathcal{P}$ describes the reproduction, and so the matrix $-\mathcal{P}\mathcal{Q}^{-1}$ is the next generation matrix. By definition, $\mathcal{R}_0$ equals its spectral radius.

One can also prove [1], [13], [35] that the following holds:

$$\begin{aligned} \mathcal{R}_0 = r(-\mathcal{P}\mathcal{Q}^{-1}) < 1 &\iff s(\mathcal{P}+\mathcal{Q}) = s(G_0) < 0, \\ \mathcal{R}_0 = r(-\mathcal{P}\mathcal{Q}^{-1}) = 1 &\iff s(\mathcal{P}+\mathcal{Q}) = s(G_0) = 0, \end{aligned} \tag{5.1}$$

where r denotes the spectral radius and s the spectral bound.

It is reasonable to assume irreducibility of the next generation matrix. This guarantees that the spectral radius is an algebraically simple eigenvalue and that we can choose a strictly positive corresponding eigenvector [1]. In biological terms the assumption of irreducibility of the next generation matrix means that the populations are well mixed; that is, for every pair of i-states j and k, the individuals of the jth subpopulation will eventually have offspring in the kth subpopulation.

Under a more strict condition, namely the primitivity [1] of the next generation matrix, the modulus of the spectral radius is strictly greater than the modulus of all other eigenvalues of $-\mathcal{P}\mathcal{Q}^{-1}$. In biological terms the assumption of primitivity means that we require that, from some generation on, individuals with birth state j can have offspring with birth state k for any two conceivable i-states at birth, j and k.

We have assumed in sections 2 and 4 that the dimension of the center subspace of G_0 equals 1. Now, in the discrete time setting the matrix G_0 is a nonnegative matrix. Its primitivity therefore guarantees that the assumption B_4 is satisfied.

In contrast with the discrete time setting we know that G_0 in the continuous time case is a nonnegative off-diagonal matrix. The Perron–Frobenius theory then tells us that already the assumption of an irreducible G_0 guarantees that A_4 holds; i.e., a zero eigenvalue is an algebraically simple eigenvalue, and all other eigenvalues have strictly negative real parts. In biological terms an irreducible G_0 means that for every pair of i-states j and k ($j \neq k$) we will eventually observe an inflow of individuals of the jth subpopulation to the kth subpopulation.

The linearization theorem of Hartman and Grobman and the center manifold theory run for discrete time dynamical systems generated by maps parallel to the one for flows generated by vector fields, which allows us to formulate the results for both

settings. The basic reproduction ratio, $\mathcal{R}_0$, provides, due to relation (5.1), a further connection and links the continuous time results directly to corresponding discrete time results.

6. Examples. In this section we present four examples to illustrate the theory presented in previous sections.

In the first example we study a continuous time model describing the dynamics in a community in which the predator selectively forages on a stage structured prey. This example is motivated by the work of de Roos, Persson, and Thieme [11] and hopefully demonstrates how little effort is needed to study the occurrence of subcritical equilibria for a class of models, in this particular case models obtained by varying the preference of the predators.

The second example is a discrete time model describing the life cycle of biennials. This example is inspired by the work of Davydova and coworkers [9], [10] and demonstrates how the theory can also be applied to studying reproductively isolated subpopulations of the same species to see whether one missing year class is, after being introduced, able to settle among the existing year classes.

The last two examples are simple continuous time epidemic models related to the author's other work, namely, modeling the spread of infectious agents that can reside in several different parts of the host's body. Though very simple in the first place, they illustrate how the determination of the direction of bifurcation can be further simplified by assuming that the total population size has reached an equilibrium and the fact that the eigenvectors in question can be chosen to be nonnegative (see Remark 3).

After determining the direction of bifurcation from a "residents only" steady state, we shall in all of these examples write some interpretation of the results for the problem at hand. We have, however, already in the remarks after Theorem 2.2 described in biological terms what can in general be said about an invasion, given that we know the direction of bifurcation. We shall therefore not repeat these general facts in the examples and rather refer the reader to section 2.

Example 2. In this first example we study a continuous time model describing interactions in a community that consists of a stage structured prey population and a population of predators that preys exclusively on one of the prey stages.

Suppose that the prey is divided into three stages, juveniles, subadults, and adults, and let their densities be denoted, respectively, by J, S, and A. The density of the predators will be denoted by P. We describe the dynamics of the predator population that forages exclusively on the adult stages of prey by the following differential equation:

$$\frac{dP}{dt} = \big(\phi f(A) - \nu\big)P. \tag{6.1}$$

Here, ϕ indicates the conversion efficiency of prey biomass into newborn predators, ν denotes the per capita death rate of the predators, and $f(.)$ stands for the predator functional response (for example, Holling type 2 or Holling type 3 response). In what follows, the function f will not be specified; we shall assume only that it is an increasing function of the adult prey density.

We shall take a closer look at a situation in which the regulation of the prey population takes place within the subadult stage. We describe the dynamics of the

prey population by the following system of differential equations:

$$\begin{aligned} \frac{dJ}{dt} &= \beta A - \rho J - \mu_J J, \\ \frac{dS}{dt} &= \rho J - \pi(S)S - \mu_S(S)S, \\ \frac{dA}{dt} &= \pi(S)S - \mu_A A - f(A)P. \end{aligned} \tag{6.2}$$

Here, the parameters have the following meaning: β denotes the adult fecundity, ρ the maturation rate from the juvenile to the subadult stage, and μ_J the per capita death rate of the juveniles. Functions $\pi(S)$ and $\mu_S(S)$ denote, respectively, the (possibly density dependent) maturation rate of subadults into adults and the per capita death rate of the subadults. The per capita death rate of the adult prey in the absence of predators is denoted by μ_A.

Regulation of the subadult prey population through maturation and/or mortality can occur if the maturation rate $\pi(.)$ decreases and/or the mortality rate $\mu_S(.)$ increases with an increase in the density of subadults. We shall therefore assume that $\pi(.)$ is a nonincreasing and $\mu_S(.)$ a nondecreasing function of S, and we exclude the situation in which the derivatives of both vanish in some point, since the population is not regulated at all in that case.

We now first calculate the steady states of (6.2) in the absence of the predators. We obtain (the steady state values are denoted by *)

$$\begin{aligned} J^* &= \frac{\beta A^*}{\mu_J + \rho} = \frac{\beta \pi(S^*)S^*}{\mu_A(\mu_J + \rho)}, \\ A^* &= \frac{\pi(S^*)S^*}{\mu_A} \end{aligned}$$

as the steady state densities of the juvenile and adult prey, and the following equilibrium equation for the (nontrivial) steady state density of the subadult prey:

$$\frac{\rho\beta\pi(S^*)}{\mu_A(\mu_J + \rho)} = \pi(S^*) + \mu_S(S^*). \tag{6.3}$$

Now, in our previous notation we would have $y = P$ and $(z_1, z_2, z_3) = (J, S, A)$ and so $G = \phi f(A) - \nu$. Since the predator population is homogeneous, we can take $w = v = 1$. Furthermore, we consider the case when the basic reproduction ratio of the predators equals one, i.e.,

$$\mathcal{R}_0 = \frac{\phi f(A^*)}{\nu} = 1,$$

and so $f(A^*) = \frac{\nu}{\phi}$.

Now, since G is a function of A only we have

$$M = -2G'(A^*)(H_z^{-1}H_y)_3, \tag{6.4}$$

where

$$H_y = \begin{bmatrix} 0 \\ 0 \\ -f(A^*) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -\frac{\nu}{\phi} \end{bmatrix}$$

and

$$H_z = \begin{bmatrix} -(\rho+\mu_J) & 0 & \beta \\ \rho & -\big(\pi(S)S+\mu_S(S)S\big)'\big|_{S=S^*} & 0 \\ 0 & \big(\pi(S)S\big)'\big|_{S=S^*} & -\mu_A \end{bmatrix}.$$

Now, since only the last component of H_y is nonzero and we need only the third component of $H_z^{-1}H_y$ it suffices to calculate $(H_z^{-1})_{33}$. We have

$$\big(H_z^{-1}\big)_{33} = \frac{1}{\det H_z}\,(\rho+\mu_J)\big(\pi(S)S+\mu_S(S)S\big)'\big|_{S=S^*},$$

and we can now rewrite (6.4) as

$$M = 2\nu(\rho+\mu_J)f'(A^*)\big(\det H_z\big)^{-1}\big(\pi(S)S+\mu_S(S)S\big)'\big|_{S=S^*}. \tag{6.5}$$

Now, ν and $(\rho+\mu_J)$ are positive. According to our assumptions, so is $f'(A^*)$. Using (6.3), we can furthermore see that

$$\det H_z = \frac{\beta\rho S^*\big(\pi'(S^*)\mu_S(S^*)-\pi(S^*)\mu_S'(S^*)\big)}{\pi(S^*)+\mu_S(S^*)},$$

which is, by our assumptions on $\pi(.)$ and $\mu_S(.)$, strictly negative. We have therefore arrived at the fact that

$$\text{sign } M = -\text{sign }\big(\pi(S)S+\mu_S(S)S\big)'\big|_{S=S^*},$$

as was also found in [11].

We have assumed that μ_S is a nondecreasing function of the subadult density. The function $\mu_S S$ is therefore an increasing function. In more biological terms we could therefore interpret the condition required for the subcritical bifurcation to occur (i.e., $M > 0$) in the following way: an emergent Allee effect (i.e., $M > 0$) is to occur in the predator population if and only if an overcompensation in the total maturation rate $\pi(S)S$ takes place, i.e., for certain values of S, an increase in the subadult density actually decreases the total maturation rate, and that this overcompensation is sufficiently strong.

In [11] the authors also studied the cases when the predator forages exclusively on either the juvenile or the subadult prey and found that the emergent Allee effect can occur (with a suitable overcompensation in the regulation) when the predators forage on one of the nonregulating stages of the prey population and can never occur when they forage on the regulating stage.

Hopefully, this example shows how little effort it would take, with the tools that we have developed in the previous sections, to consider these and also many other situations of interest.

Example 3. In this example we consider a community of strict biennials, that is, a community that consists of two age classes, with only the oldest class reproducing. Time will in this case be measured in years. We shall label the two classes by indices 0 and 1, the 0 denoting the subpopulation of individuals that have not reached age one and 1 the subpopulation of one-year-old individuals. If individuals survive till the end of their second year, they reproduce and die.

Survival and reproduction rates are described in terms of an environmental condition I, which will be taken to be the weighed sum of the two populations. More precisely, if $x_j(t)$ denotes the number (or the density) of j-year-old individuals ($j = 0, 1$) at time t, we take

$$I(t) = c_0 x_0(t) + c_1 x_1(t).$$

The weights c_0 and c_1 are also called the *impacts* of the corresponding age classes. Now let us denote by $F_0(I(t))$ the probability of surviving the first year and by $F_1(I(t))$ the reproduction rate of individuals that survive till the end of their second year. Since increasing the I means worsening the conditions for both classes, the functions F_0 and F_1 are decreasing functions. We shall also assume that they are differentiable at least once.

We can now formulate the following discrete time model:

$$\begin{aligned} x_0(t+1) &= F_1(I(t))x_1(t), \\ x_1(t+1) &= F_0(I(t))x_0(t). \end{aligned}$$

The functions F_i are also called *sensitivities to the environment*, and the index specifies how this sensitivity depends on age. Typical examples of sensitivity functions are the so-called

(i) *Ricker family*, where $F_i(I) = a_i e^{-b_i I}$;

(ii) *Beverton–Holt family*, where $F_i(I) = a_i\big(1 + b_i I\big)^{-1}$.

In order to illustrate the theory on this example we first compute the full life cycle map; that is, we apply the map

$$\begin{bmatrix} x_0(t) \\ x_1(t) \end{bmatrix} \mapsto \begin{bmatrix} 0 & F_1(I(t)) \\ F_0(I(t)) & 0 \end{bmatrix} \begin{bmatrix} x_0(t) \\ x_1(t) \end{bmatrix}$$

twice to obtain the community state after a two year time interval. We have

$$\begin{aligned} x_0(t+2) &= F_1(I(t+1))x_1(t+1) \\ &= F_1\Big(c_0 F_1(I(t))x_1(t) + c_1 F_0(I(t))x_0(t)\Big)F_0(I(t))x_0(t) \\ &= F_1(I_1(t))F_0(I(t))x_0(t), \end{aligned}$$

where

$$I_1(t) := c_0 F_1(I(t))x_1(t) + c_1 F_0(I(t))x_0(t)$$

denotes the environmental condition in the second year.

The full life cycle map is then given by

$$\begin{aligned} x_0(t+2) &= F_1(I_1(t))F_0(I(t))x_0(t), \\ x_1(t+2) &= F_0(I_1(t))F_1(I(t))x_1(t). \end{aligned}$$

Now let us assume that only the individuals with label zero are present in every second year and that the population is in a steady state, say x_0^*. This means that

$$F_1(c_1 F_0(c_0 x_0^*)x_0^*)F_0(c_0 x_0^*) = 1. \tag{6.6}$$

Furthermore, the assumption that the basic reproduction ratio of individuals with label 1 equals 1 translates into

$$F_0(c_1F_0(c_0x_0^*)x_0^*)F_1(c_0x_0^*) = 1. \tag{6.7}$$

Now, in our previous notation we have

$$\begin{aligned} G &= F_0(I_1(t))F_1(I(t)),\\ h &= F_1(I_1(t))F_0(I(t))x_0(t). \end{aligned}$$

Let us now compute the required derivatives for the case where both sensitivity functions belong to the Ricker family. We obtain

$$\begin{aligned} \frac{\partial G}{\partial x_1} &= F_0(I_1)F_1(I)\Big(b_0b_1c_0c_1F_1(I)x_1 + b_0^2c_1^2F_0(I)x_0 - b_0c_0F_1(I) - b_1c_1\Big),\\ \frac{\partial G}{\partial x_0} &= F_0(I_1)F_1(I)\Big(b_0b_1c_0^2F_1(I)x_1 + b_0^2c_0c_1F_0(I)x_0 - b_0c_1F_0(I) - b_1c_0\Big),\\ \frac{\partial h}{\partial x_1} &= F_0(I_1)F_1(I)x_0\Big(b_1^2c_0c_1F_1(I)x_1 + b_0b_1c_1^2F_0(I)x_0 - b_1c_0F_1(I) - b_0c_1\Big),\\ \frac{\partial h}{\partial x_0} &= F_0(I_1)F_1(I)\Big(x_0\big(b_1^2c_0^2F_1(I)x_1 + b_0b_1c_0c_1F_0(I)x_0 - b_1c_1F_0(I) - b_0c_0\big) + 1\Big). \end{aligned}$$

We evaluate these derivatives in $x_0 = x_0^*, x_1 = 0$; take (6.6) and (6.7) into account; and denote the results, respectively, by $\mathcal{G}_1, \mathcal{G}_0, \mathcal{H}_1$, and $\mathcal{H}_0$. We arrive at

$$\begin{aligned} \mathcal{G}_1 &= b_0^2c_1^2F_0(c_0x_0^*)x_0^* - b_0c_0F_1(c_0x_0^*) - b_1c_1,\\ \mathcal{G}_0 &= b_0^2c_0c_1F_0(c_0x_0^*)x_0^* - b_0c_1F_0(c_0x_0^*) - b_1c_0,\\ \mathcal{H}_1 &= x_0^*\Big(b_0b_1c_1^2F_0(c_0x_0^*)x_0^* - b_1c_0F_1(c_0x_0^*) - b_0c_1\Big),\\ \mathcal{H}_0 &= x_0^*\Big(b_0b_1c_0c_1F_0(c_0x_0^*)x_0^* - b_1c_1F_0(c_0x_0^*) - b_0c_0\Big) + 1. \end{aligned}$$

Now, equalities (6.6) and (6.7) in the Ricker case imply that

$$b_0 = b_1 \qquad \text{or} \qquad \left(F_0(c_0x_0^*) = \frac{c_0}{c_1} \quad \text{and} \quad F_1(c_0x_0^*) = \frac{c_1}{c_0}\right). \tag{6.8}$$

Moreover, we can take $v = w = 1$ in (4.2). The expression for the direction of bifurcation then translates into

$$M = 2\mathcal{G}_1 - 2\mathcal{G}_0\mathcal{H}_1\big(\mathcal{H}_0 - 1\big)^{-1},$$

and one can quickly see, by taking (6.8) into account, that the bifurcation is vertical (i.e., $M = 0$), as was also found in [9].

In [9], [10] it was actually shown that the bifurcation is vertical in the stronger sense that a family of period two points exists for exactly the critical parameter combination.

Example 4. Consider an infectious disease that spreads in a population of hosts that are susceptible to this infection, and assume that there are two parts of the body (the same two parts for all individuals) that can become infected. We shall assume that one of these two parts, part one, is necessarily the part where an individual's first infection occurs. Once infected at part one, the infection can spread by endogenous transmission to part two. We shall use the following notation and assumptions:

1. β_1 denotes the rate with which one individual that is infected at part one infects a susceptible individual, β_{12} the rate with which one individual that is infected at both parts infects a susceptible individual.

2. α denotes the rate of endogenous transmission of an individual's infection from part one to part two.

3. Infected individuals become infectious at the moment of infection.

4. Infected individuals retain their infection(s) until death.

5. The death rate is the same for all individuals and is denoted by μ.

6. The population birth rate is denoted by λ.

7. All newborns are susceptible.

Now let S denote the number of susceptible individuals, I_1 the number of those infected at part one, and I_{12} the number of individuals infected at both parts. If the sizes of all subpopulations are large, we can write the following system of differential equations to describe the dynamics:

$$\begin{aligned} \frac{dS}{dt} &= \lambda - \beta_1 I_1 S - \beta_{12} I_{12} S - \mu S, \\ \frac{dI_1}{dt} &= \beta_1 I_1 S + \beta_{12} I_{12} S - (\alpha + \mu) I_1, \\ \frac{dI_{12}}{dt} &= \alpha I_1 - \mu I_{12}. \end{aligned} \tag{6.9}$$

Put into our previous notation we have

$$y = (I_1, I_{12}) \in \mathbb{R}^2_+, \quad z = S \in \mathbb{R}_+,$$

and the disease-free steady state is $e = (0, 0, \frac{\lambda}{\mu})$.

Now we have only one i-state at birth in this case—all individuals are born (from an epidemiological point of view) by acquiring the infection at part one. Each individual that is infected at part one is expected to retain (only) this infection for time $1/(\mu + \alpha)$. In this time it is expected to infect $\beta_1 \lambda/\mu$ individuals. With probability $\alpha/(\alpha + \mu)$ an individual also becomes infected at part two. It is then expected to remain as such for time $1/\mu$ and in that time infects on average $\beta_{12}\lambda/\mu$ susceptibles. The basic reproduction ratio (i.e., the expected number of new infections caused by an infected individual that is introduced into a completely susceptible population, in all of its infectious period) hence equals

$$\mathcal{R}_0 = \frac{\lambda \beta_1}{\mu(\mu + \alpha)} + \frac{\lambda \alpha \beta_{12}}{\mu^2(\mu + \alpha)}.$$

The elaboration of the direction of the bifurcation can be further simplified in this case if we assume that the total population size has reached an equilibrium. The size of the whole population is then λ/μ, and we can eliminate one of the equations in (6.9). By choosing to eliminate the first and replacing S by $\lambda/\mu - I_1 - I_{12}$ in the other two equations, we see that we don't need to compute H_y and H_z in (2.8).

To compute the direction of bifurcation we write

$$G = \begin{bmatrix} \beta_1 S - \alpha - \mu & \beta_{12} S \\ \alpha & -\mu \end{bmatrix}$$

with $S = \frac{\lambda}{\mu} - I_1 - I_{12}$. Then

$$\frac{\partial G_{2i}}{\partial y_j} = 0 \quad \text{for } i, j = 1, 2$$

and

$$\frac{\partial G_{11}(e,0)}{\partial y_1} = -\beta_1, \qquad \frac{\partial G_{11}(e,0)}{\partial y_2} = -\beta_1,$$
$$\frac{\partial G_{12}(e,0)}{\partial y_1} = -\beta_{12}, \qquad \frac{\partial G_{12}(e,0)}{\partial y_2} = -\beta_{12}.$$

Hence

$$M = w_1\big(-2\beta_1 v_1^2 - 2(\beta_1+\beta_{12})v_1v_2 - 2\beta_{12}v_2^2\big).$$

Since the off-diagonal elements of G_0 are strictly positive we can choose w to be strictly positive. Since v can always be chosen to be nonnegative we see that M is negative and the bifurcation is supercritical. In other words, control measures with which we will decrease the value of $\mathcal{R}_0$ below 1 will allow us to eradicate the disease, while the infection will spread further as long as $\mathcal{R}_0$ stays above 1.

The attentive reader must have noticed that we have not specified the bifurcation parameter. As explained in section 2, we obtain the same direction of bifurcation for all bifurcation parameters, assuming that the assumptions of Theorem 2.2 are satisfied. That they indeed hold in this case can easily be verified, and we leave the details to the reader.

Example 5. Consider again an infectious agent that spreads in the population of susceptibles, and suppose again that there are two different parts of the body (the same for all individuals) at which a susceptible can become infected. These are the additional assumptions and the notation:

1. Susceptibility and infectivity of an individual have independent influences on the rate of transmission. Susceptibility to infection does not change if one is already infected at the other part of the body. That way we can write the rate with which someone, who is already infected at $\mathcal{J} \subseteq \{1,2\}$, infects someone at part j as $b_j B_{\mathcal{J}}$.
2. Once infected at one of the parts, individuals can obtain another infection only by another cross transmission.
3. Infected individuals become infectious at the moment of infection.
4. Infected individuals retain their infection(s) until death.
5. The death rate is the same for all individuals and is denoted by μ.
6. The population birth rate is denoted by λ.
7. All newborns are susceptible.

Let S denote the number of susceptibles and I_1, I_2, I_{12} the number of infected individuals that carry an infection at, respectively, the first, second, or both parts of the body.

If we assume that the sizes of all subpopulations are large, we can describe the dynamics by the following system of differential equations:

$$\frac{dS}{dt} = \lambda - \Big(\mu + (b_1+b_2)B_1I_1 + (b_1+b_2)B_2I_2 + (b_1+b_2)B_{12}I_{12}\Big)S,$$
$$\frac{dI_1}{dt} = b_1S(B_1I_1 + B_2I_2 + B_{12}I_{12}) - b_2I_1(B_1I_1 + B_2I_2 + B_{12}I_{12}) - \mu I_1,$$
$$\frac{dI_2}{dt} = b_2S(B_1I_1 + B_2I_2 + B_{12}I_{12}) - b_1I_2(B_1I_1 + B_2I_2 + B_{12}I_{12}) - \mu I_2,$$
$$\frac{dI_{12}}{dt} = (b_2I_1 + b_1I_2)(B_1I_1 + B_2I_2 + B_{12}I_{12}) - \mu I_{12}.$$

Put into our previous notation we have

$$y = (I_1, I_2, I_{12}), \qquad z = S,$$

and the disease-free equilibrium is $e = (0, 0, 0, \frac{\lambda}{\mu})$.

There are two i-states at birth in this case, i.e., becoming first infected at part one and becoming first infected at part two. We label these two birth states with 1 and 2, respectively. The next generation matrix is hence a 2×2 matrix, which, written for the case when an infected individual is introduced into a virgin environment, takes the form

$$R = \frac{\lambda}{\mu^2} \begin{bmatrix} b_1 B_1 & b_1 B_2 \\ b_2 B_1 & b_2 B_2 \end{bmatrix}.$$

The basic reproduction ratio $\mathcal{R}_0$ equals its dominant eigenvalue. Since R is a matrix of rank one, $\mathcal{R}_0$ equals its trace,

$$\mathcal{R}_0 = \frac{\lambda}{\mu^2}\Big(b_1 B_1 + b_2 B_2\Big).$$

We shall again assume that the total population has reached an equilibrium and eliminate S by taking $S = \lambda/\mu - I_1 - I_2 - I_{12}$.

To compute the direction of bifurcation we take

$$G = \begin{bmatrix} b_1 B_1 S - b_2 Y - \mu & b_1 B_2 S & b_1 B_{12} S \\ b_2 B_1 S & b_2 B_2 S - b_1 Y - \mu & b_2 B_{12} S \\ b_2 Y & b_1 Y & -\mu \end{bmatrix},$$

with $Y = B_1 I_1 + B_2 I_2 + B_{12} I_{12}$ and $S = \lambda/\mu - I_1 - I_2 - I_{12}$.

As was the case in the previous example, all bifurcation parameters yield the same direction of bifurcation from the disease-free steady state. However, we cannot determine the sign of M right away. We hence compute the left and the right zero eigenvectors of G_0 and obtain

$$w = \begin{bmatrix} B_1, & B_2, & B_{12} \end{bmatrix}^T, \qquad v = \begin{bmatrix} b_1, & b_2, & 0 \end{bmatrix}^T.$$

Then $M = M_1 + M_2 + M_3$, where

$$\begin{aligned} M_1 &= B_1\big(-2(b_2 + b_1)b_1^2 B_1 - 2(b_1 B_2 + b_2 B_2 + b_1 B_1)b_1 b_2 - 2b_1 b_2^2 B_2\big), \\ M_2 &= B_2\big(-2(b_2 + b_1)b_2^2 B_2 - 2(b_2 B_1 + b_2 B_2 + b_1 B_1)b_1 b_2 - 2b_2 b_1^2 B_1\big), \\ M_3 &= B_{12}\big(2b_1^2 b_2 B_1 + 2(b_1 B_1 + b_2 B_2)b_1 b_2 + 2b_1 b_2^2 B_2\big). \end{aligned}$$

In contrast with the previous example, bifurcation from a disease-free steady state may not always be supercritical. However, if we don't expect any "amplification" of individual's infectiousness by multiple infected parts, that is, if we assume that

$$B_{12} \le B_1 + B_2,$$

we expect a supercritical bifurcation, and indeed a simple computation (which we leave to the reader) shows that in such a case $M < 0$ and, as in the previous example, we are able to eradicate the disease by suppressing $\mathcal{R}_0$ below 1.

Appendix A. Some general considerations concerning physiologically structured population models. In this section we review some basic definitions and results that we have used in the main part of the paper.

A.1. Environmental condition. We begin with the notion of *environmental condition* [14], [15], [16]. The defining property of the environmental condition (we shall denote it by I) is that individuals are independent of one another when I is prescribed as a function of time.

The notion of the environmental condition can perhaps most easily be clarified by way of examples. We write the following ratio dependent predator-prey model and refer the reader to section 6 and [14], [15], [16] for more examples.

Example 6. Let us consider the following predator-prey model, the so-called Michaelis–Menten-type model:

$$\begin{aligned}\dot{x} &= rx\left(1-\frac{x}{K}\right)-\frac{cxy}{my+x},\\ \dot{y} &= \quad y\left(\frac{fx}{my+x}-D\right),\end{aligned}$$

where $x(t)$ and $y(t)$ denote, respectively, the prey and predator densities at time t. In the absence of the predators the prey grows with constant intrinsic growth rate r and constant carrying capacity K. The constants D, c, m, and f stand for the predators' per capita death rate, capturing rate, half saturation rate, and conversion rate, respectively.

In this case both the predator and prey densities influence the rates with which these two populations interact. Hence, by setting $I=(I_1,I_2)=(x,y)$, we can rewrite the equations so that all interactions are expressed in terms of the environmental variable I,

$$\begin{aligned}\dot{x} &= \left[r\left(1-\frac{I_1}{K}\right)-\frac{cI_2}{mI_2+I_1}\right]x,\\ \dot{y} &= \quad \left[\left(\frac{fI_1}{mI_2+I_1}-D\right)\right]y.\end{aligned}$$

Indeed, one sees that when I is prescribed as a function of time the individuals act independently of one another; i.e., equations are linear.

A.2. The next generation operator and the basic reproduction ratio $\mathcal{R}_0$. A population model is described by a collection of rules for reproduction, maturation, and survival of individuals in a given community. The traditional way to study a population model is to separate reproduction from all other processes. One of the modeler's first tasks is then to find the set of all conceivable i-states at birth and to construct the next generation operator [13], [15].

When the set of all conceivable i-states at birth is finite, the next generation operator is a matrix which we shall denote by R and which is defined as follows:

$R_{ij}(I) :=$ *the expected number of offspring with birth state i born to one individual that was born with state j, given a constant environmental condition I.*

The basic reproduction ratio, $\mathcal{R}_0(I)$, is by definition [13], [15] the spectral radius of $R(I)$.

Now, by its very definition, $R(I)$ is a nonnegative matrix. When it is irreducible (see [1] and Appendix B for the definition), its spectral radius, i.e., $\mathcal{R}_0(I)$, is a well-defined (dominant) eigenvalue, and the corresponding eigenvector can be chosen to be positive.

The literature where one can find special examples from population biology and where the next generation matrices are constructed in the context of models is vast. We refer the reader to [13], [34] and the references therein and also to section 6 for some concrete examples.

Appendix B. On the notions of species, population, and a reproductively isolated subpopulation. In the main part of the paper we have used terms such as *population* and *reproductively isolated subpopulation* in a vague, intuitive way.

The aim of this section is to describe these notions in mathematical terms (here we are inspired by an unpublished note of Gyllenberg [23]).

First, the following definition.

DEFINITION B.1. *A square matrix A is* reducible *if there exists a permutation matrix P such that*

$$P^{-1}AP = \begin{bmatrix} A_1 & 0 \\ B & A_2 \end{bmatrix}.$$

A matrix that is not reducible is irreducible.

Following [23], *we shall call a matrix A* decomposable *if the permutation matrix P can be chosen so that $B = 0$. A matrix that is not decomposable is* indecomposable.

Now, if an element of the next generation matrix, say $R_{ij}(I)$, is strictly positive for some environmental condition I, then, by definition, individuals with birth state j can have offspring with birth state i in this environment. Or, equivalently, the predecessors of individuals with birth state i may, in the environment I, be individuals with birth state j.

If $R(I)$ is indecomposable for some environmental condition I, then all the i-states at birth are in this environment related by either ancestry or descent and hence belong to one species. On the other hand, if the next generation matrix is decomposable for some environmental condition I, then the set of i-states at birth partitions into (at least) two disjoint sets of birth states that are not reproductively connected in I.

We shall speak of *reproductive isolation* of two sets of i-states at birth (and of reproductive isolation of the corresponding subpopulations) when these two sets are reproductively isolated in any conceivable environment. Two sets of i-states at birth (and the corresponding subpopulations) that are not reproductively isolated are *reproductively connected*.

A *population* is a collection of subpopulations that are reproductively connected and are at the same time the maximal connected collection in the sense that they are reproductively isolated from every subpopulation that is not included in the collection.

We now make these newly introduced terms more precise and make the following, almost mathematical definition (where "almost" refers to the unspecified "conceivable" below).

DEFINITION B.2. *Consider a finite set $\mathcal{J}$ of i-states at birth, and let $R(.)$ denote the corresponding next generation matrix. We say the following:*

1. *The set $\mathcal{J}$ of i-states at birth (and the corresponding community) is* reproductively connected *if there exists a conceivable environmental condition I in which the corresponding next generation matrix $R(I)$ is indecomposable.*

2. *If the set $\mathcal{J}$ of i-states at birth is not reproductively connected, it consists of (at least) two* reproductively isolated *subsets of i-states at birth. In other words, if $\mathcal{J}$ is not reproductively connected, the matrix $R(I)$ is decomposable for every conceivable environmental condition I. (Note, however, that in principle, different environmental conditions may yield a different number of blocks in the next generation matrix.)*

3. *Let $\mathcal{J}_1 \subseteq \mathcal{J}$. We say that individuals with i-states at birth in $\mathcal{J}_1$ form a* population *if*

(i) *$\mathcal{J}_1$ is reproductively connected and*

(ii) *if $\mathcal{J}_1 \subseteq \mathcal{J}_2 \subseteq \mathcal{J}$ and $\mathcal{J}_2$ is reproductively connected, then $\mathcal{J}_1 = \mathcal{J}_2$.*

Reproductive isolation is certainly a property that underlies the concept of species; i.e., two different species are reproductively isolated. Reproductive isolation alone, however, is not sufficient to deduce that we actually observe different species. Think of, for example, two groups of individuals that belong to the same species but live in areas that are not connected, say on two different continents.

Another display of this phenomenon would be the semelparous species, species that reproduce only once in their lives and die afterwards. Suppose that we observe a community whose individuals live for a fixed length of time, say l years. We could characterize individuals by the year of their birth. Instead of doing so, we split the whole community into year classes according to the year of birth (modulo l); for example, if $l = 2$, we divide the community into two year classes, one consisting of individuals that were born in odd numbered years, and the other of individuals born in even numbered years. Different year classes are reproductively isolated subpopulations of the same species that interact (for example, compete for food, etc.), and we can study whether a missing year class is, after being introduced into the community, able to settle among the existing year classes. The reader can find one example in this spirit in section 6.

Consider now a community, consisting of several species perhaps, whose individuals are characterized by finitely many i-states. We introduce one new population and assume that the set of conceivable i-states of this population is also finite. We find the set of all possible i-states at birth and write the next generation matrix of the combined community, which for every conceivable environmental condition I is of the form

$$R(I) = \begin{bmatrix} R(I)_{\text{new}} & 0 \\ 0 & R(I)_{\text{old}} \end{bmatrix}.$$

Since the resident community might consist of several populations, the matrix $R(I)_{\text{old}}$ may be decomposed further into indecomposable blocks. However, finding a way to deduce the number of reproductively isolated subpopulations from the next generation matrix and recognizing the set of i-states that constitute a population is not our aim here. We therefore refrain from these further decompositions.

Acknowledgment. I thank Odo Diekmann for suggesting this problem. I am very much indebted to him as well as to Hans Metz, Mats Gyllenberg, Mark Lewis, and two anonymous referees for various comments that helped to improve this paper in many ways.

REFERENCES

[1] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[2] M. G. Bulmer, *Periodical insects*, The American Naturalist, 111 (1997), pp. 1099–1117.

[3] D. Claessen, *Dwarfs and Giants. The Dynamic Interplay of Size-Dependent Cannibalism and Competition*, Ph.D. thesis, IBED/Population Biology, University of Amsterdam, The Netherlands, 2002.

[4] D. Claessen and A. M. de Roos, *Bistability in a size-structured population model of cannibalistic fish—A continuation study*, Theor. Pop. Biol., 64 (2003), pp. 49–65.

[5] M. G. Crandall and P. H. Rabinowitz, *The principle of the exchange of stability*, in Proceedings of the International Symposium on Dynamic Systems, Gainesville, FL, 1976, Bedernek and Cesari, eds., Academic Press, New York, 1977, pp. 27–41.
[6] M. G. Crandall and P. H. Rabinowitz, *Bifurcation from simple eigenvalue*, J. Funct. Anal., 8, (1971), pp. 321–340.
[7] J. M. Cushing, *An Introduction to Structured Population Dynamics*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 71, SIAM, Philadelphia, 1998.
[8] J. M. Cushing and J. Li, *Intra-specific competition and density dependent juvenile growth*, Bull. Math. Biol., 54 (1992), pp. 503–519.
[9] N. Davydova, *Old and Young. Can They Coexist?*, Ph.D. thesis, Mathematical Institute, University of Utrecht, The Netherlands, 2004; available online at http://www.library.uu.nl/digiarchief/dip/diss/2004-0115-092805/inhoud.html.
[10] N. Davydova, O. Diekmann, and S. A. van Gils, *Year class coexistence or competitive exclusion for strict biennials?*, J. Math. Biol., 46 (2003), pp. 95–131.
[11] A. de Roos, L. Persson, and H. R. Thieme, *Emergent Allee effects in top predators feeding on structured prey populations*, Proc. R. Soc. Lond. B, 270 (2003), pp. 611–618.
[12] O. Diekmann, *A beginner's guide to adaptive dynamics*, in Mathematical Modelling of Population Dynamics, Banach Center Publications, Vol. 63, Polish Academy of Science, Warsaw, 2004, pp. 47–86.
[13] O. Diekmann and J. A. P. Heesterbeek, *Mathematical Epidemiology of Infectious Diseases, Model Building, Analysis, and Interpretation*, Wiley, New York, 2000.
[14] O. Diekmann, M. Gyllenberg, and J. A. J. Metz, *Steady state analysis of structured population models*, Theor. Pop. Biol., 63 (2003), pp. 309–338.
[15] O. Diekmann, M. Gyllenberg, J. A. J. Metz, and H. R. Thieme, *On the formulation and analysis of general deterministic structured population models,* I. *Linear theory*, J. Math. Biol., 36 (1998), pp. 349–388.
[16] O. Diekmann, M. Gyllenberg, H. Huang, M. M. Kirkilionis, J. A. J. Metz, and H. R. Thieme, *On the formulation and analysis of general deterministic structured population models,* II. *Nonlinear theory*, J. Math. Biol., 43 (2001), pp. 157–189.
[17] J. J. Duistermaat and J. A. C. Kolk, *Multidimensional Real Analysis* 1: *Differentiation*, Cambridge Stud. Adv. Math., Cambridge University Press, Cambridge, UK, 2004.
[18] J. Dushoff, *Incorporating immunological ideas in epidemiological models*, J. Theor. Biol., 180 (1996), pp. 181–187.
[19] J. Dushoff, W. Huang, and C. Castillo-Chavez, *Backwards bifurcations and catastrophe in simple models of fatal disease*, J. Math. Biol., 36 (1998), pp. 227–248.
[20] L. Edelstein-Keshet, *Mathematical Models in Biology*, Random House, New York, 1988.
[21] S. A. H. Geritz, J. A. J. Metz, E. Kisdi, and G. Meszena, *The dynamics of adaptation and evolutionary branching*, Phys. Rev. Lett., 78 (1997), pp. 2024–2027.
[22] D. Greenhalgh, O. Diekmann, and M. C. M. de Jong, *Subcritical endemic steady states in mathematical models for animal infections with incomplete immunity*, Math. Biosci., 165 (2000), pp. 1–25.
[23] M. Gyllenberg, *On the Definition of Species in Structured Population Models*, unpublished manuscript, Department of Mathematics and Statistics, University of Helsinki, Finland, 2001.
[24] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.
[25] J. Hofbauer and K. Sigmund, *The Theory of Evolution and Dynamical Systems*, Cambridge University Press, Cambridge, UK, 1988.
[26] G. E. Hutchinson, *An Introduction to Population Ecology*, Yale University Press, New Haven, CT, 1978.
[27] B. W. Kooi, M. P. Boer, and S. A. L. M. Kooijman, *Resistance of a food chain to invasion by a top predator*, Math. Biosci., 157 (1999), pp. 217–236.
[28] S. Lessard and S. Karlin, *A criterion for stability-instability at fixation states involving an eigenvalue one with applications to population genetics*, Theor. Pop. Biol., 22 (1982), pp. 108–126.
[29] C.-K. Li and H. Schneider, *Applications of Perron–Frobenius theory to population dynamics*, J. Math. Biol., 44 (2002), pp. 450–462.
[30] K. G. Magnusson, *Destabilizing effect of cannibalism on a structured predator-prey system*, Math. Biosci., 155 (1999), pp. 61–75.
[31] J. A. J. Metz and O. Diekmann, *The Dynamics of Physiologically Structured Populations*, Lecture Notes in Biomath. 68, Springer, Berlin, 1986.
[32] M. G. Neubert and M. Kot, *The subcritical collapse of predator populations in discrete time predator-prey models*, Math. Biosci., 110 (1992), pp. 45–66.

[33] H. L. Smith and P. Waltman, *The Theory of the Chemostat-Dynamics of Microbial Competition,* Cambridge Stud. Math. Biol., Cambridge University Press, Cambridge, UK, 1995.
[34] H. R. Thieme, *Mathematics in Population Biology*, Princeton Ser. Theoret. Comput. Biol., Princeton University Press, Princeton, NJ, 2003.
[35] P. van den Driessche and J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci., 180 (2002), pp. 29–48.
[36] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag New York, 2003.

SIAM J. APPL. MATH.
Vol. 66, No. 4, pp. 1454–1472

AN EPIDEMIC MODEL WITH POPULATION DISPERSAL AND INFECTION PERIOD*

WENDI WANG† AND XIAO-QIANG ZHAO‡

Abstract. An epidemic model is proposed to incorporate population dispersals between patches and to include a constant infection period. A basic reproduction number of the model is established by means of a next generation matrix. It is found that a disease may spread when the population migrates in two patches, even though it dies out in each isolated patch. It is also found that the disease admits multiple exchanges between persistence and extinction for some types of migrations of individuals.

Key words. population dispersal, infection period, reproduction number, persistence, extinction

AMS subject classifications. 92D30, 34K20, 37N25

DOI. 10.1137/050622948

1. Introduction.

Population movements and the spatial structure of population communities can profoundly affect the dynamic process of an epidemic disease. This has been demonstrated by many communicable diseases. SARS was first reported in Guangdong Province of China in November of 2002. The emerging disease spread very quickly, due to the traveling of infectious persons by airplanes, trains, and buses, to some other regions in the mainland of China, as well as Hong Kong, Singapore, Vietnam, Canada, etc. By late June of 2003, it had spread to 32 countries and regions, causing about 800 deaths and more than 8000 infections (see, e.g., [21, 27]). For measles, Bartlett [5] discovered that population size was a crucial determinant of disease persistence. In large towns, measles was endemic with periodic eruptions. In cities below a population size threshold, measles displayed an epidemic pattern with complete disappearance of the disease between epidemics. These show that spatial structures and population movements, which give rise to spatio-temporal variations of population dynamics, may be of crucial importance for the prevalence of diseases. Thus, it is important to study how population movement, spatial structure, and disease transmission interact to determine the evolution of diseases. Of particular interest is the basic reproduction number of the disease under the influence of these components. The basic reproduction number is defined as the number of infections generated in a completely susceptible population by a single infected individual. In many cases, it serves as the threshold of disease transmissions, having profound implications for the control of diseases.

We choose our spatial scale so that we can represent space in a discretized way; i.e., the space is split into discrete patches. Here, one patch may represent one city or one village, and population movements in space are simulated by population dispersals among patches. The study of the effects of population dispersal on disease dynamics

*Received by the editors January 19, 2005; accepted for publication (in revised form) January 30, 2006; published electronically May 12, 2006.
http://www.siam.org/journals/siap/66-4/62294.html

†Department of Mathematics, Southwest University, Chongqing, 400715, People's Republic of China (wendi@swnu.edu.cn). This author's research was supported by the NSF of China (grant 10571143).

‡Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NF A1C 5S7, Canada (xzhao@math.mun.ca). This author's research was supported by the NSERC of Canada and the MITACS of Canada.

in the setting of patchy space has been carried out in the context of SARS [21], influenza [22], measles [6], tuberculosis [13], and malaria [20]. These papers explore the influences of population dispersal by means of computer simulations or by analytic methods. In order to simulate the evolution of spatial epidemics and calculate their basic reproduction numbers, Arino and van den Driessche [2, 3] formulated epidemic models with population traveling among cities in which the residences of individuals are maintained. Wang and Zhao [29, 30] considered epidemic models of multipatches without any record of the residence of individuals. Brauer and van den Driessche [4], Castillo-Chavez and Yakubu [7], and Wang and Mulone [28] are among other studies of epidemic models of metapopulations. Based upon their calculations of basic reproduction numbers, these papers established the thresholds of spatial disease transmissions. The models capture the essence of SIR (susceptible→infected→recovered) epidemics or SIS (susceptible→infected→susceptible) epidemics. However, one common feature of the above works is that the duration of an infectious period is described by an exponential distribution. This mathematically convenient assumption is equivalent to assuming that the chance of recovery within a given time interval is constant, regardless of the time since infection. Epidemiologically, this is quite unrealistic, as is demonstrated by statistical studies of the transmission dynamics of many diseases [19]. In practice, the infectious periods for many diseases including myxomatosis appear to be distributed fairly closely around mean survival times [1]. As a first approximation, we can assume that infectious individuals remain infectious for the same amount of time. The purpose of the present paper is to incorporate a constant infectious period into an SIR epidemic model with the spatial structure of patches and study its evolutionary behavior in terms of the basic reproduction number.

The remaining parts of this paper are organized as follows. In the next section, we present the formulation of the model. Section 3 is devoted to the establishment of the basic reproduction number of the disease. Based upon this result, we provide three examples to illustrate the effect of the population dispersal on the spread of the disease, in section 4. Section 5 gives a brief discussion of the main results.

2. Model formulations. We consider an SIR type of disease transmission. The population is divided into three classes: susceptible individuals, infectious individuals, and recovered individuals. Susceptible individuals become infective after contact with infective individuals. Infectious individuals become recovered due to treatment or when an infection age passes the infection period. Measles, rubella, smallpox, and SARS are typical diseases of SIR type. For simplicity, we consider only two patches, which is the minimal case needed to investigate the influence of population mobility on spatial epidemic transmissions. We denote the density of susceptible individuals in patch i by S_i, the density of infective individuals in patch i by I_i, the density of recovered individuals in patch i by R_i, and the population size in patch i by N_i. Thus, $N_i = S_i + I_i + R_i$. In order to accommodate more epidemic diseases that have longer time scales, we consider the demography of populations and suppose that the demographic structure of the population in patch i is described by

$$\frac{dN_i}{dt} = B_i(N_i(t))N_i(t) - \mu_i N_i(t),$$

where B_i is the per capita birth rate and μ_i the per capita death rate. This type of demography has been adopted by [9, 12, 16]. If $B_i = \mu_i$, then the population size is a constant, which is suitable when the duration of an epidemic disease is much shorter than the lifespan of the individuals in a population. As mentioned in [9], three types of birth functions $B_i(N_i)$ can be found in the biological literature:

(C1) $B_i(N_i) = H_i e^{-A_i N_i}$ with $A_i > 0$, $H_i > 0$;
(C2) $B_i(N_i) = \frac{p_i}{q_i + N_i^{m_i}}$ with $p_i, q_i, m_i > 0$;
(C3) $B_i(N_i) = \frac{A_i}{N_i} + H_i$ with $A_i > 0$, $H_i > 0$.

We assume that the disease transmission in each patch obeys the mass action incidence. If there is no population dispersal between patches, i.e., the patches are isolated, with an exponentially distributed infectious period, the dynamics of disease transmission in the ith patch is governed by

$$\begin{aligned} \frac{dS_i}{dt} &= B_i(N_i(t))N_i(t) - \mu_i S_i(t) - k_i S_i(t) I_i(t), \\ \frac{dI_i}{dt} &= k_i S_i(t) I_i(t) - (\mu_i + \gamma_i) I_i(t), \\ \frac{dR_i}{dt} &= \gamma_i I_i(t) - \mu_i R_i(t), \end{aligned} \tag{2.1}$$

where k_i is the disease transmission coefficient and γ_i is the recovery rate of infected individuals.

When the patches are connected, the dynamics of disease transmission is described by

$$\begin{aligned} \frac{dS_1}{dt} &= B_1(N_1(t))N_1(t) - (\mu_1 + d_1)S_1(t) - k_1 S_1(t) I_1(t) + d_2 S_2(t), \\ \frac{dS_2}{dt} &= B_2(N_2(t))N_2(t) - (\mu_2 + d_2)S_2(t) - k_2 S_2(t) I_2(t) + d_1 S_1(t), \\ \frac{dI_1}{dt} &= k_1 S_1(t) I_1(t) - (\mu_1 + \gamma_1 + b_1) I_1(t) + b_2 I_2(t), \\ \frac{dI_2}{dt} &= k_2 S_2(t) I_2(t) - (\mu_2 + \gamma_2 + b_2) I_2(t) + b_1 I_1(t), \\ \frac{dR_1}{dt} &= \gamma_1 I_1(t) - (\mu_1 + c_1) R_1(t) + c_2 R_2(t), \\ \frac{dR_2}{dt} &= \gamma_2 I_2(t) - (\mu_2 + c_2) R_2(t) + c_1 R_1(t), \end{aligned} \tag{2.2}$$

where d_1 represents the rate at which susceptible individuals migrate from the first patch to the second patch, d_2 the rate at which susceptible individuals migrate from the second patch to the first patch, b_1 the rate at which infectious individuals migrate from the first patch to the second patch, b_2 the rate at which infected individuals migrate from the second patch to the first patch, c_1 the rate at which recovered individuals migrate from the first patch to the second patch, and c_2 the rate at which recovered individuals migrate from the second patch to the first patch. In this model, we neglect the death and birth processes of individuals when they are dispersing and neglect the time that individuals take to move between patches. Furthermore, for the convenience of mathematical analysis, it is assumed that $b_1 > 0$ and $b_2 > 0$.

We assume that all infectious individuals have a constant length of infection τ. This means that the duration of the infectious period is described by a step function, which is one of the conventional methods for representing infectious periods [8, 11, 16], and the biological motivations are presented in the introduction section. Let a be the infection age, i.e., the time since infection, and let $I_i(a,t)$ be the density of infected individuals at time t with respect to infection age a in the ith patch. We assume that the number of individuals recovered due to treatment per unit time is proportional to the total of infectious individuals. This means that we neglect the time delay for

treatment to take effect, which is reasonable when treatment takes effect in a short time. As a consequence, infected individuals in patch i are treated at a constant rate $r_i \geq 0$ up to infection-age τ, when any remaining infected individuals of infection-age τ immediately transfer to the recovered class. For simplicity, we assume that the death rates, the disease transmission coefficients, the treatment rates, and the migration rates are independent of infection ages. Then the force of infection in patch i at time t is $k_i \int_0^\tau I_i(a,t)da$. Thus, (2.2) can be modified as

$$
\begin{aligned}
&\frac{dS_1}{dt} = B_1(N_1(t))N_1(t) - (\mu_1 + d_1)S_1(t) - \lambda_1(t)S_1(t) + d_2S_2(t),\\
&\frac{dS_2}{dt} = B_2(N_2(t))N_2(t) - (\mu_2 + d_2)S_2(t) - \lambda_2(t)S_2(t) + d_1S_1(t),\\
&\frac{\partial I_1}{\partial t} + \frac{\partial I_1}{\partial a} = -(\mu_1 + r_1 + b_1)I_1(a,t) + b_2I_2(a,t), \quad 0 < a \leq \tau,\\
&\frac{\partial I_2}{\partial t} + \frac{\partial I_2}{\partial a} = -(\mu_2 + r_2 + b_2)I_2(a,t) + b_1I_1(a,t), \quad 0 < a \leq \tau,\\
&\frac{dR_1}{dt} = r_1\int_0^\tau I_1(a,t)da + I_1(\tau,t) - (\mu_1 + c_1)R_1(t) + c_2R_2(t),\\
&\frac{dR_2}{dt} = r_2\int_0^\tau I_2(a,t)da + I_2(\tau,t) - (\mu_2 + c_2)R_2(t) + c_1R_1(t),\\
&\lambda_i(t) = k_i\int_0^\tau I_i(a,t)da, \qquad N_i(t) = S_i(t) + R_i(t) + \int_0^\tau I_i(a,t)da,\\
&I_i(0,t) = \lambda_i(t)S_i(t), \quad i = 1,2,
\end{aligned}
\tag{2.3}
$$

with the initial conditions given by

$$
\begin{aligned}
&S_i(0) = S_i^0 > 0, \qquad R_i(0) = R_i^0 \geq 0, \quad i = 1,2,\\
&I_i(a,0) = f_i(a) \geq 0 \quad \text{for } 0 \leq a \leq \tau,\ i = 1,2.
\end{aligned}
\tag{2.4}
$$

Let $P_i(t) = \int_0^\tau I_i(a,t)da$ be the total density of infected members at time t in the ith patch. We derive the equations for $P_1(t)$ and $P_2(t)$ for $t \geq \tau$. Set $V_i(a,t) = I_i(t-a,t)$ for $0 \leq t-a \leq \tau$ and $\mathbf{V}(a,t) = (V_1(a,t), V_2(a,t))^T$, where T represents a transpose of a vector. Then $\mathbf{V}$ satisfies

$$
\frac{\partial \mathbf{V}(a,t)}{\partial t} = \mathbf{B}\mathbf{V}(a,t), \quad a \leq t \leq a + \tau,
\tag{2.5}
$$

where

$$
\mathbf{B} = \begin{bmatrix} -\mu_1 - r_1 - b_1 & b_2 \\ b_1 & -\mu_2 - r_2 - b_2 \end{bmatrix}.
$$

Integrating (2.5) from a to t, we have

$$
\mathbf{V}(a,t) = \exp(\mathbf{B}(t-a))(I_1(0,a), I_2(0,a))^T, \quad a \leq t \leq a + \tau,
$$

and hence

$$
\mathbf{I}(a,t) = \mathbf{V}(t-a,t) = \exp(\mathbf{B}a)(I_1(0,t-a), I_2(0,t-a))^T, \quad a \leq \tau.
\tag{2.6}
$$

Set

$$
(b_{ij}(a)) := \exp(\mathbf{B}a), \qquad Q_i(t) := k_iS_i(t)P_i(t).
$$

It then follows from (2.3) that

$$\begin{aligned} I_1(a,t) &= b_{11}(a)Q_1(t-a) + b_{12}(a)Q_2(t-a),\\ I_2(a,t) &= b_{21}(a)Q_1(t-a) + b_{22}(a)Q_2(t-a) \end{aligned} \tag{2.7}$$

for $t \geq \tau \geq a$.

Integrating (2.7) from 0 to τ, we have

$$\begin{aligned} P_1(t) &= \int_0^\tau b_{11}(a)Q_1(t-a)da + \int_0^\tau b_{12}(a)Q_2(t-a)da, \quad t \geq \tau,\\ P_2(t) &= \int_0^\tau b_{21}(a)Q_1(t-a)da + \int_0^\tau b_{22}(a)Q_2(t-a)da, \quad t \geq \tau, \end{aligned} \tag{2.8}$$

which is equivalent to

$$\mathbf{P}(t) = \int_0^\tau \exp(\mathbf{B}a)\mathbf{Q}(t-a)da = \int_{t-\tau}^t \exp(\mathbf{B}(t-s))\mathbf{Q}(s)ds, \quad t \geq \tau, \tag{2.9}$$

where $\mathbf{P}(t) = (P_1(t), P_2(t))^T$ and $\mathbf{Q}(t) = (Q_1(t), Q_2(t))^T$. It then follows that

$$\frac{d\mathbf{P}}{dt} = \mathbf{Q}(t) - \exp(\mathbf{B}\tau)\mathbf{Q}(t-\tau) + \mathbf{B}\mathbf{P}(t), \quad t \geq \tau. \tag{2.10}$$

Define

$$\gamma_i(t) := r_i P_i(t) + b_{i1}(\tau)Q_1(t-\tau) + b_{i2}(\tau)Q_2(t-\tau), \quad i = 1, 2. \tag{2.11}$$

Consequently, we obtain the following time-delayed model:

$$\begin{aligned} \frac{dS_1}{dt} &= B_1(N_1(t))N_1(t) - (\mu_1 + d_1)S_1(t) - Q_1(t) + d_2 S_2(t),\\ \frac{dS_2}{dt} &= B_2(N_2(t))N_2(t) - (\mu_2 + d_2)S_2(t) - Q_2(t) + d_1 S_1(t),\\ \frac{d\mathbf{P}}{dt} &= \mathbf{Q}(t) - \exp(\mathbf{B}\tau)\mathbf{Q}(t-\tau) + \mathbf{B}\mathbf{P}(t),\\ \frac{dR_1}{dt} &= \gamma_1(t) - (\mu_1 + c_1)R_1(t) + c_2 R_2(t),\\ \frac{dR_2}{dt} &= \gamma_2(t) - (\mu_2 + c_2)R_2(t) + c_1 R_1(t),\\ N_i(t) &= S_i(t) + R_i(t) + P_i(t), \quad i = 1, 2, \end{aligned} \tag{2.12}$$

for $t \geq \tau$. In view of (2.9), we need to impose the following condition on initial functions:

$$\mathbf{P}(\tau) = \int_0^\tau \exp(\mathbf{B}(\tau - s))\mathbf{Q}(s)ds. \tag{2.13}$$

It is easy to see that (2.12) is an autonomous functional differential system defined on $C([0,\tau], R_+^6)$. Without loss of generality, we will consider (2.12) on $C([-\tau, 0], R_+^6)$ (after a time translation) under the condition

$$\mathbf{P}(0) = \int_{-\tau}^0 \exp(-\mathbf{B}s)\mathbf{Q}(s)ds. \tag{2.14}$$

We finish this section with discussion of the well-posedness of system (2.12), with initial values subject to condition (2.14), and the positivity of its solutions. Assume that each $B_i(N_i)N_i$ extends to a continuously differentiable function $G_i(N_i)$ on $[0,\infty)$ with $G_i(0) \geq 0$. Let $\mathbf{u}(t) = (\mathbf{S}(t), \mathbf{P}(t), \mathbf{R}(t))$ be a continuous function from $[-\tau, \sigma)$ to R^6_+ for some $\sigma > 0$. For each $t \in [0, \sigma)$, we define $\mathbf{u}_t \in C([-\tau, 0], R^6_+)$ by $\mathbf{u}_t(s) = \mathbf{u}(t+s)$ for all $s \in [-\tau, 0]$. Set

$$X := \left\{ (\mathbf{S}, \mathbf{P}, \mathbf{R}) \in C([-\tau, 0], R^6_+) : \mathbf{P}(0) = \int_{-\tau}^{0} \exp(-\mathbf{B}s)\mathbf{Q}(s)ds \right\}.$$

By the standard theory of functional differential equations (see, e.g., [14, Theorems 2.1 and 2.3] or [18, Theorems 2.1 and 2.2]), it follows that for any $\boldsymbol{\phi} \in C([-\tau, 0], R^6_+)$ there exists a unique solution $\mathbf{u}(t, \boldsymbol{\phi})$ of system (2.12) satisfying $\mathbf{u}_0 = \boldsymbol{\phi}$, which is defined on its maximal interval of existence $[0, \sigma_{\boldsymbol{\phi}})$. We first show that X is positively invariant for solutions of system (2.12) in the sense that for any $\boldsymbol{\phi} \in X$ we have $\mathbf{u}_t(\boldsymbol{\phi}) \in X$ for all $t \in [0, \sigma_{\boldsymbol{\phi}})$. Let $\mathbf{u}(t, \boldsymbol{\phi}) = (\mathbf{S}(t), \mathbf{P}(t), \mathbf{R}(t))$, and define

$$\mathbf{W}(t) := \int_{t-\tau}^{t} \exp(\mathbf{B}(t-s))\mathbf{Q}(s)ds \quad \forall t \in [0, \sigma_{\boldsymbol{\phi}}).$$

It then follows that

$$\frac{d\mathbf{W}(t)}{dt} = \mathbf{Q}(t) - \exp(\mathbf{B}\tau)\mathbf{Q}(t-\tau) + \mathbf{B}\mathbf{W}(t) \quad \forall t \in [0, \sigma_{\boldsymbol{\phi}}).$$

Thus, we obtain

$$\frac{d(\mathbf{P}(t) - \mathbf{W}(t))}{dt} = \mathbf{B}(\mathbf{P}(t) - \mathbf{W}(t)) \quad \forall t \in [0, \sigma_{\boldsymbol{\phi}}).$$

Since $\boldsymbol{\phi} \in X$, we have $\mathbf{P}(0) = \mathbf{W}(0)$, and hence

$$\mathbf{P}(t) - \mathbf{W}(t) = \exp(\mathbf{B}t)(\mathbf{P}(0) - \mathbf{W}(0)) = \mathbf{0} \quad \forall t \in [0, \sigma_{\boldsymbol{\phi}}).$$

This means that

(2.15)
$$\mathbf{P}(t) = \int_{t-\tau}^{t} \exp(\mathbf{B}(t-s))\mathbf{Q}(s)ds = \int_{-\tau}^{0} \exp(\mathbf{B}(-s))\mathbf{Q}(t+s)ds \quad \forall t \in [0, \sigma_{\boldsymbol{\phi}}).$$

By (2.15) and the differential equations for $S_1(t)$, $S_2(t)$, $R_1(t)$, and $R_2(t)$, it then easily follows that for any $\boldsymbol{\phi} \in X$, $\mathbf{u}(t, \boldsymbol{\phi})$ is (componentwise) nonnegative on $[0, \sigma_{\boldsymbol{\phi}})$, and $\mathbf{u}_t(\boldsymbol{\phi}) \in X$ for all $t \in [0, \sigma_{\boldsymbol{\phi}})$.

3. Basic reproduction number. For simplicity, from this point on we always assume that two birth functions are of type (C3), that is, $B_i(N_i) = \frac{A_i}{N_i} + H_i$, $i = 1, 2$. Thus, the model (2.12) reduces to

(3.1)
$$\begin{aligned}
\frac{dS_1}{dt} &= A_1 + H_1 N_1(t) - (\mu_1 + d_1)S_1(t) - Q_1(t) + d_2 S_2(t),\\
\frac{dS_2}{dt} &= A_2 + H_2 N_2(t) - (\mu_2 + d_2)S_2(t) - Q_2(t) + d_1 S_1(t),\\
\frac{d\mathbf{P}}{dt} &= \mathbf{Q}(t) - \exp(\mathbf{B}\tau)\mathbf{Q}(t-\tau) + \mathbf{B}\mathbf{P}(t),
\end{aligned}$$

$$\frac{dR_1}{dt} = \gamma_1(t) - (\mu_1 + c_1)R_1(t) + c_2R_2(t),$$
$$\frac{dR_2}{dt} = \gamma_2(t) - (\mu_2 + c_2)R_2(t) + c_1R_1(t),$$
$$N_i(t) = S_i(t) + R_i(t) + P_i(t), \quad i = 1, 2,\ t \geq 0.$$

The objective of this section is to establish a basic reproduction number for model (3.1) and to show that it is a threshold for the disease invasion. Assume that we have the following:

(H) $A_i, H_i, \mu_i, k_i, b_i$, $i = 1, 2$, are positive constants with $H_i < \mu_i$; d_i and c_i are nonnegative constants for $i = 1, 2$.

Define

$$X_L := \left\{ \boldsymbol{\phi} = (\phi_1, \ldots, \phi_6) \in X : \sum_{i=1}^{6} \phi_i(0) \leq L \right\} \quad \forall L \geq 0$$

and

$$m := \min\{\mu_1 - H_1, \mu_2 - H_2\}, \qquad L^* := \frac{A_1 + A_2}{m}.$$

Then we have the following preliminary result.

LEMMA 3.1. *Let* (H) *hold. Then for any $L > L^*$, the set X_L is positively invariant for solutions of* (3.1)*, and every solution $\mathbf{u}(t, \boldsymbol{\phi})$ of* (3.1) *with $\boldsymbol{\phi} \in X$ eventually enters into $[0, L]^6$.*

Proof. By (3.1) and (2.11), we obtain

$$\begin{aligned} \frac{dN_1}{dt} &= A_1 - (\mu_1 - H_1)N_1 - d_1S_1 + d_2S_2 - b_1P_1 + b_2P_2 - c_1R_1 + c_2R_2, \\ \frac{dN_2}{dt} &= A_2 - (\mu_2 - H_2)N_2 - d_2S_2 + d_1S_1 - b_2P_2 + b_1P_1 - c_2R_2 + c_1R_1. \end{aligned} \tag{3.2}$$

Let $N = N_1 + N_2$. It then follows that

$$\frac{dN}{dt} \leq A_1 + A_2 - mN, \quad t \geq 0.$$

Now the standard comparison theorem [24] completes the proof. ∎

Let $\boldsymbol{\Phi}(t) : X \to X$ be the solution semiflow associated with (3.1); that is, $\boldsymbol{\Phi}(t)\boldsymbol{\phi} = \mathbf{u}_t(\boldsymbol{\phi})$, $\boldsymbol{\phi} \in X$, $t \geq 0$. By Lemma 3.1, solutions of (3.1) are uniformly bounded and ultimately bounded. Thus, the semiflow $\boldsymbol{\Phi}(t)$ is point dissipative on X, and $\boldsymbol{\Phi}(t) : X \to X$ is compact for each $t > \tau$. By [15, Theorem 3.4.8], it then follows that $\boldsymbol{\Phi}(t)$ admits a global attractor, which attracts every bounded set in X.

Under the assumption (H), it is easy to see that (3.1) has a unique disease-free equilibrium $E_0 = (S_1^*, S_2^*, 0, 0, 0, 0)$, where

$$S_1^* = -\frac{-H_2\,A_1 + \mu_2\,A_1 + d_2\,A_1 + d_2\,A_2}{-H_1\,H_2 + H_1\,\mu_2 + H_1\,d_2 + \mu_1\,H_2 - \mu_1\,\mu_2 - \mu_1\,d_2 + d_1\,H_2 - d_1\,\mu_2},$$
$$S_2^* = -\frac{d_1\,A_1 - H_1\,A_2 + \mu_1\,A_2 + d_1\,A_2}{-H_1\,H_2 + H_1\,\mu_2 + H_1\,d_2 + \mu_1\,H_2 - \mu_1\,\mu_2 - \mu_1\,d_2 + d_1\,H_2 - d_1\,\mu_2}.$$

To determine the basic reproduction number of (3.1), we assume that the population is at the disease-free equilibrium E_0. It then follows from (2.15) that

$$
\begin{aligned}
P_1(t) &= k_1 S_1^* \int_0^\tau b_{11}(a) P_1(t-a)da + k_2 S_2^* \int_0^\tau b_{12}(a) P_2(t-a)da,\\
P_2(t) &= k_1 S_1^* \int_0^\tau b_{21}(a) P_1(t-a)da + k_2 S_2^* \int_0^\tau b_{22}(a) P_2(t-a)da.
\end{aligned} \tag{3.3}
$$

Set

$$
\mathbf{U} = \begin{bmatrix} k_1 S_1^* \int_0^\tau b_{11}(a)da & k_2 S_2^* \int_0^\tau b_{12}(a)da \\ k_1 S_1^* \int_0^\tau b_{21}(a)da & k_2 S_2^* \int_0^\tau b_{22}(a)da \end{bmatrix}.
$$

Since $\mathbf{U}$ is a positive matrix, its spectral radius $\rho(\mathbf{U})$ is a simple eigenvalue with a positive eigenvector. Indeed, two eigenvalues of $\mathbf{U}$ are real, and $\rho(\mathbf{U})$ is the maximum of positive eigenvalues. Let $\boldsymbol{\psi}(a) = (\psi_1, \psi_2)^T$ be an initial distribution of infected members in the patches during the infection period, where ψ_1 and ψ_2 are constants. If we set

$$
\boldsymbol{\mathcal{F}} = \begin{bmatrix} k_1 S_1^* & 0 \\ 0 & k_2 S_2^* \end{bmatrix},
$$

then $\boldsymbol{\mathcal{F}}\boldsymbol{\psi}$ represents the emerging rate of new infectious individuals in the patches. From (2.6), we see that $b_{ij}(a)$ is the probability that an infective initially in patch j at infection age 0 is in patch i at infection age a. Then $\mathbf{U}\boldsymbol{\psi} = \int_0^\tau \exp(\mathbf{B}a)\boldsymbol{\mathcal{F}}\boldsymbol{\psi}da$ gives the members of infected individuals in the patches when the infection period ends. Motivated by [10, 26], we call $\mathbf{U}$ the next infection matrix and define $\rho(\mathbf{U})$ as the basic reproduction number R_0 of (3.1).

The following result shows that the disease is persistent if $R_0 > 1$.

THEOREM 3.2. *Let* (H) *hold. If $R_0 > 1$, then the disease is uniformly persistent in the sense that there is a positive number ϵ such that every solution $(\mathbf{S}(t), \mathbf{P}(t), \mathbf{R}(t))$ of* (3.1) *with $\mathbf{S}(0) \geq \mathbf{0}$, $\mathbf{P}(0) > \mathbf{0}$ (i.e., $\mathbf{P}(0) \geq \mathbf{0}$ and $\mathbf{P}(0) \neq \mathbf{0}$), and $\mathbf{R}(0) \geq \mathbf{0}$ satisfies $\liminf_{t\to\infty} P_i(t) \geq \epsilon$, $i = 1, 2$.*

Proof. As mentioned before, the solution semiflow $\boldsymbol{\Phi}(t)$ of (3.1) has a global attractor on X. In order to use persistence theory, we define

$$
X_0 := \{\boldsymbol{\phi} \in X : \phi_3(0) > 0, \phi_4(0) > 0\}, \qquad \partial X_0 := X \setminus X_0.
$$

In view of (2.15), it is easy to see that X_0 is positively invariant for $\boldsymbol{\Phi}(t)$. Clearly, ∂X_0 is relatively closed in X, and

$$
\partial X_0 = \{\boldsymbol{\phi} \in X : \phi_3(0) = 0 \text{ or } \phi_4(0) = 0\}.
$$

Let $L \in (L^*, \infty)$ be fixed. Then Lemma 3.1 implies that every solution of (3.1) enters $[0, L]^6$ ultimately. Define

$$
M_\partial := \{\boldsymbol{\phi} \in X : \boldsymbol{\Phi}(t)\boldsymbol{\phi} \in \partial X_0 \ \forall t \geq 0\}.
$$

Note that $b_1 > 0$ and $b_2 > 0$ imply that $\exp(\mathbf{B}a) > 0$ (see, e.g., [23]). Let $(\mathbf{S}(t, \boldsymbol{\phi}), \mathbf{P}(t, \boldsymbol{\phi}), \mathbf{R}(t, \boldsymbol{\phi}))$ be the solution of (3.1) satisfying $(\mathbf{S}_0(\boldsymbol{\phi}), \mathbf{P}_0(\boldsymbol{\phi}), \mathbf{R}_0(\boldsymbol{\phi})) = \boldsymbol{\phi}$. It then follows that

$$
M_\partial = \{\boldsymbol{\phi} \in \partial X_0 : \mathbf{P}(t, \boldsymbol{\phi}) = 0 \ \forall t \geq 0\}.
$$

Set

$$\mathbf{U}_\epsilon = \begin{bmatrix} k_1(S_1^*-\epsilon)\int_0^\tau b_{11}(a)da & k_2(S_2^*-\epsilon)\int_0^\tau b_{12}(a)da \\ k_1(S_1^*-\epsilon)\int_0^\tau b_{21}(a)da & k_2(S_2^*-\epsilon)\int_0^\tau b_{22}(a)da \end{bmatrix}.$$

Since the spectral radius of $\mathbf{U}_\epsilon$ is continuous in ϵ, we can restrict $\epsilon > 0$ small enough such that $\mathbf{U}_\epsilon$ is a positive matrix and $\rho(\mathbf{U}_\epsilon) > 1$. Let us consider the following linear system:

$$\begin{aligned} \frac{du_1(t)}{dt} &= A_1 - \xi + H_1 u_1(t) - (\mu_1 + d_1)u_1(t) + d_2 u_2(t), \\ \frac{du_2(t)}{dt} &= A_2 - \xi + H_2 u_2(t) - (\mu_2 + d_2)u_2(t) + d_1 u_1(t), \end{aligned} \tag{3.4}$$

where $\xi > 0$ is a small number. It is easy to see that (3.4) admits an equilibrium $(u_1^*(\xi), u_2^*(\xi))$, which is globally stable and satisfies $(u_1^*(\xi), u_2^*(\xi)) \to (S_1^*, S_2^*)$ as $\xi \to 0$. Thus, we can restrict $\xi > 0$ small enough such that every solution $(u_1(t), u_2(t))$ of (3.4) satisfies $u_i(t) > S_i^* - \epsilon$, $i = 1, 2$, for all large t.

Let $\delta = \min\{\frac{\xi}{k_1 L}, \frac{\xi}{k_2 L}\}$. We then have the following claim.

Claim. $\limsup_{t\to\infty} \max\{P_1(t, \boldsymbol{\phi}), P_2(t, \boldsymbol{\phi})\} \geq \delta$ for any $\boldsymbol{\phi} \in X_0$.

Assume, by contradiction, that the claim does not hold for some $\boldsymbol{\phi} \in X_0$. Then $P_i(t) := P_i(t, \boldsymbol{\phi}) < \delta$, $i = 1, 2$, for all large t. It follows that for all large t,

$$\begin{aligned} \frac{dS_1(t)}{dt} &> A_1 - \xi + H_1 S_1(t) - (\mu_1 + d_1)S_1(t) + d_2 S_2(t), \\ \frac{dS_2(t)}{dt} &> A_2 - \xi + H_2 S_2(t) - (\mu_2 + d_2)S_2(t) + d_1 S_1(t). \end{aligned} \tag{3.5}$$

By the comparison theorem of cooperative systems (see, e.g., [23, 24]), it follows that $S_i(t) > S_i^* - \epsilon$, $i = 1, 2$, for all large t. Thus, (2.8) implies that there is a $t_0 > 0$ such that for all $t \geq t_0$,

$$\begin{aligned} P_1(t) &> k_1(S_1^*-\epsilon)\int_0^\tau b_{11}(a)P_1(t-a)da + k_2(S_2^*-\epsilon)\int_0^\tau b_{12}(a)P_2(t-a)da, \\ P_2(t) &> k_1(S_1^*-\epsilon)\int_0^\tau b_{21}(a)P_1(t-a)da + k_2(S_2^*-\epsilon)\int_0^\tau b_{22}(a)P_2(t-a)da. \end{aligned} \tag{3.6}$$

Let $\mathbf{v} = (v_1, v_2)^T$ be a positive right eigenvector of $\mathbf{U}_\epsilon$ with respect to $\rho(\mathbf{U}_\epsilon)$. Choose $l > 0$ small enough such that $lv_i < \min\{P_i(t) : t_0 \leq t \leq t_0 + \tau\}$ for $i = 1, 2$. We further show that

$$lv_i < P_i(t), \quad i = 1, 2, \quad \forall t \geq t_0. \tag{3.7}$$

Otherwise, we set

$$t_1 = \inf\{t \in [t_0, \infty) : lv_1 = P_1(t) \text{ or } lv_2 = P_2(t)\}.$$

Clearly, $t_1 > t_0 + \tau$. It then follows that $lv_i < P_i(t)$, $i = 1, 2$, for $t_0 \leq t < t_1$, and $lv_1 = P_1(t_1)$ or $lv_2 = P_2(t_1)$. However, (3.6) implies that for all $t \in [t_0 + \tau, t_1]$,

$$\begin{aligned} P_1(t) &> k_1(S_1^*-\epsilon)lv_1\int_0^\tau b_{11}(a)da + k_2(S_2^*-\epsilon)lv_2\int_0^\tau b_{12}(a)da = \rho(\mathbf{U}_\epsilon)lv_1, \\ P_2(t) &> k_1(S_1^*-\epsilon)lv_1\int_0^\tau b_{21}(a)da + k_2(S_2^*-\epsilon)lv_2\int_0^\tau b_{22}(a)da = \rho(\mathbf{U}_\epsilon)lv_2, \end{aligned} \tag{3.8}$$

which contradicts $lv_1 = P_1(t_1)$ or $lv_2 = P_2(t_1)$. Thus, (3.7) holds.

Now suppose that for some $n \geq 1$,

$$\rho^{n-1}(\mathbf{U}_\epsilon)lv_i < P_i(t), \quad i = 1, 2, \quad \forall t \geq t_0 + (n-1)\tau. \tag{3.9}$$

We want to prove that

$$\rho^{n}(\mathbf{U}_\epsilon)lv_i < P_i(t), \quad i = 1, 2, \quad \forall t \geq t_0 + n\tau. \tag{3.10}$$

Note that (3.6) and (3.9) imply that $\rho^n(\mathbf{U}_\epsilon)lv_i < P_i(t_0 + n\tau)$, $i = 1, 2$. If (3.10) is not true, then there is a $t_2 > t_0 + n\tau$ such that $\rho^n(\mathbf{U}_\epsilon)lv_i < P_i(t)$, $i = 1, 2$, for $t_0 + n\tau \leq t < t_2$, and $\rho^n(\mathbf{U}_\epsilon)lv_1 = P_1(t_2)$ or $\rho^n(\mathbf{U}_\epsilon)lv_2 = P_2(t_2)$. By (3.6) and (3.9), it follows that for $t \in (t_0 + n\tau, t_2]$,

$$\begin{aligned}
P_1(t) &> k_1(S_1^* - \epsilon)\rho^{n-1}(\mathbf{U}_\epsilon)lv_1 \int_0^\tau b_{11}(a)da + k_2(S_2^* - \epsilon)\rho^{n-1}(\mathbf{U}_\epsilon)lv_2 \int_0^\tau b_{12}(a)da \\
&= \rho^n(\mathbf{U}_\epsilon)lv_1, \\
P_2(t) &> k_1(S_1^* - \epsilon)\rho^{n-1}(\mathbf{U}_\epsilon)lv_1 \int_0^\tau b_{21}(a)da + k_2(S_2^* - \epsilon)\rho^{n-1}(\mathbf{U}_\epsilon)lv_2 \int_0^\tau b_{22}(a)da \\
&= \rho^n(\mathbf{U}_\epsilon)lv_2,
\end{aligned}$$

which contradicts $\rho^n(\mathbf{U}_\epsilon)lv_1 = P_1(t_2)$ or $\rho^n(\mathbf{U}_\epsilon)lv_2 = P_2(t_2)$. By induction, we conclude that (3.10) holds for all $n \geq 0$. Since $\rho(\mathbf{U}_\epsilon) > 1$, we then obtain

$$\lim_{t\to\infty} P_i(t) \geq \lim_{n\to\infty} \rho^n(\mathbf{U}_\epsilon)lv_i = \infty,$$

a contradiction. This proves our claim.

Define $p : X \to R_+$ by

$$p(\boldsymbol{\phi}) = \min\{\phi_3(0), \phi_4(0)\} \quad \forall \boldsymbol{\phi} \in X.$$

Clearly, $X_0 = p^{-1}(0, \infty)$ and $\partial X_0 = p^{-1}(0)$. Note that p is a generalized distance function for the semiflow $\mathbf{\Phi}(t) : X \to X$ (see [25]). It is easy to see that any forward orbit of $\mathbf{\Phi}(t)$ in M_∂ converges to E_0. By the claim above, we see that E_0 is an isolated invariant set in X, and that $W^s(E_0) \cap X_0 = \emptyset$, where $W^s(E_0)$ is the stable manifold of E_0. By [25, Theorem 3.1], it then follows that there exists $\delta > 0$ such that $\min\{p(\boldsymbol{\psi}) : \boldsymbol{\psi} \in \omega(\boldsymbol{\phi})\} > \delta$ for any $\boldsymbol{\phi} \in X_0$. This implies our required uniform persistence of solutions of system (3.1). □

Next we show that a small invasion of the disease is unsuccessful if $R_0 < 1$.

THEOREM 3.3. *Let* (H) *hold. If* $R_0 < 1$*, then for every* $L \geq L^*$ *there exists a* $\zeta = \zeta(L) > 0$ *such that for any* $\boldsymbol{\phi} \in X_L$ *with* $(\phi_3(0), \phi_4(0)) \in [0, \zeta]^2$ *the solution* $(\mathbf{S}(t, \boldsymbol{\phi}), \mathbf{P}(t, \boldsymbol{\phi}), \mathbf{R}(t, \boldsymbol{\phi}))$ *of* (3.1) *converges to* E_0 *as* $t \to \infty$.

Proof. Let $L \geq L^*$ be given. By Lemma 3.1 and its proof, X_L is positively invariant for the solution semiflow of (3.1). We then have

$$(\mathbf{S}(t, \boldsymbol{\phi}), \mathbf{P}(t, \boldsymbol{\phi}), \mathbf{R}(t, \boldsymbol{\phi})) \in [0, L]^6 \quad \forall t \geq 0, \ \boldsymbol{\phi} \in X_L. \tag{3.11}$$

Set

$$\mathbf{V}_\epsilon = \begin{bmatrix} k_1(S_1^* + \epsilon)\int_0^\tau b_{11}(a)da & k_2(S_2^* + \epsilon)\int_0^\tau b_{12}(a)da \\ k_1(S_1^* + \epsilon)\int_0^\tau b_{21}(a)da & k_2(S_2^* + \epsilon)\int_0^\tau b_{22}(a)da \end{bmatrix}.$$

By the continuity of the spectral radius of $\mathbf{V}_\epsilon$ with respect to ϵ, we can restrict $\epsilon > 0$ small enough such that $\rho(\mathbf{V}_\epsilon) < 1$. Arguing as before, we can choose a small number $\xi_1 > 0$ and a large number $T_1 = T_1(L) > 0$ such that, for any solution $(u_1(t), u_2(t))$ of the system

$$\begin{aligned} \frac{du_1(t)}{dt} &= A_1 + \xi_1 + H_1 u_1(t) - (\mu_1 + d_1)u_1(t) + d_2 u_2(t), \\ \frac{du_2(t)}{dt} &= A_2 + \xi_1 + H_2 u_2(t) - (\mu_2 + d_2)u_2(t) + d_1 u_1(t) \end{aligned} \tag{3.12}$$

with $(u_1(0), u_2(0)) \in [0, L]^2$, we have $u_i(t) < S_i^* + \epsilon$, $i = 1, 2$, for all $t \geq T_1$. Similarly, we can select a small number $\xi_2 > 0$ and a large number $T_2 = T_2(L) > 0$ such that, for any solution $(w_1(t), w_2(t))$ of the system

$$\begin{aligned} \frac{dw_1(t)}{dt} &= \xi_2 - (\mu_1 + c_1)w_1(t) + c_2 w_2(t), \\ \frac{dw_2(t)}{dt} &= \xi_2 - (\mu_2 + c_2)w_2(t) + c_1 w_1(t) \end{aligned} \tag{3.13}$$

with $(w_1(0), w_1(0)) \in [0, L]^2$, we have $H_i w_i(t) < \xi_1/2$, $i = 1, 2$, for all $t \geq T_2$.

Let $\mathbf{v} = (v_1, v_2)^T$ be a positive right eigenvector of $\mathbf{V}_\epsilon$ associated with $\rho(\mathbf{V}_\epsilon)$. Choose $\xi_3 > 0$ small enough such that

$$\xi_3 \left(r_i v_i + b_{i1}(\tau)k_1 L v_1 + b_{i2}(\tau)k_2 L v_2\right) < \xi_2, \quad \xi_3 v_i H_i < \xi_1/2, \quad i = 1, 2. \tag{3.14}$$

Let $T_3 = T_3(L) := \max\{T_1, T_2\} + \tau$ and $\mathbf{W} := \operatorname{diag}(k_1 L, k_2 L)$. Then there exists $\zeta = \zeta(L) > 0$ such that for every solution $(P_1(t), P_2(t))$ of the linear system

$$\frac{d\mathbf{P}(t)}{dt} = (\mathbf{W} + \mathbf{B})\mathbf{P}(t), \quad t \geq 0, \tag{3.15}$$

with $(P_1(0), P_2(0)) \in [0, \zeta]^2$, we have $P_i(t) < \xi_3 v_i$, $i = 1, 2$, for all $t \in [0, 2T_3]$.

For a given $\boldsymbol{\phi} \in X_L$ with $(\phi_3(0), \phi_4(0)) \in [0, \zeta]^2$, we let $(\mathbf{S}(t), \mathbf{P}(t), \mathbf{R}(t)) = (\mathbf{S}(t, \boldsymbol{\phi}), \mathbf{P}(t, \boldsymbol{\phi}), \mathbf{R}(t, \boldsymbol{\phi}))$. By (3.1) and (3.11), we then have

$$\frac{d\mathbf{P}(t)}{dt} \leq (\mathbf{W} + \mathbf{B})\mathbf{P}(t) \quad \forall t \geq 0.$$

Since $\mathbf{P}(0) \in [0, \zeta]^2$, the comparison principle implies that

$$P_i(t) < \xi_3 v_i \quad \forall t \in [0, 2T_3], \quad i = 1, 2. \tag{3.16}$$

We further claim that (3.16) holds for all $t \geq 0$. If the claim is not true, then there exists a $T_4 = T_4(\boldsymbol{\phi}) > 2T_3$ such that $P_i(t) < \xi_3 v_i$ for $0 \leq t < T_4$, $i = 1, 2$, and $P_j(T_4) = \xi_3 v_j$ for $j = 1$ or $j = 2$. It follows from (3.1) and (3.14) that

$$\begin{aligned} \frac{dR_1(t)}{dt} &\leq \xi_2 - (\mu_1 + c_1)R_1(t) + c_2 R_2(t), \\ \frac{dR_2(t)}{dt} &\leq \xi_2 - (\mu_2 + c_2)R_2(t) + c_1 R_1(t) \end{aligned} \tag{3.17}$$

for $\tau \leq t \leq T_4$. By the comparison principle and the properties of system (3.13), we have $H_i R_i(t) < \xi_1/2$, $i = 1, 2$, for all $t \in [T_3, T_4]$. It follows from (3.1) that

$$\begin{aligned} \frac{dS_1(t)}{dt} &< A_1 + \xi_1 + H_1 S_1(t) - (\mu_1 + d_1)S_1(t) + d_2 S_2(t), \\ \frac{dS_2(t)}{dt} &< A_2 + \xi_1 + H_2 S_2(t) - (\mu_2 + d_2)S_2(t) + d_1 S_1(t) \end{aligned} \tag{3.18}$$

for all $t \in [T_3, T_4]$. By the comparison principle and the properties of system (3.12), we obtain

$$S_i(t) < S_i^* + \epsilon \quad \forall t \in [T_3 + T_1, T_4], \quad i = 1, 2.$$

Hence, (2.8) implies that for any $t \in [2T_3, T_4]$ we have

$$\begin{aligned} P_1(t) &< k_1(S_1^* + \epsilon) \int_0^\tau b_{11}(a) P_1(t-a)da + k_2(S_2^* + \epsilon) \int_0^\tau b_{12}(a) P_2(t-a)da, \\ P_2(t) &< k_1(S_1^* + \epsilon) \int_0^\tau b_{21}(a) P_1(t-a)da + k_2(S_2^* + \epsilon) \int_0^\tau b_{22}(a) P_2(t-a)da. \end{aligned} \tag{3.19}$$

It then follows that

$$\begin{aligned} P_1(t) &< k_1(S_1^* + \epsilon)\xi_3 v_1 \int_0^\tau b_{11}(a)da + k_2(S_2^* + \epsilon)\xi_3 v_2 \int_0^\tau b_{12}(a)da = \rho(\mathbf{V}_\epsilon)\xi_3 v_1, \\ P_2(t) &< k_1(S_1^* + \epsilon)\xi_3 v_1 \int_0^\tau b_{21}(a)da + k_2(S_2^* + \epsilon)\xi_3 v_2 \int_0^\tau b_{22}(a)da = \rho(\mathbf{V}_\epsilon)\xi_3 v_2 \end{aligned}$$

for all $t \in [2T_3, T_4]$. Since $\rho(\mathbf{V}_\epsilon) < 1$, we obtain $P_j(T_4) < \xi_3 v_j$ for $j = 1, 2$, which contradicts $P_j(T_4) = \xi_3 v_j$ for $j = 1$ or $j = 2$. This shows that $P_i(t) < \xi_3 v_i$, $i = 1, 2$, for all $t \geq 0$, and hence (3.19) holds for all $t \geq 2T_3$. By an induction argument similar to that in the proof of Theorem 3.2, it follows that

$$P_i(t) < \rho^n(\mathbf{V}_\epsilon)\xi_3 v_i \quad \forall t \geq 2T_3 + n\tau,\ n \geq 0,\ i = 1, 2,$$

which implies that $\lim_{t\to\infty} P_i(t) = 0$, $i = 1, 2$. By using the theory of chain transitive sets in [17] (see, e.g., the proof of [29, Theorem 2.2]), we further obtain that $(S_1(t), S_2(t), R_1(t), R_2(t)) \to (S_1^*, S_2^*, 0, 0)$ as $t \to \infty$. □

4. Examples. The objective of this section is to analyze the effect of population dispersal on the spread of the disease. Since $b_1 > 0$ and $b_2 > 0$ imply $\exp(\mathbf{B}a) > 0$, it follows that the matrix $\mathbf{U}$ is monotonically increasing with respect to τ. Thus, the basic reproduction number R_0, the spectral radius of the positive matrix $\mathbf{U}$, is an increasing function of the infection period τ. Therefore, the longer the infection period is, the more likely it is that the disease will spread, which is in accordance with our intuition.

In order to analyze the influences of the population dispersal, we begin with the case where two patches are isolated, i.e., $d_i = b_i = c_i = 0$, $i = 1, 2$. Then we have two decoupled subsystems:

$$\begin{aligned} &\frac{dS_1}{dt} = A_1 + H_1 N_1(t) - \mu_1 S_1(t) - Q_1(t), \\ &\frac{dP_1(t)}{dt} = Q_1(t) - e^{-(\mu_1 + r_1)\tau} Q_1(t-\tau) - (\mu_1 + r_1) P_1(t), \\ &\frac{dR_1}{dt} = r_1 P_1(t) + e^{-(\mu_1 + r_1)\tau} Q_1(t-\tau) - \mu_1 R_1(t), \\ &N_1(t) = S_1(t) + R_1(t) + P_1(t), \quad t \geq 0, \\ &P_1(0) = \int_{-\tau}^0 e^{(\mu_1 + r_1)s} Q_1(s)ds, \end{aligned} \tag{4.1}$$

and

$$
\begin{aligned}
&\frac{dS_2}{dt} = A_2 + H_2 N_2(t) - \mu_2 S_2(t) - Q_2(t),\\
&\frac{dP_2(t)}{dt} = Q_2(t) - e^{-(\mu_2+r_2)\tau} Q_2(t-\tau) - (\mu_2+r_2)P_2(t),\\
&\frac{dR_2}{dt} = r_2 P_2(t) + e^{-(\mu_2+r_2)\tau} Q_2(t-\tau) - \mu_2 R_2(t),\\
&N_2(t) = S_2(t) + R_2(t) + P_2(t), \quad t \geq 0,\\
&P_2(0) = \int_{-\tau}^{0} e^{(\mu_2+r_2)s} Q_2(s)ds.
\end{aligned} \tag{4.2}
$$

Note that system (4.1) has a disease-free equilibrium $E_{01} = (S_{01}^*, 0, 0)$ with $S_{01}^* = A_1/(\mu_1 - H_1)$, and that system (4.2) has a disease-free equilibrium $E_{02} = (S_{02}^*, 0, 0)$ with $S_{02}^* = A_2/(\mu_2 - H_2)$. Let R_{0i} be the basic reproduction number of the disease in patch i. By arguments similar to those for model (3.1), it follows that

$$R_{0i} = k_i S_{0i}^* \int_0^\tau e^{-(\mu_i+r_i)a} da = \frac{k_i A_i \left(1 - \exp(-(\mu_i + r_i)\tau)\right)}{(\mu_i - H_i)(\mu_i + r_i)},$$

and $R_{0i} > 1$ implies that the disease is uniformly persistent in the isolated patch i. The following result shows that the disease dies out in the isolated patch i if $R_{0i} < 1$.

PROPOSITION 4.1. *Let two patches be disconnected. Then the disease-free equilibrium E_{0i} is globally attractive if $R_{0i} < 1$.*

Proof. We consider only patch 1, since the proof for patch 2 is similar. As argued for general model (2.12), we have

$$P_1(t) = \int_{t-\tau}^{t} e^{-(\mu_1+r_1)(t-s)} Q_1(s)ds = \int_{-\tau}^{0} e^{(\mu_1+r_1)s} Q_1(t+s)ds \quad \forall t \geq 0. \tag{4.3}$$

It easily follows from (4.1) that

$$\frac{dN_1}{dt} = A_1 - (\mu_1 - H_1)N_1(t).$$

Thus,

$$N_1(t) \to A_1/(\mu_1 - H_1) = S_{01}^* \quad \text{as } t \to \infty. \tag{4.4}$$

Since $R_{01} < 1$, we can choose $\epsilon > 0$ small enough such that

$$R_{01}^\epsilon := \frac{k_1(S_{01}^* + \epsilon)\left(1 - \exp(-(\mu_1 + r_1)\tau)\right)}{\mu_1 + r_1} < 1. \tag{4.5}$$

By (4.4), we can choose $\bar{t} > 0$ large enough such that

$$S_1(t) < S_{01}^* + \epsilon \quad \text{for } t \geq \bar{t}. \tag{4.6}$$

It follows from (4.3) that

$$P_1(t) < k_1(S_{01}^* + \epsilon) \int_0^\tau e^{-(\mu_1+r_1)a} P_1(t-a)da \tag{4.7}$$

for $t \geq \bar{t} + \tau$. Fix a $\bar{v} > 0$ such that $P_1(t) < \bar{v}$ for $\bar{t} + \tau \leq t \leq \bar{t} + 2\tau$. By an induction argument similar to that in the proof of Theorem 3.2, it follows that

$$P_1(t) < (R_{01}^\epsilon)^n \bar{v} \quad \forall t \geq \bar{t} + (n+1)\tau,\ n \geq 0.$$

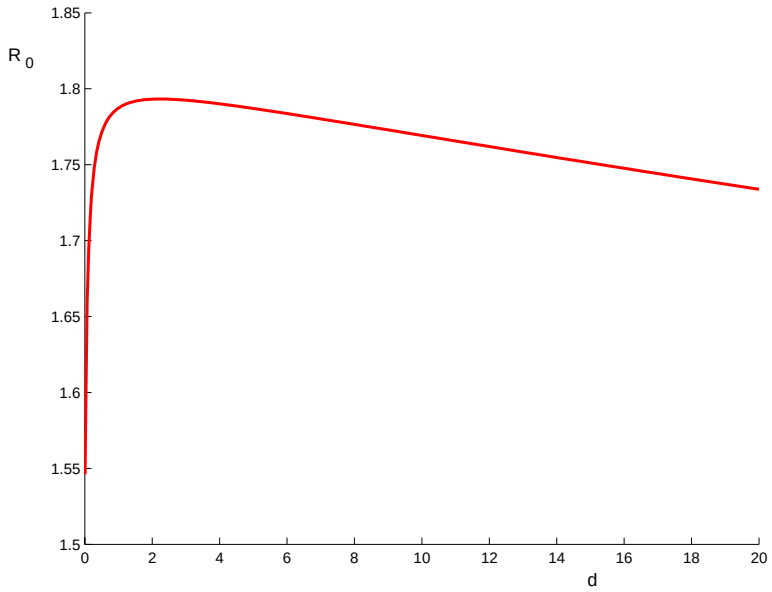


FIG. 1. *The graph of* R_0 *when* $\alpha = 0.5$.

Since $R_{01}^{\epsilon} < 1$, we have $P_1(t) \to 0$ as $t \to \infty$. By using the theory of chain transitive sets in [17] (see, e.g., the proof of [29, Theorem 2.2]), we further obtain that $(S_1(t), R_1(t)) \to (0, 0)$ as $t \to \infty$. □

When the patches are connected, we use numerical computations to find interesting dynamical behavior.

Example 4.1. We fix $A_1 = 1$, $A_2 = 0.6$, $\tau = 1$, $k_1 = 0.25$, $k_2 = 0.1$, $\mu_1 = \mu_2 = 0.2$, $H_1 = 0.05$, $H_2 = 0.1$, $r_1 = r_2 = 0$. Then we set $d_2 = d$ and take $d_1 = \alpha d$, $b_1 = 0.01\alpha d$, $b_2 = 0.01d$, where α represents the ratio of d_1 to d_2 and 0.01 means that 99 percent of infectious individuals in migration cannot pass through the borders of the two patches, due to control strategies, and only one percent can pass through, due to the failure of screening. Then $R_{01} = 1.5106$ and $R_{02} = 0.5438$. This means that if the two patches are isolated, the disease is persistent in the first patch and dies out in the second patch. If $\alpha = 0.5$, which means that individuals in the second patch have a higher migration rate than those in the first patch, from Figure 1 we see that $R_0 > \max\{R_{01}, R_{02}\}$. Thus, these kinds of mobilities facilitate disease propagation. When $\alpha = 1$, R_0 remains greater than 1 but has become a decreasing function of d. When $\alpha = 3$, from Figure 2 we see that the disease will die out when d is suitably large. Hence, the increase of mobility leads to the elimination of disease propagation. One interesting phenomenon occurs when α is larger. From Figure 3, which is typical for $\alpha \geq 4$, we see that R_0 goes through the stages of "above 1," "below 1," "above 1" as d increases, which means that the disease admits switches from persistence to extinction and then from extinction to persistence. These epidemiological changes result from the changes of population sizes in each patch under the influences of population dispersals. Indeed, if we regard S_i^* in the disease-free equilibrium as a function of d, we have $S_1^*(0) = 6.6667$ and $S_2^*(0) = 6$, which means that population sizes in the two patches are initially almost the same. For $\alpha < 0.9$, S_1^* is an increasing

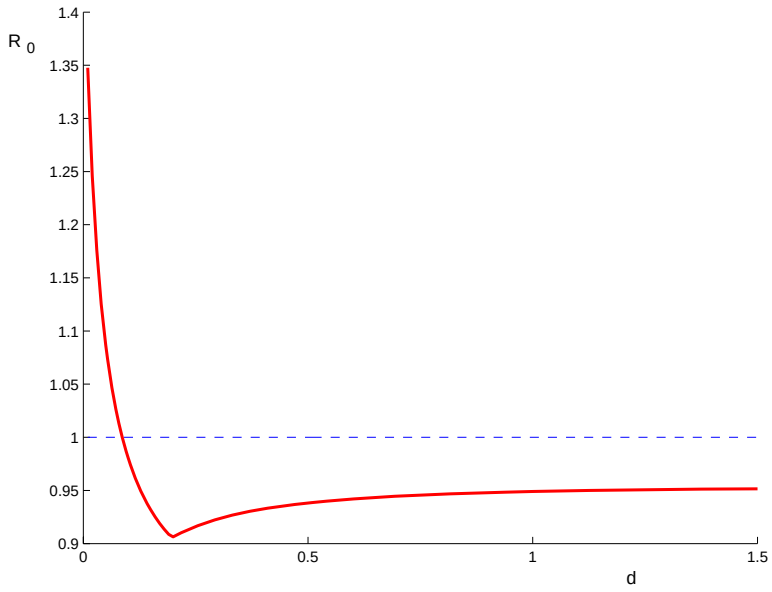


FIG. 2. *The graph of* R_0 *when* $\alpha = 3$.

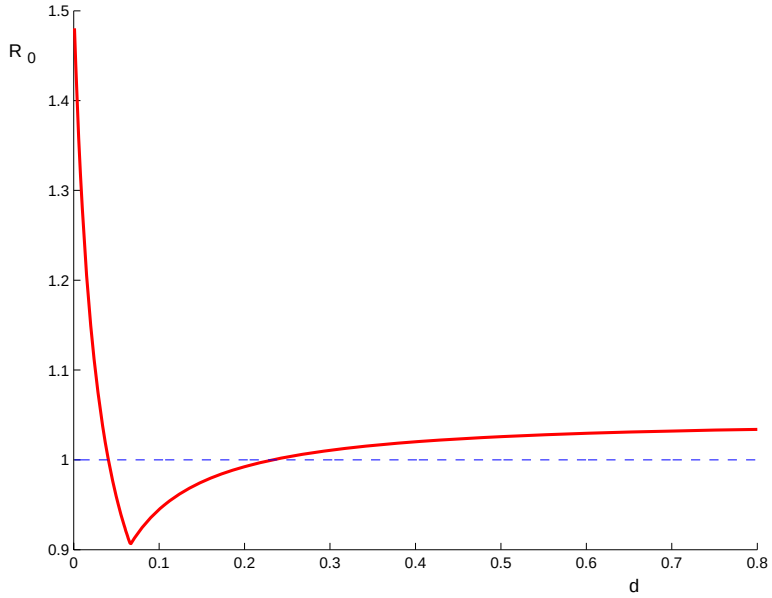


FIG. 3. *The graph of* R_0 *when* $\alpha = 4$.

function of d, and S_2^* is a decreasing function of d. For $\alpha > 0.9$, S_1^* is a decreasing function of d, and S_2^* is an increasing function of d. Thus, individuals aggregate towards the first patch when $\alpha < 0.9$ and aggregate towards the second patch when $\alpha > 0.9$. Note that $k_2 < k_1$ implies that the second patch is better than the first patch from the epidemiological perspective. Thus, it is easy to understand why the basic reproduction number for $d > 0$ is higher than that for $d = 0$ when $\alpha = 0.5$. For

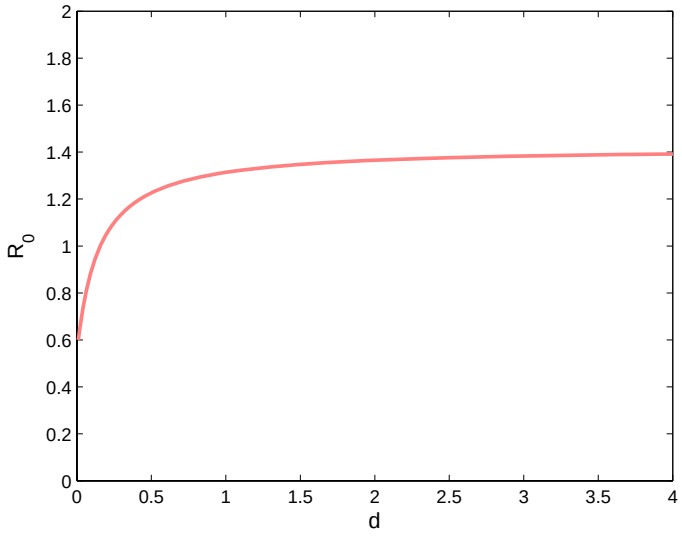


FIG. 4. $R_0 > 1$ *when* $d > 0.1574$.

the cases where $\alpha > 0.9$, the basic reproduction numbers tend to go down because the net spatial flow of population mobility is directed toward the better patch. Note that the basic reproduction number when $\alpha = 3$ is lower than 1 when the dispersal rate d is greater than a critical value. However, if $\alpha \geq 4$, the basic reproduction number fluctuates about the threshold value 1 for medium d. This cannot be interpreted in the manner above, and we should consider the collective effect of population sizes of two patches and the immigration of infectives. Hence, some unexpected effects of spatial disease transmission can be produced by population mobility.

It should be noted that there are distinctions in the demographies of the populations in two patches in the last example. However, we can find similar behavior even when the demographies of the populations in two patches are the same. Such an example can be obtained if we fix $A_1 = A_2 = 0.8$, $H_1 = H_2 = 0.9$, $\tau = 1$, $\mu_1 = \mu_2 = 0.2$, $r_1 = r_2 = 0$, $d_2 = d$ and $d_1 = \alpha d$, $b_1 = 0.01\alpha d$, $b_2 = 0.01d$.

Example 4.2. We fix $A_1 = 1$, $A_2 = 0.6$, $\tau = 1$, $k_1 = 0.1$, $k_2 = 0.1$, $\mu_1 = \mu_2 = 0.2$, $H_1 = 0.05$, $H_2 = 0.1$, $r_1 = 0$, $r_2 = 0$ and $b_1 = 0.01d$, $b_2 = 0.002d$, $d_1 = d$, $d_2 = 0.01d$. These values are different from those in the last example only in k_1 and migration rates. Now, $R_{01} = 0.6432$ and $R_{02} = 0.5438$. This means that if the two patches are isolated, the disease dies out in the two patches. When we have the migration rates, we see that the disease spreads in the two patches when $d > 0.1574$ (see Figure 4). Hence, the mobility of individuals in this case facilitates the spread of the disease. Note that the disease transmission coefficients of two patches are the same. We need to check the demography of two patches to reveal the mechanism of epidemiological changes. When the patches are disconnected, the condition for the prevalence of the disease in the ith patch is $S_i^* > 11.0333$. When $d > 0$, we see that S_1^* is a decreasing function of d, and S_2^* is an increasing function of d. Thus, when d increases, individuals aggregate towards the second patch, and S_2^* reaches 11.0333 at $d = 0.1573$. However, the basic reproduction number R_0 of the coupled system remains less than 1 for $0.1573 < d < 0.1574$. This seems to be related to the spatial heterogeneity. When

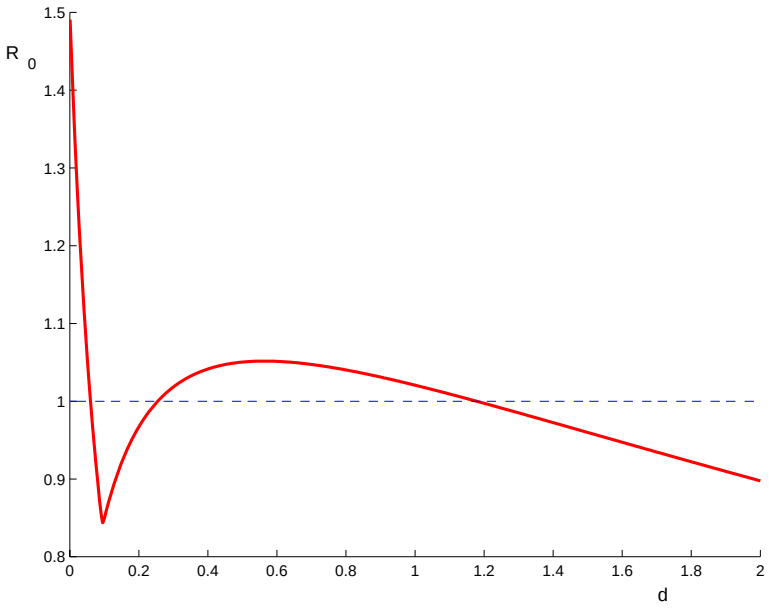


FIG. 5. *The graph of R_0 where there are multiple exchanges between persistence and extinction of the disease.*

$d > 0.1574$, we have $R_0 > 1$, which is a consequence of the aggregation of individuals in the second patch.

Example 4.3. We choose $A_1 = 1$, $A_2 = 0.6$, $\tau = 1$, $k_1 = 0.2$, $k_2 = 0.1$, $\mu_1 = \mu_2 = 0.21$, $H_1 = 0.09$, $H_2 = 0.1$ and $b_1 = 0.06d$, $b_2 = 0.4d$, $d_1 = d$, $d_2 = 0$. Numerical calculations show that R_0 crosses the critical value 1 three times at $d = 0.0601, 0.2546, 1.1784$ (see Figure 5). Hence, the disease admits three exchanges between persistence and extinction.

5. Discussions. In this paper, we have proposed an epidemic model to simulate the dynamics of disease transmission under the influence of a population dispersal among patches. The population dispersal among patches can be interpreted as the movements by which people travel or migrate from one city to another city or from one country to another country. We have incorporated a constant infection period into the model. Under the assumptions that the death rates, the disease transmission coefficients, the treatment rates, and the migration rates for infected individuals are constants, by using characteristic methods, we reduce the model to a time-delayed differential system. For this model, we have established a formula to compute the basic reproduction number, which extends the method in [26] of computing the basic reproduction number for multiple compartments governed by ordinary differential equations.

We have found that the longer the infection period is, the more likely it becomes that the disease will spread. We have shown that the disease cannot invade the population distributed in n patches if $R_0 < 1$ and invasion intensity is not strong, and the disease is uniformly persistent when $R_0 > 1$. If the disease spreads in the first isolated patch and dies out in the second isolated patch, by numerical calculations we have shown, in Example 4.1, that an increase of the ratio α between migration rates may lead to an increase of the basic reproduction number when α is small, lead

to the extinction of the disease in the two patches when α is larger, and lead to two switches between the persistence of the disease and extinction of the disease. Another example further shows that multiple switches between the persistence of the disease and extinction of the disease are possible. If the disease dies out in each isolated patch, we have shown that suitable migration rates may result in the outbreak of the disease in both patches.

We have assumed that the death rates, the infection force, the treatment rates, and the migration rates for infected individuals are independent of infection age. It is interesting to study the case where they are age-dependent. We leave this as future work.

Acknowledgments. We are very grateful to two anonymous referees for their careful reading and valuable comments which led to improvements of our manuscript.

REFERENCES

[1] V. Andreasen, *Disease regulation of age-structured host populations*, Theoret. Popul. Biol., 36 (1989), pp. 214–239.

[2] J. Arino and P. van den Driessche, *A multi-city epidemic model*, Math. Popul. Stud., 10 (2003), pp. 175–193.

[3] J. Arino and P. van den Driessche, *The basic reproduction number in a multi-city compartmental epidemic model*, in Positive Systems (Rome, 2003), Lecture Notes in Control and Inform. Sci. 294, Springer, Berlin, 2003, pp. 135–142.

[4] F. Brauer and P. van den Driessche, *Models for transmission of disease with immigration of infectives*, Math. Biosci., 171 (2001), pp. 143–154.

[5] M. S. Bartlett, *Measles periodicity and community size*, J. Roy. Statist. Soc. Ser. A, 120 (1957), pp. 48–70.

[6] B. Bolker and B. Grenfell, *Space, persistence and dynamics of measles epidemics*, Philos. Trans. Roy. Soc. London B, 348 (1995), pp. 309–320.

[7] C. Castillo-Chavez and A. A. Yakubu, *Dispersal, disease and life-history evolution*, Math. Biosci., 173 (2001), pp. 35–53.

[8] K. Cooke and P. van den Driessche, *Analysis of an SEIRS epidemic model with two delays*, J. Math. Biol., 35 (1996), pp. 240–260.

[9] K. Cooke, P. van den Driessche, and X. Zou, *Interaction of maturation delay and nonlinear birth in population and epidemic models*, J. Math. Biol., 39 (1999), pp. 332–352.

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

[11] Z. Feng and H. R. Thieme, *Endemic models with arbitrarily distributed periods of infection* I: *Fundamental properties of the model*, SIAM J. Appl. Math., 61 (2000), pp. 803–833.

[12] E. Fromont, D. Pontiera, and M. Langlais, *Disease propagation in connected host populations with density-dependent dynamics: The case of the feline leukemia virus*, J. Theoret. Biol., 223 (2003), pp. 465–475.

[13] G. R. Fulford, M. G. Roberts, and J. A. P. Heesterbeek, *The metapopulation dynamics of an infectious disease: Tuberculosis in possums*, Theoret. Popul. Biol., 61 (2002), pp. 15–29.

[14] J. K. Hale and S. M. V. Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.

[15] J. K. Hale, *Asymptotic Behavior of Dissipative Systems*, Math. Surveys Monogr. 25, AMS, Providence, RI, 1988.

[16] H. W. Hethcote and P. van den Driessche, *Two SIS epidemiologic models with delays*, J. Math. Biol., 40 (2000), pp. 3–26.

[17] W. M. Hirsch, H. L. Smith, and X.-Q. Zhao, *Chain transitivity, attractivity, and strong repellors for semidynamical systems*, J. Dynam. Differential Equations, 13 (2001), pp. 107–131.

[18] Y. Kuang, *Delay Differential Equations with Applications in Population Dynamics*, Math. Sci. Engrg. 191, Academic Press, New York, 1993.

[19] A. L. Lloyd, *Realistic distributions of infectious periods in epidemic models: Changing patterns of persistence and dynamics*, Theoret. Popul. Biol., 60 (2001), pp. 59–71.

[20] D. J. Rodríguez and L. Torres-Sorando, *Models of infectious diseases in spatially heterogenous environments*, Bull. Math. Biol., 63 (2001), pp. 547–571.
[21] S. Ruan, W. Wang, and S. A. Levin, *The effect of global travel on the spread of SARS*, Math. Biosci. Eng., 3 (2006), pp. 205–218.
[22] L. Sattenspiel and D. A. Herring, *Simulating the effect of quarantine on the spread of the* 1918-19 *flu in central Canada*, Bull. Math. Biol., 65 (2003), pp. 1–26.
[23] H. L. Smith, *Monotone Dynamical Systems. An Introduction to the Theory of Competitive and Cooperative Systems*, Math. Surveys Monogr. 41, AMS, Providence, RI, 1995.
[24] H. L. Smith and P. Waltman, *The Theory of the Chemostat*, Cambridge University Press, Cambridge, UK, 1995.
[25] H. L. Smith and X.-Q. Zhao, *Robust persistence for semidynamical systems*, Nonlinear Anal., 47 (2001), pp. 6169–6179.
[26] P. van den Driessche and J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci., 180 (2002), pp. 29–48.
[27] W. Wang and S. Ruan, *Simulating the SARS outbreak in Beijing with limited data*, J. Theoret. Biol., 227 (2004), pp. 369–379.
[28] W. Wang and G. Mulone, *Threshold of disease transmission on a patch environment*, J. Math. Anal. Appl., 285 (2003), pp. 321–335.
[29] W. Wang and X.-Q. Zhao, *An epidemic model in a patchy environment*, Math. Biosci., 190 (2004), pp. 39–69.
[30] W. Wang and X.-Q. Zhao, *An age-structured epidemic model in a patchy environment*, SIAM J. Appl. Math., 65 (2005), pp. 1597–1614.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1473–1481

TRAVELING FRONTS IN PRESSURE-DRIVEN COMBUSTION*

FATHI DKHIL† AND K. P. HADELER‡

Abstract. Brailovsky and Sivashinsky have proposed a model for pressure-driven combustion in the form of a degenerate parabolic system for temperature, concentration, and pressure. It is shown that the existence and uniqueness problem for traveling front solutions can be completely solved by exploiting the existing invariants and by phase plane methods. The approach yields exact propagation speeds which are noticeably larger than the approximations obtained so far.

Key words. combustion, deflagration, traveling front, reaction diffusion

AMS subject classifications. 35K55, 35K57, 37C29, 76L05, 80A25

DOI. 10.1137/040611148

1. Introduction. Brailovsky and Sivashinsky [2], [3] and Brailovsky, Frankel, and Sivashinsky [4] have proposed a model for pressure-driven subsonic combustion in a porous medium (deflagration as opposed to detonation). The model neglects inertia effects, and it does not produce shocks but traveling fronts; see Figure 1.1. This model and some modifications have been studied in a series of subsequent papers [5], [12], [13], [11], [7]; see also [6]. The analysis of the model is centered around the existence and uniqueness of traveling front solutions and their stability. Brézis, Kamin, and Sivashinsky [5] impose an additional condition derived from the behavior at the leading edge of the front and thus obtain a simplified system in the form of a degenerate diffusion system for temperature and pressure alone. Then they reduce the traveling

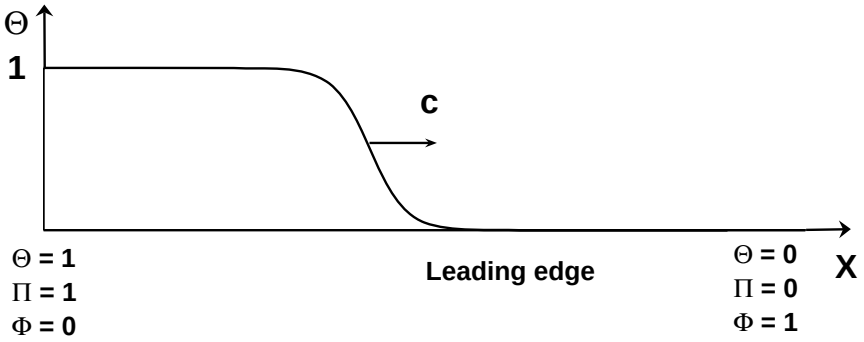


FIG. 1.1. *Deflagration front. Before the leading edge (right): temperature* $\Theta = 0$*, pressure* $\Pi = 0$*, and concentration* $\Phi = 1$*. Behind the front (left): temperature* $\Theta = 1$*, pressure* $\Pi = 1$*, concentration* $\Phi = 0$*. All quantities are given in nondimensionalized units.*

*Received by the editors July 6, 2004; accepted for publication (in revised form) January 31, 2006; published electronically May 19, 2006.

http://www.siam.org/journals/siap/66-5/61114.html

†Max-Planck-Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany. Current address: Département de Mathématiques, Institut Supérieur d'Informatique (ISI), Université de Tunis El Manar, 2 rue Abou Raihane Bairouni, 2080 Ariana, Tunisia (fathi.dkhil@isi.rnu.tn).

‡Mathematics, University of Tübingen, Auf der Morgenstelle 10, D-72076 Tübingen, Germany (hadeler@uni-tuebingen.de) and Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287.

front problem to the discussion of a nonautonomous scalar differential equation. In the present paper we find a conservation law for the system of partial differential equations. In the four-dimensional system of ordinary differential equations for the shapes and speeds of traveling fronts we can use this conservation law and a second invariant to reduce the problem to a system in dimension two, which can be treated by phase plane methods.

In section 2 we define the full problem and the simplified problem and exhibit a conservation law. Section 3 consists of two parts, reduction of the four-dimensional traveling front problem to a two-dimensional problem (which works for general nonlinearities) and solution of the specific combustion problem. In section 4 we discuss the differences between the two models in quantitative terms (with proofs deferred to an appendix). In section 5 we present some results on a limiting case (gas near ignition temperature everywhere) without giving the proofs in detail. Finally, in a conclusion, section 6, we sketch the essentials of the analytic approach and mark the point from which one must use numerics.

2. The combustion problem. The original problem [2] reads

$$\begin{aligned} \gamma\Theta_t - (\gamma-1)\Pi_t &= \Omega(\Phi,\Theta), \\ \Phi_t &= -\Omega(\Phi,\Theta), \\ \Pi_t - \Theta_t &= \Delta\Pi, \end{aligned} \tag{2.1}$$

with

$$\Omega(\Phi,\Theta) = \Phi g(\Theta), \tag{2.2}$$

where $g(\Theta)$ is a nondecreasing function and $\gamma > 1$.

The variables Θ, Φ, and Π correspond, after suitable normalization, to the temperature $\Theta \in [0,1]$, the concentration of the deficient reactant $\Phi \in [0,1]$, and the pressure $\Pi \in [0,1]$. The function g describes the reaction rate as a function of temperature, and the constant γ is the specific heat ratio.

The function g is continuous and piecewise differentiable. There is some $\beta \in (0,1)$ such that $g(\Theta) = 0$ for $0 \le \Theta \le \beta$ and $g(\Theta) > 0$ for $\beta < \Theta \le 1$.

The number β is the ignition temperature. There are various stationary states of the gas with temperature below ignition and any pressure and concentration. In [2] it is assumed that prior to combustion (at the leading edge of the front; see Figure 1.1) the state is

$$\Theta = 0, \quad \Phi = 1, \quad \Pi = 0 \tag{2.3}$$

and that

$$\Theta = 1, \quad \Phi = 0, \quad \Pi = 1 \tag{2.4}$$

after the combustion process has been completed. As a deflagration front passes through, the state of the gas passes from (2.3) to (2.4). Hence a traveling front (traveling from left to right) is a wave solution of the system (2.1) which satisfies (2.3) at $x = +\infty$ and (2.4) at $x = -\infty$.

By linearly combining the equations, the system (2.1) can be written in the form of a standard dynamical system with only one time derivative in each equation:

$$\begin{aligned} \Pi_t &= \gamma\Delta\Pi + \Phi g(\Theta), \\ \Theta_t &= (\gamma-1)\Delta\Pi + \Phi g(\Theta), \\ \Phi_t &= -\Phi g(\Theta). \end{aligned} \tag{2.5}$$

Hence it is obvious that the model (2.1) is a degenerate parabolic system.

The approach in [5] is based on the following idea. If the explosion were homogeneous, then the spatial derivatives would vanish. Then, whatever the reaction term $\Omega(\Phi, \Theta)$ is, $\Theta + \Phi$ and $\Pi + \Phi$ are constants. Using the initial conditions (2.3), one finds

$$\Phi = 1 - \Theta. \tag{2.6}$$

If this equation is used to eliminate Φ from (2.5), then one arrives at

$$\begin{aligned} \Pi_t &= \gamma \Delta \Pi + (1 - \Theta) g(\Theta), \\ \Theta_t &= (\gamma - 1) \Delta \Pi + (1 - \Theta) g(\Theta). \end{aligned} \tag{2.7}$$

The authors of [5] discuss the problem of traveling fronts for the system (2.7), subject to the boundary condition $\Pi = \Theta = 0$ at the leading edge and $\Pi = \Theta = 1$ after the front has passed. But we observe that the system (2.5) has a conservation law, and hence we can improve on [5].

PROPOSITION 2.1. *For a given solution of the system* (2.5) *the function*

$$\Phi + \gamma \Theta - (\gamma - 1) \Pi \tag{2.8}$$

does not depend on the variable t.

Proof. In (2.5) check the equation $\Phi_t + \gamma \Theta_t - (\gamma - 1) \Pi_t \equiv 0$. □

The physical interpretation of (2.8) is evident. It says that, for a given solution, at a given space point, the expression (2.8) does not change in time, or, in other words, the pressure is determined by the remaining reactant and the temperature. Equation (2.8) could be used to eliminate the variable Φ from (2.5) and to obtain a system for the two variables Π and Θ which, however, has coefficients that explicitly depend on the space variable.

Of course (2.6) is based on a rather crude assumption since, away from the leading edge, the front is far from equilibrium. It is known that for certain diffusive traveling front problems (Fisher–KPP with concave nonlinearity) the speed is determined by the linearization at the leading edge (see [1], [8], [9], [10], [14]), but this is not so for combustion problems in general. For these reasons we treat the full problem here, whereby we can use (2.8).

3. Traveling fronts for the full problem. We look for traveling fronts of the problem (2.5) and for the simplified problem (2.7). We show the following result.

THEOREM 3.1. *If* $\gamma(1 - \beta) \leq 1$, *then there are no traveling fronts for the full system nor for the simplified system since the specific heat ratio is too small or the ignition temperature is too high. If* $\gamma(1 - \beta) > 1$, *then the full system has a traveling front solution with speed* $c(1)$, *and the simplified system has a traveling front solution with speed* $c(0)$. *In either case the front is unique up to translation. Furthermore,* $c(0) < c(1)$.

3.1. Reduction of the problem. For the proof we make a traveling wave ansatz in (2.5): $\Pi = \Pi(x - ct)$, $\Theta = \Theta(x - ct)$, $\Phi = \Phi(x - ct)$ with $c > 0$. We get a system of order four (counting the derivatives):

$$-c\Pi' = \gamma \Pi'' + \Phi g(\Theta), \tag{3.1}$$

$$-c\Theta' = (\gamma - 1)\Pi'' + \Phi g(\Theta), \tag{3.2}$$

$$-c\Phi' = -\Phi g(\Theta). \tag{3.3}$$

The $'$ denotes the derivative with respect to the traveling wave coordinate $x - ct$. We look for solutions of this system which satisfy the boundary conditions (2.3) at $+\infty$ and (2.4) at $-\infty$. We recover (2.8) as an invariant of motion and find another invariant.

PROPOSITION 3.2. *The system* (3.1), (3.2), *and* (3.3) *has one invariant of motion independent of the parameter c,*

$$\Phi + \gamma\Theta - (\gamma - 1)\Pi = \text{const}, \tag{3.4}$$

and another invariant of motion depending on the parameter c,

$$\Pi' + c(\Pi - \Theta) = \text{const}. \tag{3.5}$$

Proof. We subtract (3.2) from (3.1), get $-c\Pi' + c\Theta' = \Pi''$, integrate, and obtain (3.5). We add (3.1) and (3.3), get $-c\Pi' - c\Phi' = \gamma\Pi''$, integrate, and get

$$\gamma\Pi' + c\Pi + c\Phi = \text{const}. \tag{3.6}$$

Combining (3.5) and (3.6) yields (3.4). □

Using the condition (2.3) at the leading edge, we get from (3.4) and (3.5) for traveling fronts the equations

$$\Pi' = c\theta - c\Pi \tag{3.7}$$

and

$$\Phi = 1 - \Theta + (\gamma - 1)(\Pi - \Theta), \tag{3.8}$$

from which the difference from (2.6) is obvious. Now we use (3.7), (3.2), and (3.8) to get a system for the variables Π and Θ,

$$\begin{aligned} \Pi' &= c\Theta - c\Pi, \\ \Theta' &= c\frac{\gamma - 1}{\gamma}(\Theta - \Pi) - \frac{1 - \Theta + (\gamma - 1)(\Pi - \Theta)}{c\gamma} g(\Theta). \end{aligned} \tag{3.9}$$

We have used two invariants of motion, and hence this system is of order two only.

On the other hand, for front solutions of the simplified system (2.7) we get another (three-dimensional) system for the variables Π and Θ,

$$\begin{aligned} -c\Pi' &= \gamma\Pi'' + (1 - \Theta)g(\Theta), \\ -c\Theta' &= (\gamma - 1)\Pi'' + (1 - \Theta)g(\Theta). \end{aligned} \tag{3.10}$$

Integration as before (using only one invariant) leads to a two-dimensional system

$$\begin{aligned} \Phi' &= c\Theta - c\Pi, \\ \Theta' &= c\frac{\gamma - 1}{\gamma}(\Theta - \Pi) - \frac{1 - \Theta}{c\gamma} g(\Theta). \end{aligned} \tag{3.11}$$

Now we are ready to compare the systems (3.9) and (3.11). We scale the independent variable by a factor c, and we introduce a parameter $\epsilon \in [0, 1]$ such that we can write both systems in the same form

$$\begin{aligned} \Pi' &= \Theta - c\Pi, \\ \Theta' &= \frac{\gamma - 1}{\gamma}(\Theta - \Pi) - \frac{1 - \Theta + \epsilon(\gamma - 1)(\Pi - \Theta)}{c^2\gamma} g(\Theta), \end{aligned} \tag{3.12}$$

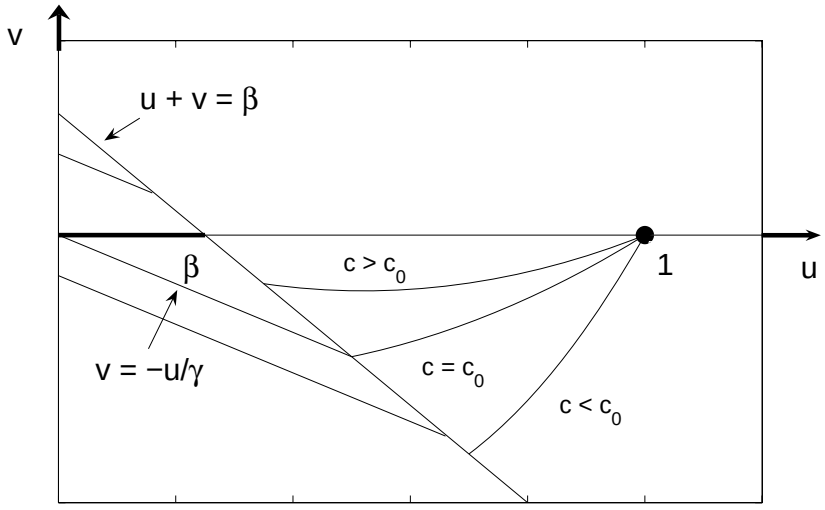


FIG. 3.1. *Schematic phase plane of the u, v-system for several values of c. There is an interval of stationary points between the origin and $(\beta, 0)$ and an isolated stationary point $(1, 0)$. For $c = c_0$ the unstable manifold of the saddle point $(1, 0)$ meets the straight line through the origin, which is the stable manifold of $(0, 0)$.*

with $\epsilon = 1$ for the full system (2.1) and $\epsilon = 0$ for the simplified system (2.7).

The system (3.12) can be further transformed by introducing new dependent variables (a similar substitution has been used in [5]):

$$u = \Pi, \qquad v = \Theta - \Pi. \tag{3.13}$$

The new system is

$$\begin{aligned} u' &= v, \\ v' &= -\frac{1}{\gamma} v - \frac{1}{c^2 \gamma}(1 - u - (1 + \epsilon(\gamma - 1))v) g(u + v). \end{aligned} \tag{3.14}$$

Now the boundary conditions read

$$\begin{aligned} u(+\infty) &= 0, \qquad v(+\infty) = 0, \\ u(-\infty) &= 1, \qquad v(-\infty) = 0. \end{aligned}$$

3.2. Qualitative analysis. We enter the qualitative analysis of the system (3.14). The phase plane is shown in Figure 3.1. There is an interval of stationary points $(u, 0)$ with $u \le \beta$ and an isolated stationary point $(1, 0)$, which is a saddle point. We are looking for a heteroclinic orbit, i.e., for a trajectory, for some value of c, which leaves the point $(1, 0)$ with the unstable manifold and ends at $(0, 0)$.

In the domain $u + v \le \beta$ the system is linear, $u' = v$, $v' = -v/\gamma$, and hence $-dv/du = 1/\gamma < 1$. Hence in this domain any trajectory follows a straight line which ends at one of the stationary points with $u \le \beta$. In particular, the trajectory ending at $(0, 0)$ lies on the straight line $v = -u/\gamma$. If we follow this line (backward in time), then it meets the line $u + v = \beta$ at the point

$$P_0 = (u_0, v_0) = (\beta\gamma/(\gamma - 1), -\beta/(\gamma - 1)). \tag{3.15}$$

The point $(1,0)$ is a saddle point. Consider the (negative part of the) unstable manifold of the saddle point. For any given $c>0$ it arrives at some point of the line $u+v=\beta$. Along the unstable manifold we have $(1-u-(1+\epsilon(\gamma-1))v)g(u+v)>0$ in view of $v<0$. Hence the factor of $-1/c^2\gamma$ in (3.14) is positive. Using this fact, one can show that the unstable manifolds for different c do not intersect (e.g., by the method of positively invariant sets; see [8]). Hence the unstable manifold depends on c in a monotone way. For very large c it arrives at the line $u+v=\beta$ at a point with v small. For very small positive c the unstable manifold arrives at that line near $u=1$, $v=-1/\gamma$. If $u_0\geq 1$, i.e., if $\gamma(1-\beta)\leq 1$, then there cannot be a heteroclinic connection. Next assume

$$\gamma(1-\beta)>1. \tag{3.16}$$

If c runs from 0 to $+\infty$, then there is exactly one value $c_0>0$ where the unstable manifold meets the point P_0 and thus forms a heteroclinic connection to $(0,0)$. Hence we have shown the following proposition.

PROPOSITION 3.3. *Suppose that* (3.16) *holds. For any choice of* $\epsilon\in[0,1]$ *there is exactly one value* $c=c(\epsilon)$ *such that the system* (3.14) *has a heteroclinic orbit connecting* $(1,0)$ *to* $(0,0)$.

If ϵ is increased from 0 to 1, then in (3.14), for given u, v, and c, the quantity v' is increasing, and hence the trajectory is moved upward; we have to increase c to move it down again to meet the point P_0. This argument shows that $c(\epsilon)$ is an increasing function. This argument can be worked out in detail using positively invariant sets. Hence Theorem 3.1 has been proved. □

4. Quantitative comparison. We have shown for the full system as well as for the simplified system the existence and uniqueness of the desired front solutions. The speeds $c(0)<c(1)$ are clearly different. The question arises whether the difference is small in some sense and whether $c(0)$ can be seen as a useful approximation. We just mention that in the limiting case $\beta=0$ (see below) the full problem and the simplified problem yield the same speed because that problem is linearly determined. However, for $\beta>0$ the two speeds are clearly different. To have something concrete at hand, we consider a caricature where at ignition temperature β the function $g(\Theta)$ jumps to some positive value κ and then stays constant,

$$g(\Theta)=\begin{cases}0, & 0\leq\Theta\leq\beta,\\ \kappa, & \beta<\Theta\leq 1.\end{cases} \tag{4.1}$$

Then the equations (3.14) become piecewise linear and can be solved explicitly. We get an explicit formula for $c(\epsilon)$,

$$c^2(\epsilon)=\kappa\frac{1+(1+\epsilon(\gamma-1))q}{q(1+\gamma q)},\quad \text{where}\quad q=\frac{\beta}{\gamma-1-\gamma\beta}. \tag{4.2}$$

Hence, as stated before, by choosing $\epsilon=0$ instead of $\epsilon=1$, the speed in underestimated. The effect becomes evident in the limit of large γ, where

$$c^2(\epsilon)\approx\kappa\gamma(1-\beta)\left(\frac{1-\beta}{\beta}+\epsilon\right). \tag{4.3}$$

5. The limiting case. The limiting case $\beta=0$ of the combustion problem can be interpreted as a situation of a traveling front moving into a gas which is at ignition

temperature but not yet burning. Then we have to be more specific about the shape of the nonlinearity. We assume

$$g(0) = 0, \quad g'(\Theta) > 0 \quad \text{for} \quad \Theta \geq 0. \tag{5.1}$$

Once the problem has been reduced to the system (3.14) for the variables u and v, the subsequent qualitative analysis, though complicated in the details, follows essentially the approach to the Fisher–KPP problem (see [8], [9]).

Linearization at the leading edge of the front yields two critical values for the speed c,

$$c_\pm = \sqrt{g'(0)}(\sqrt{\gamma} \pm \sqrt{\gamma - 1}). \tag{5.2}$$

PROPOSITION 5.1. *There is a number $c_0 \geq c_+$ such that for every $c \geq c_0$ the unstable manifold of $(1,0)$ connects to $(0,0)$ in such a way that the corresponding front is monotone. There are no monotone fronts for $c < c_0$.*

The speeds of monotone fronts form a half-line $[c_0, \infty)$. There are no such fronts for $c < c_+$, particularly not for $c < c_-$. Indeed, for $c \in (c_-, c_+)$ the point $(0,0)$ is a stable focus, and hence fronts (if they exist) are oscillating. For $c \in (0, c_-)$ the point $(0,0)$ is again a stable node. If the unstable manifold connects to $(0,0)$, then it first leaves the domain $u > 0$.

Now it remains to check whether the lower bound c_+ can be achieved. Here we use the subtangential condition (which in the Fisher case is sufficient for linear determinacy; see [9], [14]),

$$g(u) \leq g'(0)u. \tag{5.3}$$

The following proposition can be shown by comparison arguments.

PROPOSITION 5.2. *Assume that the function g satisfies the subtangential condition* (5.3). *Then $c_0 = c_+$.*

We remark that a unique traveling front can also be shown to exist for "bistable" nonlinearities such as $g(\Theta) = \Theta(\Theta - \alpha)$ with some $\alpha > 0$.

6. Conclusion. The combustion model of Brailovsky and Sivashinsky is, in mathematical terms, a degenerate parabolic system. Therefore one expects deflagration fronts traveling with asymptotically constant shape and speed. Shape and speed can be found from a nonlinear boundary value problem for a system of ordinary differential equations of order four with boundary conditions at infinity. In this paper we have found a new conservation law for the full system of partial differential equations, which yields an invariant of motion for the ordinary differential equations. The latter have a second invariant of motion independent of the first. With these two invariants and some invertible transformation we can reduce the four-dimensional problem to a two-dimensional problem, which can be treated by phase plane methods. These analytical steps work for the given system as well as for the simplified system which was studied by Brézis, Kamin, and Sivashinsky and further for a one-parameter family of problems connecting the full system and the simplified system. This approach yields the exact speed of propagation. A comparison property shows that it is always larger than the previously obtained estimates. Concrete examples show that discrepancy between the correct solution and these estimates can be very large.

In contrast to the traveling front problem for the Fisher equation where the well-known formula $c_0 = 2\sqrt{Df'(0)}$ holds under rather general concavity assumptions on the nonlinearity f, here in the combustion problem there is no explicit simple formula

(and cannot be); i.e., in concrete examples (other than in limiting cases as in the appendix) the two-dimensional boundary value problem (3.14) must be solved by a numerical method, e.g., by a shooting method starting at the unstable manifold of the stationary point behind the front $(1,0)$.

7. Appendix. Proof of formula (4.2). First observe that c^2 scales with κ, and hence we can choose $\kappa = 1$. In the domain $u + v > \beta$ we have

$$\begin{aligned} u' &= v, \\ v' &= -\frac{1}{c^2\gamma} + \frac{u}{c^2\gamma} - \frac{1}{c^2\gamma}(1 + \epsilon(\gamma - 1) - c^2)v. \end{aligned}$$

This is an inhomogeneous linear system. The stationary point $(1,0)^T$ yields a special constant solution. The homogeneous system has a 2×2 matrix with a positive and a negative eigenvalue (recall that $(1,0)$ is a saddle point). The positive eigenvalue λ_1 corresponds to the unstable manifold:

$$\lambda_1 = \frac{1 + \epsilon(\gamma - 1) - c^2}{2c^2\gamma} + \frac{1}{2c^2\gamma}\left[(1 + \epsilon(\gamma - 1) - c^2)^2 + 4c^2\gamma\right]^{1/2}.$$

The eigenvector is $(1, \lambda_1)^T$; hence the unstable manifold is given by

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ \lambda_1 \end{pmatrix} e^{\lambda_1 t}. \tag{7.1}$$

This manifold is supposed to meet the point P_0 given in (3.15) for some value of t. This condition yields two equations,

$$1 - e^{\lambda_1 t} = \frac{\gamma}{\gamma - 1}\beta, \qquad -\lambda_1 e^{\lambda_1 t} = -\frac{1}{\gamma - 1}\beta. \tag{7.2}$$

We eliminate $e^{\lambda_1 t}$ and get

$$\lambda_1\left(1 - \frac{\gamma}{\gamma - 1}\beta\right) = \frac{1}{\gamma - 1}\beta. \tag{7.3}$$

Recall that λ_1 depends on c. In (7.3) we solve for c and get the explicit formula (4.2).

REFERENCES

[1] D. G. Aronson and H. F. Weinberger, *Nonlinear diffusion in population genetics, combustion, and nerve pulse propagation*, in Partial Differential Equations and Related Topics (Tulane University, 1974), Lecture Notes in Math. 446, Springer, Berlin, 1975, pp. 5–49.

[2] I. Brailovsky and G. I. Sivashinsky, *On deflagration-to-detonation transition*, Comb. Sci. Tech., 130 (1997), pp. 201–231.

[3] I. Brailovsky and G. I. Sivashinsky, *Momentum loss as a mechanism for deflagration-to-detonation transition*, Combust. Theory Model., 2 (1998), pp. 429–447.

[4] I. Brailovsky, M. Frankel, and G. Sivashinsky, *Galloping and spinning modes of subsonic detonation*, Combust. Theory Model., 4 (2000), pp. 47–60.

[5] H. Brézis, S. Kamin, and G. Sivashinsky, *Initiation of subsonic detonation*, Asymptotic Anal., 24 (2000), pp. 73–90.

[6] H. Berestycki, S. Kamin, and G. Sivashinsky, *Metastability in a flame front evolution equation*, Interfaces Free Bound., 3 (2001), pp. 361–392.

[7] F. Dkhil, *Travelling wave solutions in a model for filtration combustion*, Nonlinear Anal., 58 (2004), pp. 395–415.

[8] K. P. Hadeler and F. Rothe, *Travelling fronts in nonlinear diffusion equations*, J. Math. Biol., 2 (1975), pp. 251–263.

[9] K. P. Hadeler, *Nonlinear propagation in reaction transport systems*, in Differential Equations with Applications to Biology, S. Ruan and G. Wolkowicz, eds., The Fields Institute Lecture Series 21, The Fields Institute, Toronto, 1999, pp. 251–257.

[10] K. P. Hadeler and M. A. Lewis, *Spatial dynamics of the diffusive logistic equation with a sedentary compartment*, Canad. Appl. Math. Quart., 10 (2004), pp. 473–499.

[11] S. Minaev, I. Kagan, G. Joulin, and G. Sivashinsky, *On self-drifting flame balls*, Combust. Theory Model., 5 (2001), pp. 609–622.

[12] P. V. Gordon, S. Kamin, and G. I. Sivashinsky, *On initiation of subsonic detonation in porous media combustion*, Asymptotic Anal., 29 (2002), pp. 309–321.

[13] P. Gordon, L. S. Kagan, and G. I. Sivashinsky, *Fast subsonic combustion as a free-interface problem*, Interfaces Free Bound., 5 (2003), pp. 47–62.

[14] H. F. Weinberger, M. A. Lewis, and B. Li, *Analysis of linear determinacy for spread in cooperative models*, J. Math. Biol., 45 (2002), pp. 183–218.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1482–1511

SMALL- AND WAITING-TIME BEHAVIOR OF A DARCY FLOW MODEL WITH A DYNAMIC PRESSURE SATURATION RELATION*

J. R. KING† AND C. M. CUESTA‡

Abstract. We address the small-time evolution of interfaces (fronts) for the pseudoparabolic generalization

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(u^\alpha \frac{\partial u}{\partial x} + u^\beta \frac{\partial^2 u}{\partial x \partial t}\right)$$

of the porous-medium equation, identifying regimes in which the local behavior remains fixed for some finite time and others in which it changes instantaneously. A number of phenomena beyond those exhibited by the porous-medium equation are elucidated, including retreating fronts and novel types of local behavior. Related results for the important limit case

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(u^\beta \frac{\partial^2 u}{\partial x \partial t}\right)$$

are also described.

Key words. degenerate pseudoparabolic equation, small-time behavior, waiting-time phenomena, time-reversibility

AMS subject classifications. 35K70, 35K65, 76S05

DOI. 10.1137/040610969

1. Introduction.

In this paper we consider the Cauchy problem for the degenerate pseudoparabolic equation

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(u^\alpha \frac{\partial u}{\partial x} + u^\beta \frac{\partial^2 u}{\partial x \partial t}\right), \tag{1.1}$$

where α and β are positive constants, with initial condition $u(x,0) = u_0(x)$ having $u_0(x) = 0$ for $x \geq a$ which, for definiteness, we take to satisfy

$$u_0(x) \sim A(a-x)_+^p \quad \text{as } x \to a \tag{1.2}$$

for some positive constants A and p. Since we are concerned here with the local behavior near the right-hand interface $x = a$, that near the left-hand one (if there is one) need not concern us. We shall need to defer detailed discussion of the boundary conditions which can hold at an interface $x = s(t)$ until we have analyzed the permissible local behavior of solutions. However, we shall throughout impose the conservation of mass condition

$$u^\alpha \frac{\partial u}{\partial x} + u^\beta \frac{\partial^2 u}{\partial x \partial t} = 0 \quad \text{at } x = s(t)\,. \tag{1.3}$$

*Received by the editors July 3, 2004; accepted for publication (in revised form) February 21, 2006; published electronically May 19, 2006. This work was supported by the RTN project "Front-singularities."

http://www.siam.org/journals/siap/66-5/61096.html

†Division of Theoretical Mechanics, School of Mathematical Sciences, University of Nottingham, NG7 2RD Nottingham, UK (john.king@nottingham.ac.uk). The research of this author was supported by the EPSRC.

‡Faculty of Mathematics, University of Vienna, Nordbergstrasse 15, 1090 Wien, Austria (pmxcc@maths.nottingham.ac.uk).

Aspects of the physical background to (1.1) are given in [21] and [10]; see also [11]. We summarize these here. The model equation for unsaturated flow in porous media with a dynamic capillary pressure relation has the general form

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(K(u)\,\frac{\partial}{\partial x}\left(-p_c(u) + L\,\frac{\partial u}{\partial t}\right)\right). \tag{1.4}$$

The unknown is the water saturation u in a horizontally placed one-dimensional porous medium, $K(u)$ is the hydraulic conductivity, $p_c(u)$ is the capillary pressure function, and L can be regarded as a damping coefficient. Under the assumption that the water saturation u is small, adopting power laws for these nonlinear functions of u leads to (1.1). Typically $p_c(u)$ is a nonincreasing function, such that $p_c'(0^+) = \infty$, and $K(u)$ is increasing, with $K(0) = K'(0) = 0$. According to these criteria, when (1.1) provides a small u approximation to (1.4) the exponents necessarily satisfy $\alpha < \beta$; nevertheless we consider here all possible positive α and β for completeness and because of other possible applications. We note that (1.4) combines conservation of mass and Darcy's law with the pressure-saturation relation

$$p_c(u) = p_a - p_w + L\,\frac{\partial u}{\partial t} \quad (L > 0)\,, \tag{1.5}$$

where p_a and p_w denote the air and water pressures. This equation is the classical capillary-pressure relation extended by a relaxation term ($L > 0$), and models capillary forces, taking into account dynamic effects. This extension of the Darcy flow model is based on the approach introduced in [19], motivated by previous experimental work by Stauffer [32], among others; see [17] and [18] for an overview. Earlier related models can be found in [6], [4], and [3]. For derivations of (1.1) on the basis of homogenization and hysteresis ideas, see also [7] and [20], respectively. Third order mixed derivatives terms also appear as regularizations of forward-backward diffusion equations as in [30], and in the viscous Cahn–Hilliard equation [29]; other degenerate regularizations of forward-backward equations arise in the case of the viscous Cahn–Hilliard equation with degenerate mobility (see [14]) and as a model for poroviscous thin-film flows, which relates closely to (1.1); see [26]. Equation (1.1) can be obtained as a (somewhat ad hoc) model of thin-film flows of viscoelastic fluids, with $\alpha = \beta = 3$ (no slip) or $\alpha = \beta = 2$ (strong slip), if instead of surface tension (see [31]) gravity is considered as the driving force, the usual Reynolds equation being replaced for a "Maxwell" fluid by

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(u^\alpha\,\frac{\partial}{\partial x}\left(p + \tau\,\frac{\partial p}{\partial t}\right)\right)$$

(where u is the film thickness, p the pressure, and τ the relaxation time). Finally we mention the work by Düll [13], where (1.1) with $\alpha = \beta$ appears as a model of case II diffusion in polymers.

Observe that the classical capillary pressure relation ($L = 0$ in (1.5)) yields the familiar porous-medium equation (PME)

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(u^\alpha\,\frac{\partial u}{\partial x}\right). \tag{1.6}$$

This equation is very widely studied as a paradigm degenerate-diffusion equation; see [1]. An interesting property of the PME is the occurrence of fronts: interfaces

that, in terms of saturation, separate wet ($u > 0$) and dry ($u = 0$) regions. Many of the intriguing phenomena we shall discuss for (1.1) arise even in the absence of the diffusion term, i.e., for

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(u^{\beta}\frac{\partial^2 u}{\partial x \partial t}\right), \tag{1.7}$$

though in the case of (1.7) the evolution has to be forced somehow since the solution to the Cauchy problem is simply $u = u_0(x)$. Equation (1.7) with $\beta = 1$ governs (see [15] and the references therein) the thin-film limit of the Hele–Shaw problem with linear kinetic undercooling (similarly, (1.1) with $\alpha = \beta = 1$ arises when both gravity and kinetic undercooling are present) and, to the best of our knowledge, first arose [23] in the description of the growth of oxide layers during the fabrication of silicon integrated circuits. It is also a limit case of (1.1) for large α (for $u < 1$), and it is striking that, because only the time derivative of u appears inside the highest spatial derivative, it typically does not smooth nonanalyticities in the initial data. Equation (1.7) is formally time-reversible (i.e., invariant under $t \to -t$), but it is instructive to highlight straight away one of the noteworthy consequences that follow from the analysis later in the paper, namely, that this time-reversibility is in fact illusory. Thus we consider (1.7) with $\beta = 1$, say, subject to suitable compactly supported nonnegative initial data in $x > 0$ and to

$$u(t) = \int_0^t P(\tau)d\tau \quad \text{at } x = 0\,, \tag{1.8}$$

with the inlet pressure $P(t)$ prescribed. Were the problem time-reversible, then if, say, the sign of P changes at $t = t_1$ such that the right-hand side of (1.8) returns to zero at $t = t_2 > t_1$, then at $t = t_2$ we would simply recover the initial data for all x. If $P > 0$ for $0 < t < t_1$ and $P < 0$ for $t_1 < t < t_2$ (for brevity we consider only a single change in sign), this is indeed what transpires for the problem formulation described later in the paper, which implies here that u has zero slope at the interface throughout $0 < t < t_2$ (corresponding to the motion of a thin film of fluid in a Hele–Shaw cell over a flat substrate which it completely wets, in the sense of the contact angle at an advancing contact line being zero; advancing and retreating contact lines have distinct behaviors for reasons which are to some extent implicit in the analysis of [15] and are addressed in detail below). However, in the converse case the interface retreats for $0 < t < t_1$ (while P is negative; we assume it is not so negative that the interface reaches $x = 0$) with u in general having finite slope there, implying that this retraction is nonreversible: while the interface advances again for $t_1 < t < t_2$ (with $P > 0$), it does so with zero contact angle and the initial conditions are not recovered. The unregularized Hele–Shaw problem and the Hele–Shaw problem with linear kinetic undercooling share formal time-reversibility. In the former case the suction (negative-pressure) problem typically leads to finite-time singularities through which the solution cannot usually be meaningfully continued (see, for example, [12] and the references therein); in the latter case the kinetic-undercooling regularization ensures that, even if singularities form, the solution can be continued through them and the failure of time-reversibility has important implications for this and other such problems, some of which are explored briefly below. Thus the formulation (1.7) is an important problem in its own right, the theory for which is in its infancy compared to that of (1.6), and our results, notably those of the appendix, encompass this special case of (1.1).

In the context of occurrence of fronts, Hulshof and King [21] studied the behavior of solutions near fronts for (1.1). They show that for $2 \le \beta < 2\alpha$ the interface remains fixed (at $x = a$) for all time and the solution develops a discontinuity there with

$$u(a^-, t) = U(t),\ u(a^+, t) = 0\,. \tag{1.9}$$

We emphasize that the solution is then genuinely discontinuous, with $u = 0$ holding for all t throughout $x > a$. By contrast, in the other two ranges of (α, β) the front moves, and behaves locally as a traveling wave. In the range $\beta > 2\alpha$ the moving-front behavior is of PME type, whereas in the range $\beta < \min(2, 2\alpha)$ the second order term in (1.1) is irrelevant at the front, and the behavior of the moving front is "new." In the fixed-front cases (1.3) with $s(t) = a$ suffices to furnish a correctly specified problem, but in the moving-front cases additional conditions involving appropriate prescriptions of the local behavior are required; these are formulated in section 3 below.

It is well known for (1.6) that interfaces can either move immediately or remain fixed for a finite time (the waiting time) and then start to move; see, for instance, [8] and [2]. The purpose of this paper is to investigate such phenomena for (1.1). In some fixed-front cases we shall find that $U(t) > 0$ for all $t > 0$, while in the remainder (which we term "waiting-time" cases) we have

$$U(t) = 0 \quad \text{for} \quad 0 \le t \le t_w, \quad U(t) > 0 \quad \text{for} \quad t > t_w \tag{1.10}$$

for some waiting time $t_w > 0$. ($U(t)$ may in principle return to zero for some $t > t_w$, but we shall not investigate such cases here.) In other ranges of (α, β) the interfaces move, at least for sufficiently large t. In some of these cases we shall see that the right-hand interface, $x = s(t)$ with $s(0) = a$, satisfies $s(t) > a$ for all $t > 0$, while for the waiting-time regimes we have

$$s(t) = a \quad \text{for} \quad 0 \le t \le t_w\,. \tag{1.11}$$

We note that in some regimes both retreating and advancing fronts are possible so that, in contrast to (1.6), $s(t) > a$ need not hold for $t > t_w$. Waiting-time solutions should not be confused with fixed-front cases, even though (1.11) is satisfied by both; in view of the first expression in (1.10), fixed- and moving-front regimes are indistinguishable during the waiting period, but $U > 0$ ultimately applies in the former case, distinguishing waiting fronts (having $U = 0$) from fixed ones.

Our analysis here aims to clarify such matters by analyzing the local behavior close to $x = a$; moreover, we shall also identify retreating-front cases whereby $s(t) < a$ for sufficiently small positive t, say. The goal of this paper is thus to develop a comprehensive small-time classification of (1.1)–(1.2) comparable to that established for the PME in [16], [27] (see also [28]) and, in doing so, to identify new types of qualitative behavior which are of significance beyond the small-time regime; we shall see that the range of possibilities for (1.1) is significantly richer than that for (1.6).

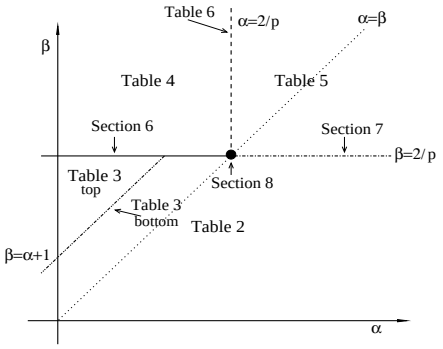
In the next two sections we give the necessary ingredients for our subsequent analysis. In section 3 we reexamine the local balances for (1.1) and identify regimes of α and β in which retreating fronts are possible. We then turn to the analysis of the different relevant small-time regimes. Most of these are for brevity summarized in tables (we refer the interested reader to [24], where the analysis of all cases is given explicitly). Section 5 examines cases in which (1.6) dominates the local behavior, a first case where there is no waiting-time, a case where there are "global" waiting-time scenarios (cf. [27] for (1.6)), and the "local" waiting-time case (cf. [28] for (1.6)).

TABLE 1

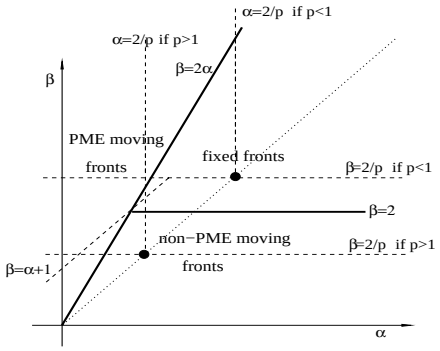
Each table or section listed below describes a distinct small-time regime. In the borderline cases we have fronts that move exponentially slowly with t. For a discussion of the distinction between "global" and "local" cases see the relevant sections below. "Immediate change" means the local solution instantaneously adjusts, but a waiting-time scenario may still ensue.

	$\beta < \alpha + 1$	$\beta = \alpha + 1$	$\beta > \alpha + 1$
$p < 2/\beta$ non-PME balance	Table 2[1]	Table 3 (bottom)	Table 3 (top)
	no waiting time		

	$p > 2/\alpha$	$p = 2/\alpha$	$p < 2/\alpha$
$p > 2/\beta$ PME balance	Table 5[1] "global" waiting time	Table 6 ($\beta > \alpha$) "local" waiting time	Table 4 ($\beta > \alpha$) no waiting time
$p = 2/\beta$ non-PME balance	Section 7 ($\beta < \alpha$) (I) immediate change (II) "global" waiting time (III) immediate change (borderline)	Section 8[1] ($\alpha = \beta$) (I) no waiting time (II) "local" waiting time (III) no waiting time (borderline)	Section 6 ($\beta > \alpha$) immediate change



(a) Regimes in the (α, β)-plane for fixed p.



(b) The regimes relevant to the small-time behavior and the moving-fixed-front cases.

FIG. 1. *The (α, β)-plane.*

Sections 5 and 6 discuss other non-waiting-time cases. Sections 7 and 8 discuss critical cases ($p = 2/\beta$ with $\beta < \alpha$ and $p = 2/\alpha$ with $\beta = \alpha$) in which the behavior is rather delicate (and novel) and we conclude in section 9 with a discussion.

Table 1 gives a summary of the results of small-time analysis. Figure 1 illustrates these regimes in the (α, β)-plane, together with the relevant moving/fixed front ranges. We observe that although in section 5 (i.e., for the case $p > 2/\beta$) the dominant balance is of PME type, the behavior may differ from PME, since this regime lies partially in the fixed-front ranges and partially in the non-PME moving-front range found in [21]; see Figure 1(b).

[1]Relevant regimes for the applications with $\alpha = \beta$; see the introduction.

2. Method/preliminaries. In order to identify the appropriate small-time behavior, we first introduce the small-time (outer) expansion

$$u \sim u_0(x) + u_1(x)t \qquad \text{as } t \to 0^+ \quad \text{with } a - x = O(1)\,, \tag{2.1}$$

which at leading order gives the equation

$$u_1 - \frac{d}{dx}\left(u_0^\beta \frac{du_1}{dx}\right) = \frac{d}{dx}\left(u_0^\alpha \frac{du_0}{dx}\right)\,. \tag{2.2}$$

The relevant boundary conditions on (2.2) will be noted as they arise. The expansion (2.1) may be nonuniform as $x \to a$, depending on the behavior of $u_0(x)$. If it is, then it is typically necessary to introduce a local similarity variable $\eta = (x-a)/t^\omega$, where ω needs to be identified, and consider the inner region $\eta = O(1)$ as $t \to 0^+$. It will be necessary to distinguish several regimes. Observe that in (2.2) the first term of the left-hand side dominates the second as $x \to a$ if $p > 2/\beta$, while the second dominates the first if $p < 2/\beta$. In the latter case, and in the borderline case $p = 2/\beta$, the sign of $\beta - (\alpha+1)$ dictates the dominant balance as $x \to a$. This leads us to distinguish regimes for which we will have different similarity (inner) behavior.

We now summarize some possible small-time balances in (1.1); these will provide the crucial ingredients in the analysis that follows. First, if we assume that the solution is of the form $u(x,t) \sim t^\gamma f(\eta)$ for small t, then by (1.1) possible small-time balances can be identified from

$$\gamma f - \omega\eta\frac{df}{d\eta} \sim t^{\alpha\gamma-2\omega+1}\frac{d}{d\eta}\left(f^\alpha\frac{df}{d\eta}\right) + t^{\beta\gamma-2\omega}\frac{d}{d\eta}\left(f^\beta\left((\gamma-\omega)\frac{df}{d\eta} - \omega\eta\frac{d^2f}{d\eta^2}\right)\right)\,. \tag{2.3}$$

This gives the following possibilities.

(i) If all the terms of (2.3) balance as $t \to 0^+$, then $\omega = \beta/2(\beta-\alpha)$, $\gamma = 1/(\beta-\alpha)$ and

$$u \sim t^{\frac{1}{(\beta-\alpha)}}\, f\left((x-a)/t^{\frac{\beta}{2(\beta-\alpha)}}\right) \quad \text{if } \beta \neq \alpha\,, \qquad u \sim (a-x)^{\frac{2}{\alpha}}\, f(t) \quad \text{if } \beta = \alpha\,, \tag{2.4}$$

which give exact similarity reductions of (1.1).

(ii) If the left-hand side and the first term on the right-hand side of (2.3) dominate as $t \to 0^+$, we get a one-parameter (γ) family of similarity solutions with $\omega = (1+\alpha\gamma)/2$, so

$$u \sim t^\gamma f\left((x-a)/t^{\frac{1}{2}(1+\alpha\gamma)}\right) \quad \text{as} \quad t \to 0^+\,. \tag{2.5}$$

This is a similarity reduction of the PME limit of (1.1), (1.6) and is consistent for (1.1) as $t \to 0^+$ (i.e., the omitted terms in (2.3) are indeed negligible) when $\gamma > 0$ if $\beta > \alpha$ with $\gamma > 1/(\beta-\alpha)$; the precise value of γ can readily be determined from that of p in (1.2) (see below).

(iii) If the left-hand side and the last term of the right-hand side in (1.1) balance and dominate as $t \to 0^+$, then the small-time behavior might be of the form

$$u \sim \Omega^{\frac{2}{\beta}}(t) f\left((x-a)/\Omega(t)\right)\,, \tag{2.6}$$

which is a similarity reduction of (1.7) for any $\Omega(t)$, and hence f satisfies the equation

$$f - \frac{\beta}{2}\eta\frac{df}{d\eta} = \frac{d}{d\eta}\left(f^\beta\frac{d}{d\eta}\left(f - \frac{\beta}{2}\eta\frac{df}{d\eta}\right)\right)\,. \tag{2.7}$$

We observe that the required solution to this equation, supplemented with appropriate boundary and initial conditions, can in a number of cases be written in closed form by noting that

$$u^\beta = \frac{\beta^2}{2(2-\beta)}(s(t)-x)^2 \tag{2.8}$$

is an exact solution of (1.7) for any $s(t)$. Writing $s(t) - a \sim \eta_0\Omega(t)$ we thus have the similarity solution

$$f^\beta(\eta) = \frac{\beta^2}{2(2-\beta)}(\eta_0-\eta)^2\,. \tag{2.9}$$

In many cases that follow it will suffice to restrict our attention to the special one-parameter (power-law) case (the parameter being again γ)

$$u \sim t^\gamma f\left((x-a)/t^{\frac{\beta\gamma}{2}}\right) \quad \text{as} \quad t \to 0^+\,, \tag{2.10}$$

whereby $\omega = \beta\gamma/2$, which, from (2.3), is consistent for (1.1) when $\gamma > 0$ if $\alpha < \beta$ with $\gamma < 1/(\beta-\alpha)$ or if $\alpha \geq \beta$.

(iv) If the two terms of the right-hand side balance and dominate as $t \to 0^+$, then $\gamma = 1/(\beta-\alpha)$ and the expected small-time behavior is of the one-parameter (namely, ω) form

$$u \sim t^{\frac{1}{(\beta-\alpha)}} h\left((x-a)/t^\omega\right) \quad \text{as} \quad t \to 0^+\,, \tag{2.11}$$

which gives a similarity reduction to

$$\frac{\partial}{\partial x}\left(u^\alpha \frac{\partial u}{\partial x}\right) \sim -\frac{\partial}{\partial x}\left(u^\beta \frac{\partial^2 u}{\partial x \partial t}\right) \tag{2.12}$$

and, from (2.3), is consistent for (1.1) if $\beta > \alpha$ with $\omega > \beta/2(\beta-\alpha)$. Requiring mass conservation, (1.3), and from the matching conditions which will subsequently follow, it follows that

$$\frac{1}{\alpha-\beta+1} u^{\alpha-\beta+1} \sim -\frac{\partial u}{\partial t} \tag{2.13}$$

for $\beta \neq \alpha + 1$. From this we obtain

$$u^{\beta-\alpha}(x,t) \sim \frac{\alpha-\beta}{\alpha-\beta+1} t + U_0^{\beta-\alpha}(x) \tag{2.14}$$

for some function $U_0(x)$ which we shall subsequently equate to the initial data $u_0(x)$; $U_0 = A(a-x)^{1/(\beta-\alpha)\omega}$ for constant A corresponds to the self-similar form (2.11).

(v) If the final term in (1.1) dominates, that is (in a slight abuse of notation),

$$\frac{\partial}{\partial x}\left(u^\beta \frac{\partial^2 u}{\partial x \partial t}\right) \sim 0\,, \tag{2.15}$$

then

$$u \sim t^\gamma h((x-a)/t^\omega) \quad \text{as } t \to 0^+ \tag{2.16}$$

provides a similarity reduction for any γ and ω, self-consistency demanding that $\beta\gamma < 2\omega$ and $(\beta-\alpha)\gamma < 1$. Imposing conservation of mass at $x = a$ implies from (2.15) that $\partial^2 u/\partial x \partial t \sim 0$ so for some constants A and B,

$$h(\eta) = A(-\eta)^{\frac{\gamma}{\omega}} + B\,. \tag{2.17}$$

(vi) Other possible small-time balances are of traveling-wave type, and will typically give inner-inner regions enabling the appropriate local behavior at the front to be attained. Setting $x = s(t) + \zeta$ gives

$$\frac{\partial u}{\partial t} - \dot{s}\frac{\partial u}{\partial \zeta} = \frac{\partial}{\partial \zeta}\left(u^\alpha \frac{\partial u}{\partial \zeta} - \dot{s}u^\beta \frac{\partial^2 u}{\partial \zeta^2} + u^\beta \frac{\partial^2 u}{\partial \zeta \partial t}\right), \tag{2.18}$$

so if the right-hand side is negligible, then the local solution scales according to

$$u \sim \dot{s}^{\frac{2}{2\alpha-\beta}}(t) G\left((x - s(t))/\dot{s}^{\frac{\beta}{2\alpha-\beta}}(t)\right), \tag{2.19}$$

the neglect of the right-hand side of (2.18) requiring that $\ddot{s} \ll \dot{s}^{\frac{(4\alpha-3\beta)}{(2\alpha-\beta)}}$; $G(Z)$ satisfies

$$\frac{d}{dZ}\left(G^\beta \frac{d^2 G}{dZ^2}\right) - \frac{d}{dZ}\left(G^\alpha \frac{dG}{dZ}\right) - \frac{dG}{dZ} = 0\,. \tag{2.20}$$

In particular, we may have for $\beta \neq 2\alpha$ that

$$u \sim t^{\frac{2(\nu-1)}{(2\alpha-\beta)}} g\left((x - \mu t^\nu)/t^{\frac{\beta(\nu-1)}{(2\alpha-\beta)}}\right) \quad \text{as} \quad t \to 0^+\,, \tag{2.21}$$

where $\nu > 0$, and if $\alpha < \beta < 2\alpha$, then $\nu > \beta/2(\beta-\alpha)$, or if $\beta > 2\alpha$, then $\nu < \beta/2(\beta-\alpha)$, in view of the condition $\ddot{s} \ll \dot{s}^{\frac{(4\alpha-3\beta)}{(2\alpha-\beta)}}$. Finally, a possible small-time balance of traveling-wave type for $\beta = 2\alpha$ is

$$u \sim t^\sigma g\left((x - \mu t)/t^{\alpha\sigma}\right) \quad \text{as} \quad t \to 0^+ \tag{2.22}$$

for $\sigma > 2/\beta$. Conserving mass at the interface $z = 0$, it follows from (2.21) that $g(z)$ satisfies

$$-\mu\nu g = g^\alpha \frac{dg}{dz} - \mu\nu g^\beta \frac{d^2 g}{dz^2}, \tag{2.23}$$

solutions to this traveling-wave ordinary differential equation (ODE) having been discussed in detail in [21]; for (2.22) we obtain (2.23) with $\nu = 1$.

When matching the similarity behavior into traveling-wave inner-inner regions, knowledge of the local behavior of fronts described in [21] will be valuable. It is convenient at this stage to summarize these results in the following form; however, we shall need to revisit this classification, which we do in section 3.

(a) Moving-front cases. The ranges in which the interfaces move, at least for sufficiently large times, and the corresponding local behaviors, are as follows:

$$u^\alpha \sim \alpha \dot{s}(t)(s(t) - x) \quad \text{as} \quad x \to s(t)^- \quad \text{if } \beta > 2\alpha\,, \tag{2.24}$$

requiring $\dot{s} \geq 0$ (here the β term in (1.1) is locally negligible);

$$u^\alpha \sim \frac{2\alpha\dot{s}(t)}{\left(1 + \sqrt{1 + 4(1-\alpha)\dot{s}(t)^2}\right)}\,(s(t) - x) \quad \text{as} \quad x \to s(t)^- \quad \text{if } \beta = 2\alpha\,, \tag{2.25}$$

requiring for all $\alpha > 0$ that $\dot{s} \geq 0$ and for $\alpha > 1$ that $\dot{s} \leq 1/2(\alpha - 1)^{\frac{1}{2}}$ (both terms on the right-hand side of (1.1) contribute to this local balance, except when $\alpha = 1$, $\beta = 2$, in which case (2.24) is recovered);

$$u^\beta \sim \frac{\beta^2}{2(2-\beta)}(s(t)-x)^2 \quad \text{as} \quad x \to s(t)^- \quad \text{if } \beta < \min(2, 2\alpha) \tag{2.26}$$

(cf. (2.8)), where $\dot{s}$ can in principle have either sign (in this case the α term from (1.1) is locally negligible).

(b) Fixed-front cases, $2 \leq \beta < 2\alpha$. As already noted (see (1.9) and (1.10)), the front remains fixed in the regime, and u satisfies the conservation of mass condition

$$u^\alpha \frac{\partial u}{\partial x} + u^\beta \frac{\partial^2 u}{\partial x \partial t} = 0 \quad \text{at} \quad x = a\,. \tag{2.27}$$

3. Classification of solution branches in moving-front regimes.

3.1. Local analysis. Equation (1.1) is of second order in x, so in fixed-front cases the no-flux condition (2.27) is sufficient to specify the solution uniquely. However, in the frame of reference of a moving front, we have (2.18), which is third order in ζ, so that the usual mass-conserving interface conditions

$$u \to 0, \quad u^\alpha \frac{\partial u}{\partial x} + u^\beta \frac{\partial^2 u}{\partial x \partial t} \to 0 \quad \text{as } x \to s^-(t) \tag{3.1}$$

may not suffice to completely specify the moving-boundary problem. Since (2.18) is third order, solutions to (1.1) can have up to three degrees of freedom in their local expansion about a moving front; we pursue below a classification of the local behavior which clarifies how the problem can be made correctly specified. We now describe the various possible solution branches and, in particular, identify new classes of retreating fronts; these did not arise in [21] because only advancing fronts were considered there. (It is worth remarking that for PME retreating fronts are not possible and such behavior thus represents a qualitatively distinct, and very important, effect of including the β term in (1.1); it can occur for $\beta < \min(2, \alpha + 1)$, which includes some of the cases of physical interest, e.g., $0 < \alpha = \beta < 2$.)

(A) $\beta > \max(\alpha+1, 2\alpha)$. Here the PME local form (2.24) is correctly specified, containing the single degree of freedom s. The balance

$$u^\alpha \frac{\partial u}{\partial \zeta} \sim \dot{s} u^\beta \frac{\partial^2 u}{\partial \zeta^2} \quad \text{as } \zeta \to 0^-$$

is also possible in (2.18), implying

$$\dot{s}\frac{\partial u}{\partial \zeta} \sim -\frac{u^{-(\beta-\alpha-1)}}{\beta-\alpha-1} + \Phi(t) \tag{3.2}$$

containing two degrees of freedom, s and Φ (other terms may intrude between the degrees of freedom, but we shall not keep track of such terms here), and having

$$u^{\beta-\alpha} \sim \frac{\beta-\alpha}{\beta-\alpha-1}\frac{s(t)-x}{\dot{s}(t)} \quad \text{as } x \to s^-(t)\,. \tag{3.3}$$

As might be expected from its having an additional degree of freedom, (3.3) is less smooth at $\zeta = 0$ than is (2.24); the signs imply that $\dot{s} > 0$ (advancing front) is required in both cases. The case $\beta = \alpha + 1$, in which (3.2) becomes

$$\dot{s}\frac{\partial u}{\partial \zeta} \sim \ln u + \Phi(t)$$

so that

$$u \sim \frac{1}{\dot{s}(t)}(s(t) - x) \ln\left(\frac{1}{s(t) - x}\right) \quad \text{as } x \to s^-(t)\,,$$

in consequence represents a new borderline regime for (1.1).

(B) $\boldsymbol{\beta < \min(2, 2\alpha)}$. Here the balance (2.26), for which the α term is locally negligible, contains one degree of freedom, s, and gives no constraints on the sign of $\dot{s}$ (so advancing and retreating fronts are both possible, at least in principle). From the local balance

$$u \sim u^\beta \frac{\partial^2 u}{\partial \zeta^2} \quad \text{as } \zeta \to 0^-$$

in (2.18) we also have

$$\frac{1}{2}\left(\frac{\partial u}{\partial \zeta}\right)^2 \sim \frac{u^{(2-\beta)}}{2 - \beta} + \frac{1}{2}\Lambda^2(t)\,, \tag{3.4}$$

which provides a two-degree-of-freedom (s and Λ) branch; when $\Lambda = 0$ we recover (2.26), while for $\Lambda > 0$ we have finite-slope local behavior,

$$u \sim \Lambda(t)(s(t) - x) \quad \text{as } x \to s^-(t), \tag{3.5}$$

and advancing and retreating fronts are again both possible. The local balance (3.5) is self-consistent only for $\beta < \min(2, \alpha + 1)$.

(C) $\boldsymbol{2\alpha < \beta < \alpha + 1}$. This is the most subtle regime. First, the two-degree-of-freedom branch (3.5) again represents a self-consistent balance, as does the one-degree-of-freedom (and zero-slope) branch (2.24), in which $\dot{s} > 0$ necessarily holds. However, the one-degree-of-freedom branch (3.3), in which we now require $\dot{s} < 0$, is also possible. For (3.3) to hold with $\beta < \alpha + 1$ we require $\Phi = 0$ in (3.2) (setting $\Phi \equiv -\dot{s}\Lambda \neq 0$ leads in the current regime to (3.5) rather than (3.2)).

3.2. Classification of advancing and retreating fronts. We focus exclusively here (and in the rest of the paper) on the smoothest (and, in consequence, correctly specified) solution branches, namely, in the case of advancing fronts, the one-degree-of-freedom branches (2.24) for (A) (in this case a distinct one-degree-of-freedom branch arises by enforcing $\Phi = 0$ in (3.2)) and for (C), and (2.26) for (B); see [21] and [10] for a discussion (in terms of a lifting regularization) of why these solution branches may be expected on physical grounds typically to be the relevant ones. However, there is an important distinction between advancing and retreating fronts which is perhaps most easily seen by rewriting (1.1) in the form

$$\frac{\partial}{\partial x}\left(u^\beta \frac{\partial w}{\partial x} + u^\alpha \frac{\partial u}{\partial x}\right) = w\,, \quad w = \frac{\partial u}{\partial t}\,. \tag{3.6}$$

The first of these represents an elliptic two-point boundary value problem for w (treating u as known) which is to be solved subject to (1.3) (with an appropriate second boundary condition elsewhere; in the case of the Cauchy problem this will of course take the form (1.3) at the left-hand interface). In fixed and contracting domain cases the ODE for u (now treating w as known) requires no further boundary condition (with the evolution of $s(t)$ then being determined according to where u reaches zero); by contrast, for advancing fronts initial data on u is not available to the right of the current front location (i.e., in $x > s$), which is why a further boundary condition (such as zero slope) is needed there. Indeed, in order to specify the moving-boundary problem correctly (3.1) should be supplemented by

$$u^{\alpha-1}\frac{\partial u}{\partial x} \to -\dot{s}(t) \quad \text{as } x \to s^-(t)$$

in cases (A) and (C), and by

$$\frac{\partial u}{\partial x} \to 0 \quad \text{as } x \to s^-(t)$$

in case (B). Thus the results of section 3.1 need a different interpretation when $\dot{s} < 0$; retreating fronts are not possible in regime (A), but in regimes (B) and (C) they are and we need the two-degree-of-freedom branch (3.5) when $\dot{s} < 0$, in which (by the above argument) s and Λ are determined as part of the solution, i.e., (3.5) is a correctly specified branch when $\dot{s} < 0$ and (3.1) on its own specifies the moving-boundary problem correctly when the interface is retreating. In regime (C), the one-degree-of-freedom branch (3.3) has the status of a nongeneric (overspecified) branch which separates advancing fronts (2.24) from retreating ones (3.5). Since (1.1) is of second order in x, it is worth clarifying further why three boundary conditions are needed at an advancing front (this is associated with the convected version of (2.18) being of third order in ζ; this issue of the order of the highest spatial derivative that appears depending on the frame of reference is common for equations with mixed derivatives and has here important implications). The key point is that the nonanalyticity at $x = a$ of the initial data is not instantaneously smoothed. We have continuity conditions

$$[u]^+_- = \left[u^\alpha \frac{\partial u}{\partial x} + u^\beta \frac{\partial^2 u}{\partial x \partial t}\right]^+_- = 0 \quad \text{at } x = a \tag{3.7}$$

and, defining $u(a,t) = u_a(t)$ (which is positive for an advancing front), we have that the degrees of freedom in the local expansion are (assuming for brevity that $du_0/dx = 0$ at $x = a^-$) the functions $u_a(t)$, $\lambda_a(t)$, and $\Phi_\pm(x)$ in

$$u \sim u_a(t) + \lambda_a(t)(x-a) + \Phi_\pm(x) e^{-\int_0^t u_a^{\alpha-\beta}(t')\,dt'} \quad \text{as } x \to \pm a, \tag{3.8}$$

where the continuity of the u_a and λ_a terms corresponds precisely to the conditions (3.7) and $\Phi_-(x)$ can be thought of as being determined by the initial data, in which case it is the determination of the arbitrary function $\Phi_+(x)$ as part of the solution that requires a third boundary condition to hold at $x = s(t)$, the latter in effect serving to provide the initial data in the expanding domain $a < x < s(t)$. We remark that for (1.7), the explicit solution (2.8) implies that in this special case $\Phi_+(x) \equiv 0$; this will not be the case in general, however. Since for retreating fronts we have $u = 0$ for $x > s(t)$, where $s(t) < a$, the nonanalyticity at $x = a$ *is* lost in the case; this important distinction is implicit in some of the discussion that follows.

We now expand upon such matters for the limit case (1.7), in which the α term is absent, with $\beta < 2$. The initial value problem for (1.7) of course has the solution $u = u_0(x)$ for all t so we need to force the solution to evolve by, for example, imposing

$$u = u_b(t) \quad \text{at } x = 0 \tag{3.9}$$

for some prescribed $u_b(t)$. As noted in [23], if $s(t) > a$, then the solution in $x > a$ is given exactly by (2.8) (being the required one-degree-of-freedom (s) solution), with $s(t)$ to be determined. In $0 < x < a$ we are thus required to solve (1.7) subject to continuity of u and of flux, i.e., to

$$u = \left(\frac{\beta^2 s(t)^2}{2(2-\beta)}\right)^{\frac{1}{\beta}}, \quad u^\beta \frac{\partial^2 u}{\partial x \partial t} = -\left(\frac{\beta^2 s(t)^2}{2(2-\beta)}\right)^{\frac{1}{\beta}} \dot{s}(t) \quad \text{at } x = a^- ; \tag{3.10}$$

in other words, we have a fixed domain ($0 < x < a$) problem for (1.7), subject to (3.9) and (by elimination of s from (3.10))

$$u^{\frac{\beta}{2}} \frac{\partial^2 u}{\partial x \partial t} = -\sqrt{\frac{2-\beta}{2}} \frac{\partial u}{\partial t} \quad \text{at } x = a^- , \tag{3.11}$$

which also determines s via (3.10). (Because (1.7) does not smooth nonanalyticities in the initial data, (2.8) need not, and does not in general, apply at $x < a$.)

However, the above formulation applies only if $u_a(t) \geq 0$, and it can easily be seen that this need not apply. Thus, if picking $u_b(t) = \phi(t)$ leads to $\dot{u}_a > 0$ initially, then (in view of the apparent time-reversibility of (1.7)) taking $u_b(t) = \phi(-t)$ would lead to $\dot{u}_a < 0$ for small time, and a contracting front of finite slope will instead result, as in (3.5). This confirms that the moving-boundary problem for (1.7) is not in fact time-reversible, since if a front switches from retreating (which it will generically do with nonzero slope) to advancing, then the slope becomes zero instantaneously on reversal of direction. An exception to this is that for initial data of the extremely special form (2.8) at $t = 0$, for which contraction with zero slope occurs, reversibility is possible; zero-slope retraction will of course occur while $s(t) > a$ for u_a decreasing if u_a first increases, so that (2.8) holds in $a < x < s$, and then decreases again—as already indicated, such behavior is time-reversible, in contrast to cases in which u_a decreases and then increases again.

We now turn to the description of the small-time behavior, distinguishing several parameter regimes. We shall only analyse in any detail the retreating front cases, and advancing and fixed front cases where the behavior of the front or discontinuity cannot be directly inferred from the small-time similarity balances listed in section 2. All other cases are summarized in tables, in which we note the inner balances and the associated front behavior. The expansion (2.1) holds in the outer region. One of the self-similar forms (i)–(v) of section 2 provides the inner solution and in some cases yet narrower inner-inner and possibly inner-inner-inner regions are required to match the corresponding front behavior, i.e., (2.24), (2.25), or (2.26) for advancing front cases and (3.5) for retreating ones. In most cases in the narrowest region a traveling-wave balance (section 2 (vi)) provides the solution. In some cases there are multiple inner structures, whereby an inner region generates a new local exponent $p' \neq p$ and one then needs to consult the table appropriate to that new value to obtain the full asymptotic structure. The tables should be read from left to right and the last column in blocks from top to bottom (starting at cases where the (widest) inner region matches the front behavior). Thus the solution appearing at the top/right (specified by writing (inner)) also holds in the inner region of the subsequent cases in the same block of the table, and so on down the column.

TABLE 2

The non-PME balance $p < 2/\beta$, $\beta < \alpha + 1$, in which the β-term in (1.1) dominates the small-time behavior of the front. An immediate change occurs with, in the moving-front regime, the interface advancing (retreating) if $B > 0$ ($B < 0$).

<table>
<tr><td rowspan="5">$\beta < \alpha + 1$

No waiting time

Similarity form (2.16) with $\gamma = 1$, $\omega = 1/p$ (inner)</td><td rowspan="3">$B > 0$</td><td>$2 \le \beta < \alpha + 1$
(fixed fronts)</td><td>$U(t) \sim Bt$ as $t \to 0^+$</td></tr>
<tr><td colspan="2">Similarity form (2.16) with $\gamma = 1$, $\omega = 1/p$ (inner)</td></tr>
<tr><td>$\beta < 2$, $\beta \le 2\alpha$
(moving fronts)</td><td>$s(t) - a \sim \eta_0 t^{\frac{\beta}{2}}$ as $t \to 0^+$

$(\eta_0 = (2(2-\beta))^{1/2} B^{\beta/2}/\beta)$</td></tr>
<tr><td colspan="2">Similarity form (2.10) with $\gamma = 1$ (inner-inner)</td></tr>
<tr><td>$2\alpha < \beta < \alpha + 1$
(moving fronts)</td><td>$g^\alpha(z) \sim \alpha\mu\nu(-z)$ as $z \to 0^-$
(inner-inner-inner)</td></tr>
<tr><td colspan="3">Traveling-wave form (2.21) with $\nu = \beta/2$, $\mu = \eta_0$</td></tr>
<tr><td rowspan="3">$B < 0$</td><td>$2 < \beta < \alpha + 1$
(fixed fronts)</td><td>$u \sim C(a-x)^{\frac{2(1-p)}{(\beta-2)}}$ as $t \to 0$ for $\sigma < x < a$
(cf. (4.6) and (4.7))</td></tr>
<tr><td colspan="2">Similarity form (2.16) with $\gamma = 1$, $\omega = 1/p$ (inner)</td></tr>
<tr><td>$\beta < \min(2, \alpha + 1)$

(moving fronts)</td><td>$s(t) - a \sim -\left(\frac{-B}{A}\right)^{\frac{1}{p}} t^{\frac{1}{p}}$ as $t \to 0^+$
and $\Lambda(t) \sim pA\left(\frac{-B}{A}\right)^{\frac{p-1}{p}} t^{\frac{p-1}{p}}$ as $t \to 0^+$
(cf. (3.5))</td></tr>
<tr><td colspan="3">Similarity form (2.16) with $\gamma = 1$, $\omega = 1/p$ (inner)</td></tr>
</table>

4. The case $p < 2/\beta$. In this case (2.2) is to be solved subject to the mass conservation condition

$$u_0^\alpha \frac{du_0}{dx} + u_0^\beta \frac{du_1}{dx} \to 0 \quad \text{as } x \to \pm a\,. \tag{4.1}$$

The regimes $p < \alpha + 1$, $p > \alpha + 1$, and $p = \alpha + 1$ are summarized in Tables 2 and 3. We first consider the regime $\beta < \alpha + 1$, $p < 2/\beta$. Here the second term of the left-hand side of (2.2) dominates as $x \to a^-$, so that $u_1(x) \sim B$ as $x \to a$, where the constant B, which may take either sign, is determined by solving (2.2) subject to (4.1). The expansion (2.1), which becomes

$$u \sim A(a-x)^p + Bt \quad \text{as } t \to 0^+\,, \tag{4.2}$$

is nonuniform as $x \to a$. For $B > 0$, in $x < a$ the local expansion (4.2) is of the similarity form (2.16) with $\gamma = 1$, $\omega = 1/p$, simply giving (2.17). In the fixed-front cases $2 \le \beta < \alpha + 1$ this completes the small-time structure. For the moving-front cases, however, this structure is clearly incomplete since (2.17) does not become zero, and a further inner-inner region is needed; see Table 2.

We now turn to the case $B < 0$, starting with the moving-front regimes. Since $\beta < \alpha + 1$, we lie in class (B) or (C) of section 3, so in either case a contracting front occurs, whereby

$$s(t) - a \sim -\left(\frac{-B}{A}\right)^{\frac{1}{p}} t^{\frac{1}{p}} \quad \text{as } t \to 0^+\,, \tag{4.3}$$

with the local expansion (4.2) holding right up to the front, so that

$$\Lambda(t) \sim pA\left(\frac{-B}{A}\right)^{\frac{p-1}{p}} t^{\frac{p-1}{p}} \quad \text{as } t \to 0^+ \tag{4.4}$$

TABLE 3

The non-PME balance $p < 2/\beta$, $\beta \geq \alpha + 1$. The α and β terms in (1.1) dominate the left-hand side of the front. An immediate change occurs, with the interface advancing in the moving-front cases.

$\beta > \alpha + 1$	$\alpha + 1 < \beta < 2\alpha$ (fixed fronts)	$U(t) \sim h(0) t^{\frac{1}{(\beta-\alpha)}}$ as $t \to 0^+$
		$(h(0) = ((\beta-\alpha)/(\beta-\alpha-1))^{1/(\beta-\alpha)})$
No waiting time	Similarity form (2.16) with $\omega = 1/(\beta-\alpha)p$ (inner)	
	$\beta > 2\alpha$ (moving fronts)	$s(t) - a \sim \eta_0\, t^{\frac{\beta}{2(\beta-2)}}$ as $t \to 0^+$
	Similarity form (2.4) (inner-inner)	
$\beta = \alpha + 1$	$\alpha > 1$ (fixed fronts)	$U(t) \sim t \ln(1/t)$ as $t \to 0^+$
	Similarity form $u \sim t \ln(1/t) h\big((x-a)/t^{\frac{1}{p}} \ln^{\frac{1}{p}}(1/t)\big)$, see (2.16) (inner)	
No waiting time	$\alpha = 1$, $\beta = 2$ (moving fronts)	$s(t) - a \sim t \ln^{\frac{1}{2}}(1/t)$ as $t \to 0^+$
	Similarity form $u \sim t \ln(1/t) f\big((x-a)/t \ln^{\frac{1}{2}}(1/t)\big)$ (inner-inner)	
	$\alpha < 1$ (moving fronts)	$G^\alpha(Z) \sim \alpha(-Z)$ as $Z \to 0^-$ (inner-inner-inner)
	Traveling-wave form (2.19) with $s(t) - a = \eta_0 t^{\frac{\beta}{2}} \ln^{\frac{\beta}{2}}(1/t)$	

holds in (3.5). In the fixed-front regime $\beta > 2$ (for which $p < 2/\beta$ evidently implies $p < 1$), the situation is very different. Here (4.2) holds in $x < \sigma(t)$, where we introduce $\sigma(t)$ such that

$$\sigma(t) - a \sim -\left(\frac{-B}{A}\right)^{\frac{1}{p}} t^{\frac{1}{p}} \quad \text{as } t \to 0^+, \tag{4.5}$$

as in (4.3), and setting $\zeta = x - \sigma(t)$ leads to the "traveling-wave" balance

$$u - u_\infty(t) \sim u^\beta \frac{\partial^2 u}{\partial \zeta^2}, \quad \begin{matrix} u \to u_\infty(t) & \text{as } \zeta \to +\infty, \\ u \sim \Lambda(t)(-\zeta) & \text{as } \zeta \to -\infty, \end{matrix} \tag{4.6}$$

where $\Lambda(t)$ in the second matching condition is again given by (4.4) and $u_\infty(t)$ is to be determined as part of the solution of (4.6) (from (4.4), it follows that the relevant scalings in (4.6) are $\zeta \propto t^{\frac{\beta(1-p)}{(\beta-2)p}}$, $u, u_\infty \propto t^{\frac{2(1-p)}{(\beta-2)p}}$). In $\sigma < x < a$ we have $\partial u/\partial t = 0$ at leading order, subject (by (4.4)) to

$$u = \left(\frac{(\beta-1)(\beta-2)p^2A^2}{2}\right)^{-\frac{1}{\beta-2}} \left(\frac{-B}{A}\right)^{\frac{2(1-p)}{(\beta-2)p}} t^{\frac{2(1-p)}{(\beta-2)p}} \quad \text{at } x = \sigma$$

and hence, by (4.5),

$$u \sim \left(\frac{(\beta-1)(\beta-2)p^2A^2}{2}\right)^{-\frac{1}{\beta-2}} (a-x)^{\frac{2(1-p)}{(\beta-2)}} \quad \text{as } t \to 0 \text{ for } \sigma < x < a, \tag{4.7}$$

with $a - x = O(t^{\frac{1}{p}})$. A front thus attempts to retreat, leaving behind a power-law profile (4.7) with $p' = 2(1-p)/(\beta-2)$; since $p' > 2/\beta$ for $p < 2/\beta$, the small-time response at $x = a$ generated by (4.7) is as described in section 5 by substituting p' for p, with $p' = 2/\alpha$ corresponding to $\beta = (1-p)\alpha + 2$. Note that for $p = 2/\beta$ we have

TABLE 4

The PME balance $2/\beta < p < 2/\alpha$, $\beta < \alpha$, the α term being dominant in (1.1) in the left-hand side of the front. There is no waiting-time behavior, as for PME.

$\beta < \alpha$, $p < 2/\alpha$	$\beta/2 < \alpha < \beta$, $\beta \geq 2$ (fixed fronts)	$U(t) \sim f(0)t^{p/(2-\alpha p)}$ as $t \to 0^+$
	Similarity form (2.5) with $\gamma = p/(2-\alpha p)$ (inner)	
	$\alpha \leq \beta/2$ (moving fronts)	$s(t) - a \sim \eta_0 t^{1/(2-\alpha p)}$ as $t \to 0^+$
	Similarity form (2.5) with $\gamma = p/(2-\alpha p)$ (inner)	
No waiting time	$\alpha < 1$ (moving fronts)	$g^\beta(z) \sim \frac{\beta^2}{2(2-\beta)}(-z)^2$ as $z \to 0^-$
	Traveling-wave form (2.21) with $\nu = 1/(2-\alpha p)$, $\mu = \eta_0$ (inner-inner)	

TABLE 5

The PME balance $p > \max(2/\alpha, 2/\beta)$. "Global" waiting-time behavior occurs for $0 < t < t_w$, and the local form (1.2) persists for $0 < t \leq t_w$. In the fixed front cases ($2 \leq \beta < 2\alpha$) a discontinuity at $x = a$ will emerge at t_w, while in the remaining (moving-front) regimes the interfaces can be expected to move on arrival of a "shock" (essentially as in [27] for PME; see also [22]).

$p > \max(2/\beta, 2/\alpha)$	$2 \leq \beta < 2\alpha$ (fixed fronts)	(1.10) holds
"Global" waiting time	$\alpha \leq \beta/2$ (moving fronts)	arrival of a "shock" at $t = t_w$

TABLE 6

The PME balance $p = 2/\alpha$, $\beta > \alpha$. A local waiting-time solution is available as for PME. This case differs from the PME in that fixed-front cases arise (for $\alpha > 1$ with $2 \leq \beta < 2\alpha$; see Figure 1).

$p = 2/\alpha$ "Local" waiting time	Fixed and moving fronts	$t_w \leq \frac{\alpha}{2(\alpha+2)A^\alpha}$
	Separable form $u \sim (a-x)^{\frac{2}{\alpha}} f(t)$ of (1.6) as $x \to a^-$ (inner)	

formally that $p' = 2/\beta$; the corresponding regime is that described in section A.5 of the appendix.

We shall not discuss here either of the borderline cases, $\beta = 2$ (in which we expect a profile that is exponentially small in $a - x$ to be left behind; cf. (4.7)) and $B = 0$ (which, in the moving-front regimes, separates advancing and retreating fronts and is particularly awkward since it requires the investigation of $u_2(x)$, the $O(t^2)$ correction term in the small-t expansion (2.1); this may lead to further subcases (and a further refined borderline), depending on the sign of u_2 at $x = a$). In case (C) of section 3 the nongeneric local solution (3.3) corresponds to the borderline behavior when $B = 0$.

5. The case $p > 2/\beta$. The second term on the left-hand side of (2.2) is negligible in this limit, thus giving the local expansion

$$u(x,t) \sim A(a-x)_+^p + p\Big((\alpha+1)p - 1\Big)A^{\alpha+1}(a-x)_+^{(\alpha+1)p-2}t \quad \text{as } x \to a\,.$$

The sign of $p - 2/\alpha$ can therefore be expected to determine whether there is a waiting time or not. The results in this regime are summarized in the Tables 4–6.

6. The critical case $p = 2/\beta$, $\alpha < \beta$. In this borderline case for the exponent p the first and second terms on the left-hand side of (2.2) are the same size as $x \to a^-$,

giving as the dominant balance the equation

$$u_1 + 2A^\beta(a-x)\frac{du_1}{dx} - A^\beta(a-x)^2\frac{d^2u_1}{dx^2} \sim A^{\alpha+1}\frac{2(2(\alpha+1)-\beta)}{\beta^2}(a-x)^{\frac{2(\alpha+1-\beta)}{\beta}} \quad \text{as } x \to a^-\,. \tag{6.1}$$

This is a nonhomogeneous Euler equation, with complementary functions of the form $(a-x)^m$, where m satisfies the quadratic equation $1 - A^\beta(m^2+m) = 0$. We specify the solution to (2.2) uniquely by excluding the negative root for m (required for the no-flux condition (4.1) to hold), leaving the acceptable root

$$m = \frac{1}{2}\left((1+4A^{-\beta})^{\frac{1}{2}} - 1\right)\,. \tag{6.2}$$

Defining $q = \frac{2(\alpha+1-\beta)}{\beta}$, then, provided $q \neq m$, a particular integral to (6.1) satisfies

$$u_1(x) \sim \frac{2A^{\alpha+1}(2\alpha+2-\beta)}{\beta^2(1-A^\beta(q^2+q))}(a-x)^q\,; \tag{6.3}$$

the dominant term in the local expansion for u_1 is thus given by (6.3) if $q < m$ and by $u_1 \sim B(a-x)^m$ for some constant B if $q > m$, with m given by (6.2). Since $q < p$, the outer expansion is nonuniform in this regime ($m < q < p$), and it follows that waiting-time behavior does not occur.

In the forthcoming analysis we let $A_1 \equiv \left(\beta^2/2(2\alpha+2-\beta)\,(\alpha+1-\beta)\right)^{\frac{1}{\beta}}$, so that q satisfies $1 - A_1^\beta(q^2+q) = 0$. There are various possibilities depending on the sign of $m-q$, and these are expressed in terms of A and A_1 when $\beta < \alpha+1$. We have the following cases.

(I) $A > A_1$ for $\alpha < \beta < \alpha+1$ ($0 < m < q$). Here the term on the right-hand side of (6.1) is locally negligible. Hence

$$u_1(x) \sim B(a-x)^m \quad \text{as } x \to a^-, \tag{6.4}$$

where m is given by (6.2) and B is a constant which may take either sign and which can be found by solving (2.2) (subject to the appropriate boundary conditions). In this case the solution in the inner region close to $x = a$ is of the similarity form (2.10) with $\gamma = 2/(2-\beta m)$ and with

$$f(\eta) \sim A(-\eta)^{\frac{2}{\beta}} + B(-\eta)^m \quad \text{as } \eta \to -\infty, \tag{6.5}$$

both terms here being needed to specify $f(\eta)$, as is to be expected since some time-dependent forcing (namely, the B term) is needed if the solution to (1.7) is not to be simply $u \sim u_0(x)$; note that $f(\eta) = A(-\eta)^{2/\beta}$ is an exact solution of (2.7), and this implies the presence of a borderline corresponding to the special case $B = 0$. Such similarity reductions of (1.7) are the subject of the appendix. We start by discussing the case $B > 0$. In the moving-front regime $\beta < 2$, $f(\eta)$ satisfies (2.7), with (2.9) holding for $0 < \eta < \eta_0$ and in $\eta < 0$ we thus have conditions

$$f^{\frac{\beta}{2}}\frac{d}{d\eta}\left(f - \frac{\beta}{2}\eta\frac{df}{d\eta}\right) = -\sqrt{\frac{2-\beta}{2}}\left(f - \frac{\beta}{2}\eta\frac{df}{d\eta}\right) \quad \text{at } \eta = 0^- \tag{6.6}$$

TABLE 7

The parameter regime $p = 2/\beta$, $\alpha < \beta$, with $A > A_1$ for $\beta < \alpha + 1$ and with $B > 0$ in (6.5). In the moving-front cases the interface immediately advances.

$A > A_1$ for $\alpha < \beta < \alpha + 1$, $B > 0$	$\beta \geq 2$ (fixed fronts)	$U(t) \sim f(0) t^{\frac{2}{(2-\beta m)}}$ as $t \to 0^+$
	Similarity form (2.10) with $\gamma = 2/(2-\beta m)$ (inner)	
	$\beta < \min(2, 2\alpha)$ (moving fronts)	$s(t) - a \sim \eta_0 \, t^{\frac{\beta}{2(\beta-2)}}$ as $t \to 0^+$
	Similarity form (2.10) with $\gamma = 2/(2-\beta m)$ (inner)	
No waiting time	$2\alpha \leq \beta < 2$ (moving fronts)	$g^\alpha(z) \sim \eta_0 \frac{\alpha\beta}{(2-\beta m)}(-z)$, as $z \to 0^-$
	Traveling-wave form (2.21) with $\nu = \beta/(2-\beta m)$ $\mu = \eta_0$ (inner-inner)	

(see (3.11)) and (6.5). Thus as $\eta \to -\infty$ we have in effect two boundary conditions (both A and B are specified in (6.5), leaving a single eigenmode, $(-\eta)^{-(m+1)}$), while at the origin the eigenmodes are $(-\eta)^0$, $(-\eta)^1$, and $(-\eta)^{2/\beta}$; a relation between the first two of these is furnished by (6.6), while the third gives a nonanalyticity whereby u at $t = 0^+$ has a term $C_0(a-x)^{2/\beta}$ as $x \to a^-$ (C_0 being the coefficient of the $(-\eta)^{2/\beta}$ eigenmode, determined by the solution $f(\eta)$ of (2.7) and not in general equal to A), but no comparable term in $x > a$. Assuming that for $t > 0$ we have $u_a > 0$, then, by balancing the terms on the right-hand side, (1.1) implies that u contains a term $C(t)(a-x)^{2/\beta}$ as $x \to a^-$, with this nonanalyticity decaying according to (cf. (3.8))

$$\frac{dC}{dt} = -u_a^{\alpha-\beta} C \,, \; C = C_0 \quad \text{at } t = 0 \,; \tag{6.7}$$

we have

$$u_a(t) \sim f(0) t^{\frac{2}{(2-\beta m)}} \quad \text{as } t \to 0^+ \tag{6.8}$$

and, since $m < 2/\beta < q$, it follows that $-1 < 2(\alpha - \beta)/(2 - \beta m) < 0$ and hence that (6.7) behaves in an appropriate fashion as $t \to 0^+$. It is worth repeating the more general point that (unlike (1.6) with $u > 0$) equation (1.1) does not instantaneously smooth nonanalyticities in u. We note that in the special case $A = A^* \equiv (\beta^2/2(2-\beta))^{1/\beta}$ (for this value of A to lie in the current regime we require $\beta < 2\alpha$) the relevant similarity solution can be constructed explicitly; we have $m = (2-\beta)/\beta$ and the solution (2.9) applies for all η, with

$$\eta_0 = \frac{\beta B}{2A^*}, \quad C_0 = A^* \tag{6.9}$$

(see (A.11)). Table 7 summarizes the front behavior. The analysis of the case $B < 0$ follows from that of the appendix, and can be summarized as follows. In the moving-front regime $\beta < 2$ the self-similarity form (2.7) applies and for $A > A^* \equiv (\beta^2/2(2-\beta))^{\frac{1}{\beta}}$ we have a contracting front with local behavior of the form (A.3) (with $\eta_0 < 0$), while for $A_1 < A < A^*$ we have a switch in the structure of the inner solution whereby (A.4) provides its local behavior (note that, in the notation of the appendix, $A_1 > A_2 \equiv (\beta^2/2(\beta+2))^{\frac{1}{\beta}}$ holds in this regime, whatever the sign of $2-\beta$). In the fixed-front regime $\beta \leq 2$ with $B < 0$ we recover the local behavior (A.4) for any $A > A_1$; see the analysis of section A.5. Since in (A.4) we have $\hat{A} < A_2 < A_1$, this is not the end of the story in these final two cases, however; $\hat{A}$ lies in the regime

TABLE 8

The parameter regime $p = 2/\beta$, $\alpha < \beta$, for $\alpha + 1 \le \beta$ or $A < A_1$ with $\alpha < \beta < \alpha + 1$. There is no waiting-time behavior, and, in particular, the interface immediately advances in the moving-front cases.

$\alpha + 1 \le \beta$ or $A < A_1$ with $\alpha < \beta < \alpha + 1$	No waiting time	$\alpha < \beta < 2\alpha$, $\beta \ge 2$ (fixed fronts)	$U(t) \sim f(0) t^{\frac{1}{(\beta-\alpha)}}$ as $t \to 0^+$
		Similarity form (2.4)	(inner)
		$\alpha < \beta \le 2\alpha$, $\beta > 2$ (moving fronts)	$s(t) - a \sim \eta_0\, t^{\frac{\beta}{2(\beta-2)}}$ as $t \to 0^+$
		Similarity form (2.4)	(inner)

TABLE 9

The parameter regime $p = 2/\beta$, $\alpha < \beta$, for $A = A_1$ with $\alpha < \beta < \alpha + 1$. There is no waiting-time behavior as in case (II) *above.*

$A = A_1$ for $\alpha < \beta < \alpha + 1$	$\beta \ge 2$ (fixed fronts)	$U(t) \sim f(0) t^{\frac{1}{(\beta-\alpha)}} \ln^{\frac{1}{\beta-\alpha}}(1/t)$ as $t \to 0^+$ with $f(0) > 0$	
	Similarity form (2.6) with $\Omega(t) = t^{\frac{\beta}{2(\beta-\alpha)}} \ln^{\frac{\beta}{2(\beta-\alpha)}}(1/t)$		(inner)
No waiting time	$\beta < 2$, $\beta \le 2\alpha$ (moving fronts)	$s(t) - a \sim \eta_0 t^{\frac{\beta}{2(\beta-\alpha)}} \ln^{\frac{\beta}{2(\beta-\alpha)}}(1/t)$ as $t \to 0^+$	
	Similarity form (2.6) with $\Omega(t) = t^{\frac{\beta}{2(\beta-\alpha)}} \ln^{\frac{\beta}{2(\beta-\alpha)}}(1/t)$		(inner)
	$2\alpha < \beta < 2$ (moving fronts)	$G^\alpha(Z) \sim \alpha(-Z)$ as $Z \to 0^-$	(inner-inner)
	Traveling-wave form (2.19) with $s(t) - a = \eta_0 t^{\frac{\beta}{2(\beta-\alpha)}} \ln^{\frac{\beta}{2(\beta-\alpha)}}(1/t)$		

governed by (II) below, so there is in the small-time behavior a further (inner-inner) region described by the similarity solution (2.4) (in the inner region (2.10) holds with $\beta\gamma/2 = \beta/(2-\beta m) < \beta/2(\beta-\alpha)$ because $m < q$ here, so (as required) the inner-inner region is much narrower than the inner one for $t \ll 1$).

(II) $\alpha + 1 \le \beta$ or $A < A_1$ with $\alpha < \beta < \alpha + 1$ ($q < m$). See Table 8.

(III) $A = A_1$ for $\alpha < \beta < \alpha + 1$ ($q = m$). See Table 9.

7. The critical case $p = 2/\beta$, $\beta < \alpha$. The dominant local balance within this section is (1.7), and the dominant balance in (2.2) is the homogeneous version of (6.1). Hence u_1 is of the form

$$u_1(x) \sim B(a - x)^m \quad \text{as } x \to a^-, \tag{7.1}$$

where m is given by (6.2), and the constant B is determined by the boundary value problem for (2.2) implied by (7.1) with (6.2). Whether (7.1) is locally negligible compared to (1.2) depends on whether m is greater or less than p ($= 2/\beta$), and the critical value of A is therefore given by A_2. We accordingly distinguish three cases.

(I) $A > A_2$. Here $m < 2/\beta$, so (2.1) is nonuniform and the small-time behavior is as outlined in case (II) of section 6 except that, since $\hat{A} < A_2$ holds in (A.4), in the relevant cases the local behavior switches into the regime (II) below, corresponding to waiting-time behavior; see also the appendix. The advancing-front cases can here lie only in the range given in (2.26), so the front behaves as

$$s(t) - a \sim \eta_0 t^{\frac{\beta}{(2-\beta m)}} \quad \text{as } t \to 0^+,$$

and a further "traveling-wave" region (i.e., (vi) of section 2) is not present.

(II) $A < A_2$. Here $m > 2/\beta$, and (1.2) is valid as $x \to a^-$ for $t > t_w$, so that "global" waiting-time behavior occurs.

(III) $A = A_2$. In this borderline case we have $m = 2/\beta$ and the appropriate matching condition thus takes the form

$$u \sim (a-x)^{\frac{2}{\beta}} (A_2 + Bt)\,. \tag{7.2}$$

There is no relevant separable solution of the balance (1.7), but adopting the "near-separable" form

$$u = (a-x)^{\frac{2}{\beta}}\, \Phi(\xi, t), \qquad \xi = -\ln(a-x)\,, \tag{7.3}$$

it can be inferred that $\Phi(\xi, t) \sim A_2 + \Psi(\xi, t)$ as $t \to 0^+$ with

$$\Psi = \frac{Bt}{1 - \frac{B(\beta+2)}{A_2(\beta+4)}\xi t}\,, \tag{7.4}$$

which for $B > 0$ blows up as $\xi t \to \frac{A_2(\beta+4)^-}{B(\beta+2)}$. Equations (7.3) and (7.4) furnish the matching condition

$$u \sim (a-x)^{\frac{2}{\beta}} \left(A_2 + \frac{A_2(\beta+4)}{(\beta+2)} \frac{1}{\frac{A_2(\beta+4)}{B(\beta+2)t} + \ln(a-x)} \right) \quad \text{as } x \to a^-\,,\ t \to 0^+ \tag{7.5}$$

for the final (inner-inner) region for $B > 0$ in which $x - a$ is exponentially small, with the asymptotic solution essentially taking the form

$$u \sim e^{-\frac{2A_2(\beta+4)}{B\beta(\beta+2)t}} f((x-a)e^{\frac{A_2(\beta+4)}{B(\beta+2)t}}), \tag{7.6}$$

where $f(\eta)$ is a similarity solution of (1.7) with, in view of (7.5),

$$f(\eta) \sim A_2(-\eta)^{\frac{2}{\beta}} + \frac{A_2(\beta+4)}{(\beta+2)} \frac{(-\eta)^{\frac{2}{\beta}}}{\ln(-\eta)} \quad \text{as } \eta \to -\infty \tag{7.7}$$

(cf. section A.3 of the appendix, in particular (A.16)). We note that, as is typical in such borderline regimes, the relevant similarity reduction could in fact include multiplicative terms which are algebraic in t, as well as the exponentials in (7.6); calculation of these terms requires higher-order matching which we shall not pursue here ((7.6) nevertheless embodies the dominant time dependence of the relevant exponentially small region).

When $B < 0$, no blow-up occurs in (7.4), the inner-inner region is absent, and waiting-time behavior ensues. The local behavior is summarized in Table 10.

This borderline regime $A = A_2$ could be refined further by an analysis of the case $B = 0$. We note that we are assuming throughout that the correction terms to (1.2) are at least algebraically smaller in $(x - a)$ and so do not influence the logarithmic correction terms in (7.5), for example.

8. The critical case $p = 2/\beta$, $\beta = \alpha$. This is the most delicate case of all, lying at the intersection of other borderlines (see Figure 1), and is therefore worth recording in detail. We also observe that, as in section 7, the only moving fronts

TABLE 10

The critical regime $p = 2/\beta$, $\beta < \alpha$ with $A = A_2$. This borderline case is noteworthy in that in the moving-front regime the interface immediately advances (though exponentially slowly) for $B > 0$, while for $B < 0$ no change in the leading-order local behavior occurs, though a logarithmic correction term immediately enters; as such this represents a natural borderline with the "switch" in local behavior which occurs for $A > A_2$.

<table>
<tr><td rowspan="3">$A = A_2$</td><td rowspan="3">$B > 0$</td><td>$\beta \geq 2$
(fixed fronts)</td><td>$U(t) \sim f(0) e^{-\frac{2A_2(\beta+4)}{B\beta(\beta+2)t}}$ as $t \to 0^+$
(inner-inner)</td></tr>
<tr><td colspan="2">Similarity form (2.6) with $\Omega(t) = e^{-\frac{A_2(\beta+4)}{B(\beta+2)t}}$</td></tr>
<tr><td>Near-separable form (7.3)
(inner)</td><td>No waiting time</td><td>$\beta < 2$, $\beta \leq 2\alpha$
(moving fronts)</td><td>$s(t) - a \sim \eta_0\, e^{-\frac{A_2(\beta+4)}{B(\beta+2)t}}$ as $t \to 0^+$
(inner-inner)</td></tr>
<tr><td colspan="3">Similarity form (2.6) with $\Omega(t) = e^{-\frac{A_2(\beta+4)}{B(\beta+2)t}}$</td></tr>
<tr><td></td><td>$B < 0$
Local waiting time</td><td colspan="2">$u \sim A_2(a-x)^{\frac{2}{\beta}} \left(1 + \frac{(\beta+4)}{(\beta+2)\ln(a-x)}\right)$ as $x \to a^-$
for $0 < t < t_w$ (inner)</td></tr>
</table>

which are possible are of the form (2.26), when $p > 1$ ($\beta < 2$); otherwise, this regime lies in the fixed-front domain.

The dominant balance in (2.2) is again given by (6.1), i.e.,

$$u_1 + 2A^\beta(a-x)\frac{du_1}{dx} - A^\beta(a-x)^2\frac{d^2u_1}{dx^2} \sim A^{\beta+1}\frac{2(\beta+2)}{\beta^2}(a-x)^{\frac{2}{\beta}} \quad \text{as } x \to a^- . \tag{8.1}$$

As before we let m denote the positive solution of the quadratic equation $1 - A^\beta(m^2 + m) = 0$. In this case q from section 6 is $2/\beta$, and the critical value of A is given by A_2, which in this case coincides with A_1. The following three subcases arise.

(I) $A > A_2$. Here $m < 2/\beta$, and $u_1(x) \sim B(a-x)^m$ as $x \to a^-$. The behavior is identical to that given in case (I) of section 7; in the cases in which (A.4) applies, the local solution subsequently evolves as in subcase (II) below with $\hat{A}$ replacing A.

(II) $A < A_2$. Here $m > 2/\beta$ and (6.3) holds with $q = 2/\beta$, and the matching condition

$$u \sim A(a-x)^{\frac{2}{\alpha}} \left(1 + \frac{A^{\alpha+1}}{A_2^\alpha - A^\alpha} t\right) \quad \text{as } x \to a^-$$

thus applies. Aiming to extend this analysis beyond the small t regime, we seek a separable solution to (1.1) with $\beta = \alpha$ of the form (2.4), finding that

$$\left(A_2 f^{-(\alpha+1)} - f^{-1}\right)\frac{df}{dt} = 1 \tag{8.2}$$

so that

$$\frac{1}{\alpha}(f/A_2)^{-\alpha} + \ln f = -t + \frac{1}{\alpha}(A/A_2)^{-\alpha} + \ln A \tag{8.3}$$

for $t \leq T$, where T is given by $f(T) = A_2$, i.e.,

$$T = \ln(A/A_2) + \left((A/A_2)^{-\alpha} - 1\right)/\alpha . \tag{8.4}$$

Thus the coefficient f of the local solution automatically increases towards the borderline A_2 between waiting-time ($f < A_2$) and non-waiting-time ($f > A_2$) scenarios, and the solution to (8.2) ceasing to exist at $t = T$, with

$$f \sim A_2(1 - (2(T-t)/\alpha)^{\frac{1}{2}}) \quad \text{as } t \to T^- . \tag{8.5}$$

This again represents a local waiting-time solution with $t_w \leq T$; unlike (7.4), however, the solution (8.3) does not blow up as the local waiting time $t = T$ is reached, and a further set of scalings needs to be considered for $t - T$ small if we are to describe the "local" waiting-time scenario in which $t_w = T$.

Again we write

$$u = (a-x)^{\frac{2}{\alpha}}\Phi(\xi,t), \quad \xi = -\ln(a-x) . \tag{8.6}$$

In view of the behavior (8.5) near $t = T$ we now seek a "backward" self-similar solution of the form

$$\Phi \sim A_2 + \Psi(\xi,t), \ \Psi(\xi,t) = (T-t)^{\frac{1}{2}}F_-(\xi(T-t)^{\frac{1}{2}}) \quad \text{as } t \to T^- \tag{8.7}$$

to give (after some manipulations and again assuming the initial data does not contain correction terms to (1.2) which are only logarithmically smaller in $(a-x)$)

$$\frac{dF_-}{d\zeta} = \frac{(\alpha+2)}{(\alpha+4)A_2}\left(F_-^2 - \frac{2A_2^2}{\alpha}\right) . \tag{8.8}$$

Matching to (8.5) requires $F_- \to -\sqrt{2/\alpha}A_2$ as $\zeta \to +\infty$, while we require F_- to behave as $1/\zeta$ as $\zeta \to 0^+$ in order that (8.7) have suitable behavior at $t = T$; this finally gives $F_- = -A_2\sqrt{2/\alpha}\coth((\alpha+2)/(\alpha+4)\sqrt{2/\alpha}\zeta)$ so that at $t = T$

$$\Phi \sim A_2\left(1 + \frac{(\alpha+4)}{(\alpha+2)\ln(a-x)}\right) \quad \text{as } x \to a^- \tag{8.9}$$

as in section 7 (III) for $B < 0$; see Table 10. In view of (8.9), we can now continue the local solution into $t > T$ by seeking a "forward" self-similar solution Ψ of the form

$$\Psi(\xi,t) = (t-T)^{\frac{1}{2}}F_+(\xi(t-T))^{\frac{1}{2}} \quad \text{as } t \to T^+, \tag{8.10}$$

so that $F_+(\zeta)$ satisfies

$$\frac{dF_+}{d\zeta} = \frac{(\alpha+2)}{(\alpha+4)A_2}\left(F_+^2 + \frac{2A_2^2}{\alpha}\right),$$

and matching with (8.9) requires that

$$F_+(\zeta) = -\sqrt{\frac{2}{\alpha}}A_2\cot\left(\frac{(\alpha+2)}{(\alpha+4)}\sqrt{\frac{2}{\alpha}}\zeta\right) . \tag{8.11}$$

Hence $F_+ \to +\infty$ as $\zeta \to \pi\sqrt{\alpha/2}\,(\alpha+4)/(\alpha+2)$ and one more region is needed with scalings essentially of the form

$$u \sim e^{-\pi\sqrt{\frac{2}{\alpha}}\frac{(\alpha+4)}{(\alpha+2)}\frac{1}{(t-T)^{1/2}}}\phi\left((x-a)e^{\pi\sqrt{\frac{\alpha}{2}}\frac{(\alpha+4)}{(\alpha+2)}\frac{1}{(t-T)^{1/2}}}\right)$$

TABLE 11

The critical regime $p = 2/\beta$, $\alpha = \beta$ with $A < A_2$, in the local waiting-time case $t_w = T$ is available, for which (unlike the PME) the evolution past the waiting time can be tracked asymptotically.

$A < A_2$ Local waiting time	$\beta \geq 2$ (fixed fronts)	$U(t) \sim \phi(0) e^{-\pi \sqrt{\frac{2}{\alpha}} \frac{(\alpha+4)}{(\alpha+2)} \frac{1}{(t-t_w)^{1/2}}}$ as $t \to t_w^+$
	Similarity form (2.6) with $\Omega(t) = e^{-\pi \sqrt{\frac{\alpha}{2}} \frac{(\alpha+4)}{(\alpha+2)} \frac{1}{(t-t_w)^{1/2}}}$ (inner-inner)	
Separable form (2.4)	$\beta < 2$ (moving fronts)	$x - a \sim \eta_0 e^{-\pi \sqrt{\frac{\alpha}{2}} \frac{(\alpha+4)}{(\alpha+2)} \frac{1}{(t-t_w)^{1/2}}}$ as $t \to t_w^+$
(inner)	Similarity form (2.6) with $\Omega(t) = e^{-\pi \sqrt{\frac{\alpha}{2}} \frac{(\alpha+4)}{(\alpha+2)} \frac{1}{(t-t_w)^{1/2}}}$ (inner-inner)	

TABLE 12

The critical regime $p = 2/\beta$, $\alpha = \beta$ with $A = A_2$. No waiting time occurs, with the interface immediately advancing in the moving-front regime albeit exponentially slowly.

$A = A_2$ No waiting time	$\beta \geq 2$ ($p \leq 1$) (fixed fronts)	$U(t) \sim \phi(0) e^{-\frac{\pi}{2} \sqrt{\frac{2}{\alpha}} \frac{(\alpha+4)}{(\alpha+2)} \frac{1}{t^{1/2}}}$ as $t \to 0^+$
	Similarity form (2.6) with $\Omega(t) = e^{\frac{\pi}{2} \sqrt{\frac{\alpha}{2}} \frac{(\alpha+4)}{(\alpha+2)} \frac{1}{t^{1/2}}}$ (inner-inner)	
Near-separable form (8.6)	$\beta < 2$ ($p > 1$) (moving fronts)	$s(t) - a \sim \eta_0 e^{-\frac{\pi}{2} \sqrt{\frac{\alpha}{2}} \frac{(\alpha+4)}{(\alpha+2)} \frac{1}{t^{1/2}}}$ as $t \to 0^+$
(inner)	Similarity form (2.6) with $\Omega(t) = e^{\frac{\pi}{2} \sqrt{\frac{\alpha}{2}} \frac{(\alpha+4)}{(\alpha+2)} \frac{1}{t^{1/2}}}$ (inner-inner)	

(again, both here and in Table 12 there may be additional algebraic terms), giving a similarity solution of the form (2.6) to (1.7), with $\alpha = \beta$ and

$$\phi(\eta) \sim A_2 \left(1 + \frac{(\alpha+4)}{(\alpha+2)\ln(-\eta)}\right)(-\eta)^{\frac{2}{\alpha}} \quad \text{as } \eta \to -\infty. \tag{8.12}$$

A striking feature of this analysis is that, unlike existing analyses of the PME (but more like Hele–Shaw corners [25]), it is possible here to follow explicitly the local analysis through the waiting-time in the "local" case $t_w = T$, the behavior being summarized in Table 11.

(III) $A = A_2$ ($m = q = 2/\beta$). See Table 12.

9. Discussion. Our small-time results serve in particular to identify, and to provide concrete illustrations of, certain phenomena not shared by PME. Specifically, for (1.1) with $\beta < \min(2, \alpha + 1)$ we have the possibility (when $B < 0$) of (finite-slope) retreating fronts or of interfaces which instantaneously switch the coefficient in their local behavior. Those local solutions which switch necessarily have $p = 2/\beta$ in (1.2) with $A \in (A_2, A^*)$ and, when they are applicable, represent a (slightly diffuse) borderline between retreating fronts and solutions in which the local behavior remains unchanged for some finite time. It is worth remarking that, by choosing initial data which exhibit different power laws over different orders of magnitude of $a - x$, it is possible to construct in the relevant (α, β) regimes solutions in which, for example, the interface advances, retreats, and then advances again. We leave the behavior at the end of a waiting time (for example, whether the interface can then begin to retreat rather than advance) as an open problem.

It is worth highlighting that we have in a number of places alluded to the formal time-reversibility of (1.7) being violated by differences between advancing and retreating fronts in terms of the nature of the local singularity at the interface. Thus despite

its apparent time-reversibility, (1.7) does exhibit a clear arrow of time in the sense of there being a marked distinction between the two directions of interface motion; a specific example of this is given in section 1 (taking $t_2 = 2t_1$, $P(t_2 - t) = -P(t)$, one could distinguish a film of fluid-thickness profile being run forward from $t = 0$ from one being run backwards from $t = t_2$ when P is negative, but not when P is positive, for $0 < t < t_1$). Such phenomena are of relevance to other formally time-reversible evolution equations; see [5] for a discussion of certain related issues. Our final remark in this regard is that, as described in section 3, the finite-slope branch (3.5), say, can also be admissible for advancing fronts, but is then underspecified. This implies that an extra physical constraint is needed to specify the solution, perhaps relating the slope Λ to the interface speed $\dot{s}$ (as may occur in other contexts, such as capillary-driven spreading). It would be worthwhile to explore the circumstances under which such solutions may be pertinent to (1.1).

Appendix. The similarity solution (2.7).

A.1. Formulation. Here we discuss relevant boundary value problems for (2.7), the results being of independent interest in view of their direct applicability to (1.7). We impose the far-field condition (6.5) with $A > 0$ and B prescribed and with m given by (6.2). We defer the fixed-front case $\beta > 2$ to section A.5, first describing the various regimes for $\beta < 2$. The requirement that m be less than $2/\beta$ implies that $A > A_2$ must hold, where $A_2 \equiv (\beta^2/2(\beta+2))^{\frac{1}{\beta}}$. The expression (6.5) then represents two boundary conditions, and imposing (6.6), with (2.9) holding in $\eta > 0$, thus yields a correctly specified system provided a solution exists in which $f(0) > 0$; the free-boundary location, $\eta_0 > 0$, in (2.9) is then given by

$$\eta_0 = \frac{(2(2-\beta)f^\beta(0))^{\frac{1}{2}}}{\beta}\,. \tag{A.1}$$

However, it is also possible for $f(\eta)$ to attain zero at $\eta = \eta_0 < 0$ with

$$f \to 0, \quad f^\beta \frac{d^2 f}{d\eta^2} \to 0 \quad \text{as } \eta \to \eta_0^- \tag{A.2}$$

leading to a correctly specified free-boundary problem, the solution of which has the local form

$$f(\eta) \sim \lambda(\eta_0 - \eta) \quad \text{as } \eta \to \eta_0^-, \eta_0 < 0\,, \tag{A.3}$$

for some constant $\lambda > 0$, corresponding to (3.5) and containing the necessary two-degree-of-freedom λ and η_0. Finally, there is a third possibility with the required two degrees of freedom, namely, that f attains zero at $\eta_0 = 0$ (so that this case represents a waiting-time scenario) with

$$f \sim \hat{A}(-\eta)^{\frac{2}{\beta}} + \hat{B}(-\eta)^{\hat{m}} \quad \text{as } \eta \to 0^-, \tag{A.4}$$

where $0 < \hat{A} < A_2$, so that

$$\hat{m} = \frac{1}{2}\left((1 + 4\hat{A}^{-\beta})^{\frac{1}{2}} - 1\right) \tag{A.5}$$

satisfies $\hat{m} > 2/\beta$. The constants $\hat{A}$ and $\hat{B}$ provide the two degrees of freedom in (A.4), with the calculation of the dependence of $\hat{A}$ upon A requiring the solution

of the boundary value problem for $f(\eta)$. We emphasize that, while the interface waits in this case, the local behavior (A.3) differs from that of the initial data, with $\hat{A} < A_2 < A$.

Since it may seem mysterious that retreating fronts, whereby (A.3) holds, satisfy one less boundary condition than advancing ones (for which $df/d\eta = 0$ at $\eta = \eta_0$) it is worth also giving an explanation for this in the similarity ODE context (noting that the ODE (2.7) is of third order in η, whereas (1.1) is second order in x). The expression (6.5) represents two boundary conditions and (A.2) two more which, given that η_0 is unknown, gives a correctly specified problem for $\eta_0 < 0$. For $\eta_0 > 0$, however, the continuity conditions

$$[f]_-^+ = \left[\frac{2}{\beta}\frac{df}{d\eta} - \eta\frac{d^2f}{d\eta^2}\right]_-^+ = 0 \quad \text{at } \eta = 0 \tag{A.6}$$

lead to the degrees of freedom f_0, f_1, $C_\pm$ in

$$f \sim f_0 + f_1\eta + C_\pm|\eta|^{\frac{2}{\beta}} \quad \text{as } \eta \to 0^\pm\,.$$

From (2.9) it happens that $C_+ = 0$; it is the need also to determine C_- that requires the third condition at $\eta = \eta_0$.

A.2. Exact results. When $B = 0$ we have the exact (borderline, steady-state) solution

$$f(\eta) = A(-\eta)^{\frac{2}{\beta}}, \quad \eta_0 = 0\,, \tag{A.7}$$

which can play the role of separating advancing and retreating fronts. When $B \neq 0$, the rescaling

$$f(\eta) = |B|^{\frac{2}{2-\beta m}}\hat{f}(\hat{\eta})\,,\ \hat{\eta} = \eta/|B|^{\frac{\beta}{2-\beta m}} \tag{A.8}$$

enables us to set $B = \pm 1$ without loss of generality, with corresponding solutions $\hat{f}_\pm(\hat{\eta})$, where we might anticipate that

$$\hat{f}_+(0) > 0,\ \eta_0 = (2(2-\beta)\hat{f}_+^\beta(0)^{\frac{1}{2}})B^{\frac{\beta}{2-\beta m}}/\beta\,, \tag{A.9}$$

and

$$\hat{f}_-(\hat{\eta}_0) = 0,\ \eta_0 = |B|^{\frac{\beta}{2-\beta m}}\hat{\eta}_0 \tag{A.10}$$

for some $\hat{\eta}_0 < 0$, so that $B > 0$ gives an advancing front and $B < 0$ a retreating one; in view of the alternative scenario (A.4) this is far from clear a priori, a point to which we shall return in section A.4.

When $A = A^* \equiv \left(\beta^2/2(2-\beta)\right)^{\frac{1}{\beta}}$ we have $m = 2/\beta - 1$ and

$$f^\beta(\eta) = \frac{\beta^2}{2(2-\beta)}(\eta_0 - \eta)^2\,,\ \eta_0 = \frac{\beta B}{2A^*} \tag{A.11}$$

holds exactly, both for $B > 0$ (giving $\eta_0 > 0$) and for $B < 0$ (with $\eta_0 < 0$); in this special case we thus have $\lambda = 0$ in (A.3). Other cases may also be analytically tractable; for example, for $\beta = 1$, $m = 1/2$ (so that $A = 4/3$) we have $f(\eta) = \frac{4}{3}(-\eta)^2 + B(-\eta)^{\frac{1}{2}}$ for $B < 0$, so that $\eta_0 = -(3|B|/4)^{\frac{2}{3}}$.

A.3. Asymptotics. The two regimes $A \to \infty$ and $A \to A_2^+$ are amenable to asymptotic investigation. In the former case we have $m \to 0$ and the right-hand side of (2.7) dominates; the boundary conditions then imply that the leading-order solution is simply

$$f(\eta) \sim A(-\eta)^{\frac{2}{\beta}} + B\,, \tag{A.12}$$

the relevant η scaling being $-\eta = O((|B|/A)^{\frac{\beta}{2}})$. Hence for $B > 0$ we have $\eta_0 > 0$, given by (A.1) with $f(0) \sim B$, while for $B < 0$ a retreated front occurs with

$$\eta_0 \sim -\left(\frac{|B|}{A}\right)^{\frac{\beta}{2}}, \quad \lambda \sim \frac{2}{\beta}\left(\frac{A}{|B|}\right)^{\frac{\beta}{2}} |B|\,. \tag{A.13}$$

Now setting $A = A_2 + \varepsilon$, with $0 < \varepsilon \ll 1$, we have

$$m \sim \frac{2}{\beta} - \varepsilon\mu,$$

where $\mu = \beta^2/A_2^{\beta+1}(4+\beta)$, and for algebraic convenience it is helpful to use the rescaling invariance to set, without loss of generality, $|B| = 2\varepsilon$. Under the change of variables

$$f = (-\eta)^{\frac{2}{\beta}}(A_2 + \varepsilon G(\zeta)), \quad \zeta = -\varepsilon\ln(-\eta)\,,$$

appropriate to the outer region $\zeta = O(1)$, we find at leading order that

$$0 = \frac{\beta}{A_2} G_0 \frac{dG_0}{d\zeta} - \frac{4+\beta}{\beta} A_2^{\beta} \frac{d^2 G_0}{d\zeta^2}\,.$$

Requiring $G_0(\zeta) \sim 1 + 2\frac{B}{|B|}e^{\mu\zeta}$ as $\zeta \to -\infty$ yields

$$\frac{dG_0}{d\zeta} = \frac{\mu}{2}(G_0^2 - 1)$$

and then

$$G_0(\zeta) = \coth\left(\frac{\mu}{2}(-\zeta)\right) \quad \text{if } B > 0\,, \tag{A.14}$$

$$G_0(\zeta) = \tanh\left(\frac{\mu}{2}(-\zeta)\right) \quad \text{if } B < 0\,. \tag{A.15}$$

For $B > 0$ (i.e., $B = 2\varepsilon$), the inner scaling is $\eta = O(1)$ (i.e., $\zeta = O(\varepsilon)$) and at leading order we obtain the full problem, except that the far-field condition (6.5) is replaced by

$$f_0(\eta) \sim (-\eta)^{\frac{2}{\beta}}\left(A_2 + \frac{2}{\mu\ln(-\eta)}\right) \quad \text{as } \eta \to -\infty \tag{A.16}$$

in order to match with (A.14); we conjecture that this condition leads to a solution with $\eta_0 > 0$. Finally, for $B < 0$ (i.e., $B = -2\varepsilon$) the outer expansion does not break down (unlike (A.14), the expression (A.15) does not become singular), and by taking the limit $\zeta \to +\infty$ ($\eta \to 0^-$) we obtain

$$f(\eta) \sim (A_2 - \varepsilon)(-\eta)^{\frac{2}{\beta}}\,, \tag{A.17}$$

consistent with our third scenario (A.4) with $\hat{A} \sim A_2 - \varepsilon$; we have $\hat{m} \sim \frac{2}{\beta} + \varepsilon\mu$ and, by (A.17), $\hat{B} \sim 2\varepsilon$.

A.4. Parameter regimes. The above results enable us to piece together as the likely picture the following scenario. The problem evidently warrants further study, not least by rigorous techniques, in order to confirm the behavior is indeed as outlined here. Further corroborating evidence for what follows can be obtained by an analysis of the limit $A \to A^*$; one then finds in particular that, setting $A = A^* - \varepsilon$ with $0 < \varepsilon \ll 1$, one has for $B < 0$

$$\hat{A} \sim \varepsilon h_\infty \quad \text{for } 0 < \beta < 1\,, \quad \hat{A} \sim \varepsilon^{\frac{1}{2-\beta}} h_\infty \quad \text{for } 1 < \beta < 2\,, \tag{A.18}$$

for some positive constant $h_\infty(\beta)$. The reasons for the transition in (A.18) at $\beta = 1$ can be identified from noting that the leading-order inner solution satisfies for $\beta \neq 1$ the traveling-wave balance

$$\frac{1}{2-\beta} h^{2-\beta} - \frac{1}{1-\beta} h_\infty h^{1-\beta} + \frac{1}{(2-\beta)(1-\beta)} h_\infty^{2-\beta} = \frac{1}{2}\left(\frac{dh}{d\zeta}\right)^2 . \tag{A.19}$$

In the outer region $\eta_0^* < \eta < 0$, where $\eta_0^* = \beta B/2A^*$ is the location of the interface for the $A = A^*$ problem (see (A.11)), we have from (2.7) that

$$h - \frac{\beta}{2}\eta \frac{dh}{d\eta} = 0$$

holds to leading order (where $f = \varepsilon h$ for $0 < \beta < 1$, $f = \varepsilon^{1/(2-\beta)} h$ for $1 < \beta < 2$), so that

$$f \sim \varepsilon h_\infty (-\eta/\eta_0^*)^{\frac{2}{\beta}} \quad \text{for } 0 < \beta < 1, \quad f \sim \varepsilon^{\frac{1}{2-\beta}} h_\infty (-\eta/\eta_0^*)^{\frac{2}{\beta}} \quad \text{for } 1 < \beta < 2$$

for $\eta_0^* < \eta < 0$, whereby $\hat{A} \to 0$, $\hat{m} \to \infty$ in (A.4) as $\varepsilon \to 0^+$. Thus the apparently abrupt transition in η_0 from $\eta_0^* < 0$ at $A = A^*$ to zero for $A < A^*$ with $A^* - A$ arbitrarily small is effected by f tending to zero in $\eta_0^* < \eta < 0$ as $\varepsilon \to 0^+$.

The main purpose of the subdivision which follows is to distinguish for $B < 0$ the ranges of A in which (A.3) holds from those in which (A.4) holds.

(1) $A > A^*$. Here we have (as illustrated by the large A solution (A.11)) $\eta_0 > 0$ for $B > 0$ and $\eta_0 < 0$ for $B < 0$ (so (A.3) applies for $B < 0$), with the borderline case $B = 0$ being given by (A.7).

(2) $A = A^*$. For $B < 0$, this case furnishes the borderline between the regime $A > A^*$ in which solutions satisfy (A.3) with $\eta_0 < 0$ and $A_2 < A < A^*$ in which (A.4) applies. It is characterized by the solution (A.10) having zero slope (and thus representing the exceptional case $\lambda = 0$ in (A.3)).

(3) $A_2 < A < A^*$. For $B > 0$ we conjecture that $\eta_0 > 0$ remains valid (with (A.9) here applying whenever $A > A_2$ for $B > 0$), while for $B < 0$ we have (A.4), as illustrated by (A.17) and (A.18) (determining the full dependence of $\hat{A}$ on A for given β would require the boundary value problem to be solved numerically), and $\eta_0 = 0$ thus holds in this regime for all $B \leq 0$.

A.5. The fixed-front regime $\beta \leq 2$. As $\beta \to 2^-$ we have $A^* \to +\infty$ and much of the behavior in the fixed-front cases $\beta \geq 2$ is naturally different from that described above. The borderline $A = A_2$ still applies, however, but for $B < 0$ we now have that (A.4) applies for all $A > A_2$ for the relevant solutions, implying that $U(t)$ remains zero for some finite time. The analysis of section A.3 for $A = A_2 + \varepsilon$ still applies, as does that for $A \to \infty$ with $B > 0$, so for $B > 0$ we have

$$U(t) \sim \Omega^{\frac{2}{\beta}}(t) f(0) \quad \text{as } t \to 0$$

with $f(0) \sim B$ as $A \to \infty$ and with the leading-order expression for $f(0)$ as $A \to A_2$ being determined by the boundary value problem in which the far-field condition is given by (A.16) (we conjecture that this leads to a solution with $f(0) > 0$). To complete the picture it remains only to analyze the limit $A \to \infty$ with $B < 0$, for which the structure is somewhat similar to that described at the end of section 4. The scalings in the first of the two outer regions, namely, $\xi = O(1)$ with $\xi < -1$, are

$$\eta = (|B|/A)^{\frac{\beta}{2}}\xi\,, \quad f = |B|g,$$

whereby to leading order as $A \to \infty$ we have

$$0 = \frac{d}{d\xi}\left(g^\beta \frac{d}{d\xi}\left(g - \frac{\beta}{2}\xi\frac{dg}{d\xi}\right)\right),$$

from which the leading-order solution

$$g = (-\xi)^{\frac{2}{\beta}} - 1$$

follows. The inner scalings are

$$\xi = \xi_0(\varepsilon) + A^{-\frac{\beta}{\beta-2}}\zeta, \quad g = A^{-\frac{\beta}{\beta-2}}h\,,$$

with $\xi_0(0) = -1$ and where, for brevity, we shall not address the borderline case $\beta = 2$ in which the relevant scalings are exponentially small. This again gives the traveling-wave balance

$$h - h_\infty \sim h^\beta \frac{d^2h}{d\zeta^2}\,, \quad \begin{array}{l} h \sim \dfrac{2}{\beta}(-\zeta) \quad \text{as } \zeta \to -\infty\,, \\ h \to h_\infty \quad \text{as } \zeta \to +\infty\,, \end{array} \tag{A.20}$$

at leading order, where the positive constant h_∞ is determined as part of the solution; since the first integral (A.19) is available, we have that

$$h_\infty = (2(\beta-2)(\beta-1)/\beta^2)^{-\frac{1}{\beta-2}}\,.$$

Finally, in the second outer region $-1 < \xi < 0$ we have

$$h - \frac{\beta}{2}\xi\frac{dh}{d\xi} = 0$$

to leading order, so matching into the interior layer $\zeta = O(1)$ we obtain $h \sim h_\infty(-\xi)^{\frac{2}{\beta}}$, and hence in (A.4) we have

$$\hat{A} \sim A^{-\frac{2}{\beta-2}}h_\infty \quad \text{as } A \to \infty\,.$$

A.6. The special case $\beta = 1$. The implications of the above results for this special case (in which $A_2 = 1/6$, $A^* = 1/2$) are worth spelling out because it also represents the thin-film limit of the Hele–Shaw problem with kinetic undercooling (see [23], [9]; the formulation applies when the problem is symmetric about the x-axis, though the results can readily be generalized). The current analysis pertains locally when the initial fluid region contains an outward-pointing cusp of the form

$$u_0(x) \sim A(a-x)_+^2 \quad \text{as } x \to a^-\,. \tag{A.21}$$

In this context $y = u(x,t)$ denotes the upper fluid interface and the thin-film problem can be written in the form

$$p = \frac{\partial}{\partial x}\left(u\frac{\partial p}{\partial x}\right), \quad p = \frac{\partial u}{\partial t},$$

where p corresponds to the fluid pressure. Thus at $t = 0$ we typically have

$$p \sim B(a-x)^m \quad \text{as } x \to a^-$$

for some constant B (so that $\Omega(t) = t^{1/(2-m)}$ applies in the small-time local solution (2.6)), where m is given by (6.2) with $\beta = 1$. For $A < 1/6$ we have $m > 2$, so a waiting-time phenomenon ensues in which (A.21) remains valid locally for some finite time. In the injection case (corresponding to $B > 0$), for $A > 1/6$ the interface immediately propagates forward, though (for reasons discussed more generally in [15]) not according to the above thin-film prescription; rather we should impose on (2.7) the (fixed-front) condition

$$f\frac{d}{d\eta}\left(f - \frac{1}{2}\eta\frac{df}{d\eta}\right) = 0 \quad \text{at } \eta = 0^-,$$

with u in $x > a$ then being asymptotically given for small t by a quarter circle of radius $t^{2/(2-m)}f(0)$ (which is not of course of large aspect ratio, so does not fall within the remit of the thin-film evolution equation).

For the suction case ($B < 0$), however, we have for $1/6 < A < 1/2$ that the local behavior flips to (A.4) and then waits for some finite time. This class of suction phenomena is strikingly similar to that exhibited by the unregularized (and hence ill-posed) Hele–Shaw problem in which the interface initially contains a corner [25]; if the corner angle is less than $\pi/2$, the local behavior is then preserved for some finite time (as it is here for $A < 1/6$), while for angles between $\pi/2$ and π the corner angle instantaneously decreases to a value less than $\pi/2$ (which depends only on the initial angle) and then waits for some finite time (akin to the instantaneous decrease described above in the relevant coefficient from A to $\hat{A}$ ($< 1/6$), where $\hat{A}$ depends only on A for given β). However, for angles greater than π, the unregularized problem has no solution for $t > 0$, whereas in the current case the interface retracts in the equivalent regime $A > 1/2$, forming a corner of small but increasing angle (as in (A.3)). This continued existence of the solution is a desirable feature of the regularization; moreover, while the Hele–Shaw problem with kinetic undercooling exhibits (not just in its thin-film limit) formal time-reversal symmetry between the suction and injection cases, this symmetry is again broken by singularity formation (whereby if a corner forms and propagates under suction, it is immediately smoothed by injection in a non-time-reversible fashion; cf. the discussion of cusps in the zero-surface-tension Stokes flow problem in [12]).

Acknowledgments. The authors thank J. Hulshof for helpful comments. They gratefully acknowledge the support of the RTN project "Front-singularities" and J. R. King that of the EPSRC.

REFERENCES

[1] D. G. Aronson, *The porous medium equation*, in Nonlinear Diffusion Problems, Lect. 2nd 1985 Sess. C.I.M.E., Montecatini Terme, Italy, 1985, Lecture Notes in Math. 1224, Springer, Berlin, 1986, pp. 1–46.

[2] D. G. Aronson, L. A. Caffarelli, and S. Kamin, *How an initially stationary interface begins to move in porous medium flow*, SIAM J. Math. Anal., 14 (1983), pp. 639–658.
[3] G. Barenblatt, V. Entov, and V. Ryzhik, *Theory of Fluid Flows through Natural Rocks*, Theory and Applications of Transport in Porous Media 3, Kluwer Academic, Dordrecht, The Netherlands, 1990.
[4] G. I. Barenblatt, M. Bertsch, R. Dal Passo, V. M. Prostokishin, and M. Ughi, *A mathematical model of turbulent heat and mass transfer in stably stratified shear flow*, J. Fluid Mech., 253 (1993), pp. 341–358.
[5] G. I. Barenblatt, M. Bertsch, R. Dal Passo, and M. Ughi, *A degenerate pseudoparabolic regularization of a nonlinear forward-backward heat equation arising in the theory of heat and mass exchange in stably stratified turbulent shear flow*, SIAM J. Math. Anal., 24 (1993), pp. 1414–1439.
[6] G. I. Barenblatt, J. García-Azorero, A. De Pablo, and J. L. Vázquez, *Mathematical model of the non-equilibrium water-oil displacement in porous strata*, Appl. Anal., 65 (1997), pp. 19–45.
[7] A. Bourgeat and M. Panfilov, *Effective two-phase flow through highly heterogeneous porous media: Capillary nonequilibrium effects*, Comput. Geosci., 2 (1998), pp. 191–215.
[8] L. A. Caffarelli and A. Friedman, *Regularity of the free boundary for the one-dimensional flow of gas in a porous medium*, Amer. J. Math., 101 (1979), pp. 1193–1218.
[9] S. J. Chapman and J. R. King, *The selection of Saffman-Taylor fingers by kinetic undercooling*, J. Engrg. Math., 46 (2003), pp. 1–32.
[10] C. Cuesta, C. J. van Duijn, and J. Hulshof, *Infiltration in porous media with dynamic capillary pressure: Traveling waves*, European J. Appl. Math., 11 (2000), pp. 381–397.
[11] C. M. Cuesta, *Pseudo-parabolic Equations with Driving Convection Term*, Ph.D. thesis, Vrije Universiteit Amsterdam, 2003.
[12] L. J. Cummings, S. D. Howison, and J. R. King, *Two-dimensional Stokes and Hele-Shaw flows with free surfaces*, European J. Appl. Math., 10 (1999), pp. 635–680.
[13] W.-P. Düll, *Theory of a Pseudoparabolic Partial Differential Equation Modeling Solvent Uptake in Polymeric Solids*, Ph.D. thesis, University of Bonn, Germany, 2004.
[14] C. M. Elliott and H. Garcke, *On the Cahn–Hilliard equation with degenerate mobility*, SIAM J. Math. Anal., 27 (1996), pp. 404–423.
[15] J. D. Evans and J. R. King, *Asymptotic results for the Stefan problem with kinetic undercooling*, Quart. J. Mech. Appl. Math., 53 (2000), pp. 449–473.
[16] R. E. Grundy, *Local similarity solutions for the initial value problem in nonlinear diffusion*, IMA J. Appl. Math., 30 (1983), pp. 209–214.
[17] S. M. Hassanizadeh, *Dynamic effects in the capillary pressure saturation relationship*, in Proceedings of the 4th Annual International Conference on Civil Engineering, Water Resources and Environmental Engineering, Vol. 4, Tehran, Iran, 1997, Sharif University of Technology, pp. 141–149.
[18] S. M. Hassanizadeh, M. A. Celia, and H. K. Dahle, *Dynamic effect in the capillary pressure-saturation relationship and their impacts on unsaturated flow*, Vadose Zone Hydrology, 1 (2002), pp. 487–510.
[19] S. M. Hassanizadeh and W. G. Gray, *Thermodynamic basis of capillary pressure in porous media*, Water Resour. Res., 29 (1993), pp. 3389–3405.
[20] S. M. Hassanizadeh, R. J. Schotting, and A. Y. Beliaev, *A new capillary pressure-saturation relationship including hysteresis and dynamic effects*, in Proceedings of the 13th Annual International Conference on Computational Methods in Water Resources, Calgary, Canada, 2000, pp. 245–251.
[21] J. Hulshof and J. R. King, *Analysis of a Darcy flow model with a dynamic pressure saturation relation*, SIAM J. Appl. Math., 59 (1998), pp. 318–346.
[22] W. L. Kath and D. S. Cohen, *Waiting-time behavior in a nonlinear diffusion equation*, Stud. Appl. Math., 67 (1982), pp. 79–105.
[23] J. R. King, *Mathematical Aspects of Semiconductors Process Modelling*, Ph.D. thesis, University of Oxford, UK, 1986.
[24] J. R. King and C. M. Cuesta, *Small and Waiting Time Behaviour of a Darcy Flow Model with a Dynamic Pressure Saturation Relation*, Technical report, RTN "Front and Singularities," 2005; available online from http://www.iac.rm.cnr.it/rtn/.
[25] J. R. King, A. A. Lacey, and J. L. Vázquez, *Persistence of corners in free boundaries in Hele-Shaw flow*, European J. Appl. Math., 6 (1995), pp. 455–490.
[26] J. R. King and J. M. Oliver, *Thin-film modelling of poroviscous free surface flows*, European J. Appl. Math., 4 (2005), pp. 493–517.
[27] A. A. Lacey, *Initial motion of the free boundary for a nonlinear diffusion equation*, IMA J.

Appl. Math., 31 (1983), pp. 113–119.
[28] A. A. Lacey, J. R. Ockendon, and A. B. Tayler, *"Waiting-time" solutions of a nonlinear diffusion equation*, SIAM J. Appl. Math., 42 (1982), pp. 1252–1264.
[29] A. Novick-Cohen, *On the viscous Cahn-Hilliard equation*, in Material Instabilities in Continuum Mechanics, Oxford University Press, New York, 1988, pp. 329–342.
[30] V. Padrón, *Sobolev regularization of a nonlinear ill-posed parabolic problem as a model for aggregating populations*, Comm. Partial Differential Equations, 23 (1998), pp. 457–486.
[31] M. Rauscher, A. Münch, B. Wagner, and R. Blossey, *A thin-film equation for viscoelastic liquids of Jeffreys type*, Eur. Phys. J. E, 17 (2005), pp. 373–379.
[32] F. Stauffer, *Time dependence of the relations between capillary pressure, water content and conductivity during drainage of porous media*, in Proceedings of the LAHR Symposium on Scale Effects in Porous Media, Thessaloniki, Greece, LAHR, Madrid, Spain, 1978.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1512–1532

PERMEABILITY HYSTERESIS IN GRAVITY COUNTERFLOW SEGREGATION*

C. E. SCHAERER†, M. SARKIS‡, D. MARCHESIN†, AND P. BEDRIKOVETSKY§

Abstract. Hysteresis effects in two-phase flow in porous media are important in applications such as waterflooding or gas storage in sand aquifers. In this paper, we develop a numerical scheme for such a flow where the permeability hysteresis is modeled by a family of reversible scanning curves enclosed by irreversible imbibition and drainage permeability curves. The scheme is based on associated local Riemann solutions and can be viewed as a modification of the classical Godunov method. The Riemann solutions necessary for the scheme are presented, as well as the criteria that guarantee the well-posedness of the Riemann problem with respect to perturbations of left and right states. The numerical and analytical results show strong influence of the permeability hysteresis on the flow. In addition, the numerical scheme accurately reproduces the available experimental data once hysteresis is taken into account in the model.

Key words. relative permeability, hysteresis, Riemann problem, conservation laws, two-phase flow in porous media, Godunov method

AMS subject classifications. 74N30, 76T10, 35L65, 74S10

DOI. 10.1137/040616061

1. Introduction.

Capillary hysteresis strongly affects two-phase flow in porous media during sequential increase and decrease of wetting phase saturation (i.e., during the so-called imbibition and drainage, respectively) [6, 8]. The alternation of imbibition and drainage occurs in several oil recovery processes. It occurs in waterflooding with displacement direction change due to redistribution of injection and production rates in a system of wells, WAG (water-alternate-gas) injection of sequences of water and gas slugs, and sequential injection and production in the same well [4]. Annual injection and production of natural gas in aquifers or in depleted petroleum reservoirs for storage purposes also result in significant hysteretic phenomena. A similar flow regime change phenomenon, from imbibition to drainage and vice versa, occurs in secondary migration of hydrocarbons during the formation of petroleum accumulations [1], in irrigation, and in soil contamination by gasoline.

Capillary hysteresis at a macroscopic scale is caused by several pore scale phenomena. The contact angle on menisci between wetting and nonwetting phases suffers hysteresis during flow changes in a single pore. Creation of new interfacial surfaces resulting in energy losses occurs during imbibition; on the contrary, energy is released during drainage due to oil droplet joining. All these phenomena result in different scenarios of porous space filling by wetting and nonwetting phases.

*Received by the editors September 30, 2004; accepted for publication (in revised form) January 26, 2006; published electronically May 26, 2006.
http://www.siam.org/journals/siap/66-5/61606.html

†Instituto de Matemática Pura e Aplicada–IMPA, E. D. Castorina 110, 22460-320, Rio de Janeiro, RJ, Brazil (cschaer@fluid.impa.br, marchesi@impa.br). The research of the first and third authors was partially supported by FAPERJ (E-26/152.254/2002, E-26/152.163/2002) and by CNPq (500075/2002-6, 301532/2003-6).

‡Worcester Polytechnic Institute, Worcester, MA 01609 (msarkis@impa.br). The research of this author was partially supported by CNPq (Brazil) under grant 305539/2003-8 and by the U.S. National Science Foundation under grant CGR 9984404.

§Petroleum Department, North Fluminense State University, R. Sebastião Lopes da Silva 56, Macaé, 27937-150 RJ, Brazil (pavel@lenep.uenf.br).

Models for multiphase flow in porous media are based on conservation of mass and Darcy's law. The associated equations contain quantities describing the rock and fluid properties, in particular relative phase permeabilities. The latter describe the capability of each phase to flow in the porous medium [2, 9]. Relative permeability of the nonwetting phase exhibits hysteresis or memory effects [10], i.e., according to the saturation tendency, the relative permeability is different [18].

The model for hysteretic relative permeability of the nonwetting phase [2, 12] follows experimental observations of drainage, imbibition, and scanning behavior of relative permeability [6, 7, 8, 13, 19]. Changes in the direction of the flow in drainage and imbibition are irreversible [6, 18]; flow in the region between drainage and imbibition curves is in general almost reversible [8, 19, 17]; however, we will make the approximation that it is exactly reversible [6, 13].

Observations and explanations of permeability hysteresis in laboratory experiments for horizontal one-dimensional flow were presented in several works [17, 6, 8]. However, mathematical understanding is insufficient, hindering the inclusion of hysteresis in numerical simulation of reservoir flow. Formulae for drainage, imbibition, and scanning relative permeabilities curves were developed in [2, 12], among others. In [15], only the imbibition and drainage curves were considered. A model that we will call the scanning hysteresis model (SHM) for the history dependence of the relative permeabilities was presented in [10] and in [18].

In the current work, we concentrate on hysteretic gravity segregation. This phenomenon occurs after waterflooding or after gas injection in thick oil formations. It also occurs in in-situ gas storage in thick formations between injection and production cycles. Estimation of the gravity separation time is necessary for planning of tertiary recovery from reformed formations.

Our goal is to develop a numerical tool for the gravitational counterflow segregation problem with a hysteretic relative permeability. Because Riemann solutions with hysteresis in the relative permeability are not unique, we introduce criteria to obtain well-posedness with respect to left and right states. In the large scale approximation formulation, we do not include the capillarity pressure and its hysteresis [4].

The paper is organized as follows: In section 2, we present the model for two-phase gravity counterflow segregation. In addition, we extend the SHM for the nonwetting relative permeability [18] to include gravity. This model associates a hysteretic parameter π in order to "remember" the value of the saturation at the last time when the saturation tendency was reversed. In the equations used to model the segregation, the capillary forces and its hysteresis affect both the transport part and the diffusive part. We concentrate on the hysteresis in relative permeability. In section 3, the Riemann solutions for the hysteretic conservation law are discussed. Criteria to select a unique well-posed solution are developed. In section 4, we propose a corrected Godunov scheme that updates both the saturation and the hysteretic states. This scheme conserves mass locally. Finally, in section 5, we show that the numerical solution of the Riemann problem converges to the exact solution. Comparisons of the numerical solution (with and without hysteresis) with laboratory data are presented. They show that hysteresis must be taken into account to obtain correct predictions of segregation. Additionally, it is demonstrated that the proposed numerical method captures adequately the experimental profiles, and the main hysteresis effects can be modeled through the relative permeability curves.

2. The two-phase model for gravity counterflow segregation. We consider a sand-packed vertical tube with a given initial saturation profile of two incom-

pressible immiscible fluids. Redistribution of the fluids with different densities occurs due to gravitational forces. The total flow along the tube is zero as the tube is closed at the top and bottom. Neglecting diffusive terms due to capillarity relative to gravitational forces, the two-phase flow model equation expressing mass conservation and Darcy's law in dimensionless variables (z,t), $\{0 \le z \le 1, t \ge 0\}$ is [4]

$$\partial_t s + \partial_z F(s,\pi) = 0, \tag{2.1}$$

where the flow function $F(s,\pi)$ is

$$F(s,\pi) := \frac{k_{rw}k_{ro}}{k_{ro} + (\mu_o/\mu_w)k_{rw}}. \tag{2.2}$$

We use the indices w and o to refer to the wetting and nonwetting phases, respectively. The quantities k_{rw} (k_{ro}) and μ_w (μ_o) are the relative permeability and the viscosity of the wetting phase w (nonwetting o), respectively. In the absence of hysteresis, k_{rw} (k_{ro}) are functions of the effective wetting phase saturation s defined as

$$s := \frac{s_w - s_{wi}}{s_{ro} - s_{wi}}, \tag{2.3}$$

where s_{wi} is the irreducible wetting saturation and s_{ro} is the residual nonwetting saturation.

2.1. The SHM: Mathematical description. To model the hysteresis phenomenon observed experimentally in the relative permeabilities [6], we extend the SHM presented in [10, 18] to include gravity. For simplicity, the nonwetting phase exhibits hysteresis, while the wetting phase does not. In order to describe the behavior due to hysteresis, a parameter π is introduced. Concretely, we generalize the permeability functions presented in [18] and use the following special permeability functions of the effective saturation s [3].

The wetting relative permeability is defined as (Figure 2.1(a))

$$k_{rw}(s) := \gamma s^{\beta},\ \beta > 1, \tag{2.4}$$

where γ is a parameter to adjust the curve (2.4) to the relative permeability curve obtained experimentally, so that $k_{rw}(1) = \gamma$, as the nonwetting permeability is normalized as 1 for $s = 0$.

The nonwetting drainage and imbibition relative permeabilities functions (expressed by k^d_{ro} and k^i_{ro}) are defined as (Figure 2.1(a)) [3]

$$\begin{aligned} k^d_{ro}(s) &= (1-s)^{\eta} \quad \text{for } 0 \le s \le 1 \quad \text{and when } \frac{\partial s}{\partial t} < 0, \\ k^i_{ro}(s) &= (1-s)^{\theta} \quad \text{for } 0 \le s \le 1 \quad \text{and when } \frac{\partial s}{\partial t} > 0, \end{aligned} \tag{2.5}$$

where $1 < \theta < \eta$. In this paper, we use $\beta = 2$, $\eta = 3$, and $\theta = 2$; therefore $k^d_{ro} \le k^i_{ro}$.

The scanning region corresponds to the region between the nonwetting drainage and imbibition relative permeabilities curves. In such a region the nonwetting permeability k_{ro} is chosen as

$$k_{ro}(s,\pi) := \frac{(1-\pi)^{\xi}}{(1-\alpha\pi)^{\zeta}}(1-\alpha s)^{\zeta}, \tag{2.6}$$

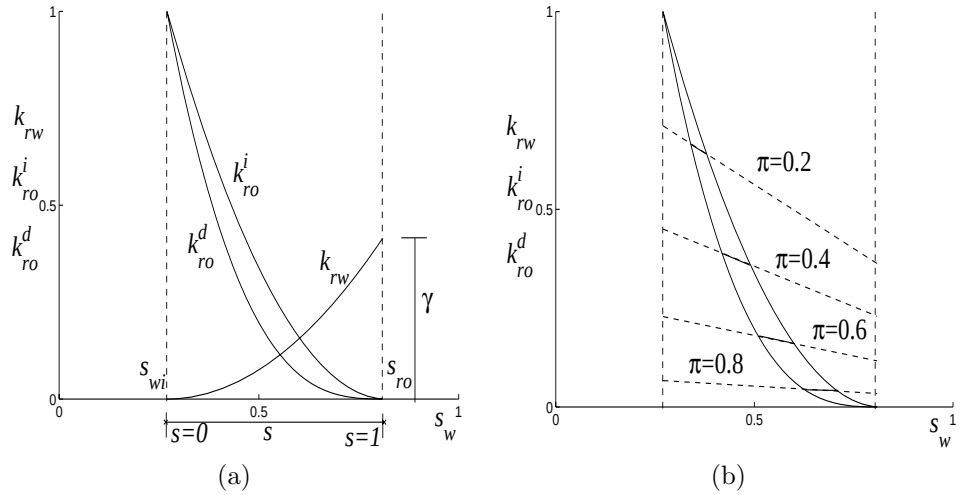


FIG. 2.1. (a) *Nonwetting imbibition and drainage permeabilities functions, and wetting permeability, and* (b) *scanning curves (as a function of the saturation of the wetting phase s_w and the hysteresis parameter π); inspired by Braun and Holland* [6].

where the parameter π ($0 \le \pi \le 1$) discriminates each scanning curve. We use $\xi = 2$ and $\zeta = 1$; therefore, k_{ro} defined above is linear in s for each fixed π; see Figure 2.1(b). Notice that the dashed lines meet at the point $(1/\alpha, 0)$, where α is another parameter used to adjust the slopes of the scanning curves to the experimental ones. We use $\alpha = 0.5$. In the SHM, the scanning curve associated to π is defined in the saturation range $s^i(\pi) < s < s^d(\pi)$. The functions $s^i(\pi)$ and $s^d(\pi)$ are defined implicitly by

$$k^d_{ro}(s^d(\pi)) = k_{ro}(s^d(\pi), \pi) \quad \text{and} \quad k^i_{ro}(s^i(\pi)) = k_{ro}(s^i(\pi), \pi). \tag{2.7}$$

On the drainage and imbibition permeability curves, expressions for the parameter π as a function of the saturation s, and vice versa, can be obtained. Thus we define from (2.7) the functions $\pi^d(s)$ and $\pi^i(s)$. Plots of these functions are shown in Figure 2.2(a).

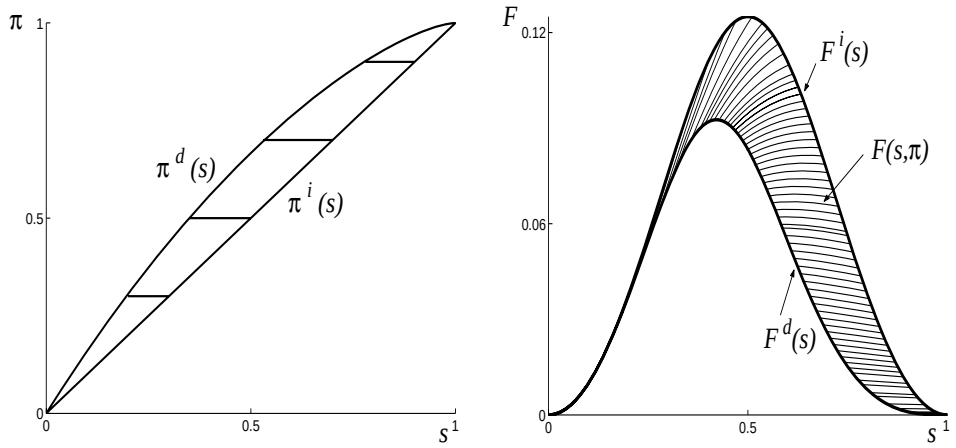
The flux (2.2) depends on the history (expressed by the parameter π) and the type of the flow (expressed through the sign of $\partial_t s$). Using the relative permeabilities k_{ro} in (2.6) and (2.5), and using k_{rw} in (2.4), the flux function (2.2) takes the form

$$F(s, \pi) \text{ in the scanning region, where } \partial_t \pi = 0, \tag{2.8}$$

$$F^d(s) := F(s, \pi^d(s)) \text{ on the drainage curve, where } \partial_t s < 0, \tag{2.9}$$

$$F^i(s) := F(s, \pi^i(s)) \text{ on the imbibition curve, where } \partial_t s > 0. \tag{2.10}$$

The fluxes $F(s, \pi)$, $F^d(s)$, and $F^i(s)$ (given by (2.2), (2.9), and (2.10), respectively) are shown in Figure 2.2(b). Notice that the drainage and imbibition curves $F^d(s)$ and $F^i(s)$ bound the admissible scanning region Ω on the plane (s, F), defined as $\Omega = \{(s, F) \in \Re^2 : F^d(s) \le F \le F^i(s)\}$. We define the drainage and imbibition curves as $dr := \{(s, F) \in \Re^2 : F = F^d(s)\}$ and $im := \{(s, F) \in \Re^2 : F = F^i(s)\}$, respectively. The permeability functions (2.4) and (2.5) lead to the following necessary properties of the fractional flow function in the scanning, drainage, and imbibition



(a) (b)

FIG. 2.2. (a) *State space.* (b) *Fractional flow curves (imbibition, drainage, and scanning).*

flows in the SHM:

$$F^d(0) = 0, \quad F^d(1) = 0, \quad F^i(0) = 0, \quad F^i(1) = 0, \tag{2.11}$$

$$\partial_s F^d(s) \le \partial_s F(s, \pi^d(s)), \quad \partial_s F^i(s) \le \partial_s F(s, \pi^i(s)); \tag{2.12}$$

of course other permeability functions satisfying (2.11) and (2.12) can be chosen. These properties ensure that the intersection of the scanning curves with the drainage (or imbibition) curve varies smoothly with π. In addition, as we will see in the next section, property (2.12) guarantees the existence of solution at the intersections of the scanning with the drainage and imbibition curves.

3. The Riemann problem. For imbibition and drainage flows the pair (s, π) lies on the imbibition and drainage curves; therefore the value of π is given by $\pi = \pi^i(s)$ and $\pi = \pi^d(s)$. In these cases, scalar conservation laws are satisfied:

$$\partial_t s + \partial_z F^j(s) = 0, \ j := i, d. \tag{3.1}$$

For scanning flow the conservation law (3.1) is extended to include the independent variable π. Hence the conservation law becomes

$$\partial_t s + \partial_z F(s, \pi) = 0, \tag{3.2}$$

$$\partial_t \pi = 0, \tag{3.3}$$

and can be written in quasi-linear form $\partial_t u + A^s \partial_z u = 0$ with Jacobian

$$A^s = \begin{bmatrix} \partial_s F(s,\pi) & \partial_\pi F(s,\pi) \\ 0 & 0 \end{bmatrix} \text{ and } u := [s, \pi]^T.$$

For scanning flow, the eigenvalues of A^s (or characteristics speeds) are zero and $\partial_s F(s, \pi)$, with corresponding eigenvectors $[\partial_\pi F(s,\pi), \ -\partial_s F(s,\pi)]^T$ and $[1, \ 0]^T$.

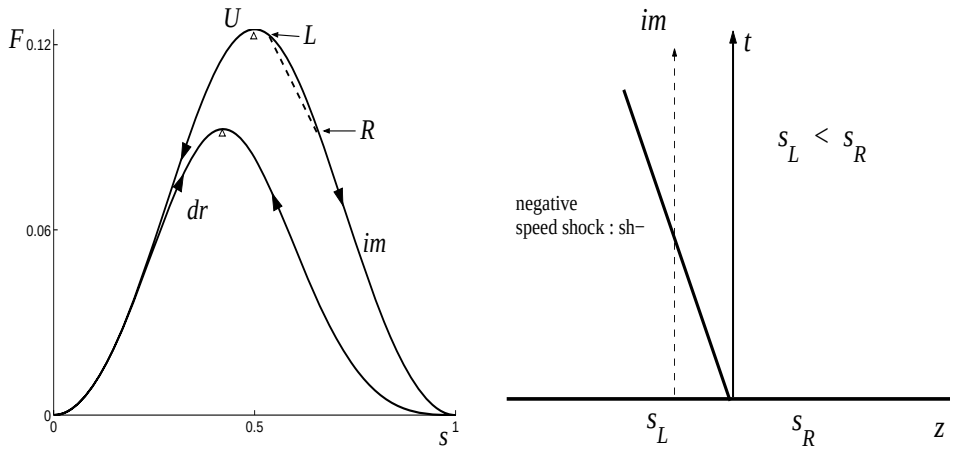


FIG. 3.1. *Flow direction in drainage and imbibition curves. dr: drainage, im: imbibition;* s_U*: saturation associated to the state* $U = \big(s_U, F^i(s_U)\big)$.

3.1. Wave families. For each flow we describe the solution of the Riemann problem:

$$(s,\pi)|_{t=0} = \begin{cases} (s_L,\pi_L) & \text{for } z < z_o, \\ (s_R,\pi_R) & \text{for } z > z_o. \end{cases} \tag{3.4}$$

Hereafter, we will use the following notation. A state L is defined as $L := (s, F) \in \Omega$, where s_L and F_L are the saturation and the flux associated to the state L. Additionally, π_L is the associated parameter π to the state L. In the scanning region, $F_L = F(s_L, \pi_L)$, and on the curve dr (im) $F_L = F^d(s_L)$ $(F_L = F^i(s_L))$.

3.1.1. Imbibition or drainage flow. Since in imbibition (drainage) flow the saturation increases (decreases) in time, i.e., $\partial_t s > 0$ $(\partial_t s < 0)$, a rarefaction wave is characterized by a continuous and monotonically increasing speed $\lambda = \partial_s F^j(s)$ from s_L to s_R. Additionally, a shock wave with speed σ satisfies the Rankine–Hugoniot (RH) condition:

$$\sigma = \frac{F^j(s_R) - F^j(s_L)}{s_R - s_L},\ j = i, d; \tag{3.5}$$

furthermore, the shock is required to satisfy the Oleinik entropy condition for scalar equations. Therefore, admissible sequences of shock and rarefaction waves can be constructed graphically using the concave and convex hull of the fractional flow curves, following [16]. We still have to impose the imbibition and drainage flow orientation. Let s_U be the saturation that maximizes $F^i(s)$ $(U := \{(s_U, F^i(s_U) : F_U = \max(F^i(s)) \text{ for } 0 \le s \le 1\})$. To select the imbibition (drainage) flow orientation the conditions $s_U \le s_L \le s_R$ or $s_R \le s_L \le s_U$ $(s_U \le s_R \le s_L$ or $s_L \le s_R \le s_U)$ must hold. For instance, a shock satisfying these conditions is shown in Figure 3.1(a), and we see that $\partial_t s > 0$ in Figure 3.1(b). For other cases, such as $s_U \le s_R \le s_L$, $s_L \le s_R \le s_U$ $(s_U \le s_L \le s_R$, $s_R \le s_L \le s_U$ for drainage), $s_L \le s_U \le s_R$, and $s_R \le s_U \le s_L$, the Riemann solution must contain scanning waves; otherwise $\partial_t s$ would be negative and therefore the solution could not be on the imbibition curve.

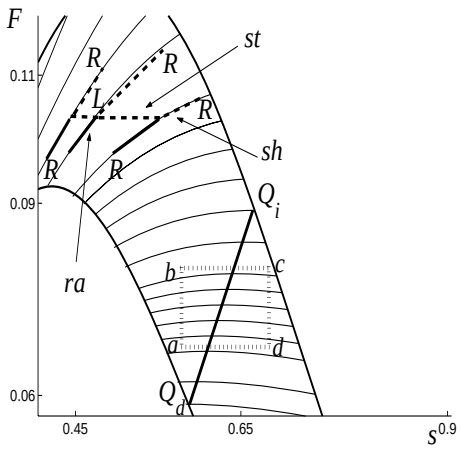


FIG. 3.2. *Scanning cases. Stationary shock, st; shock, sh; rarefaction, ra.*

3.1.2. Scanning flow. The RH condition is

$$F(s_R, \pi_R) - F(s_L, \pi_L) = \sigma(s_R - s_L), \tag{3.6}$$

$$0 - 0 = \sigma(\pi_R - \pi_L). \tag{3.7}$$

Equations (3.6) and (3.7) are satisfied by two kinds of discontinuities; see Figure 3.2. The first one is a shock with speed σ:

$$F(s_R, \pi) - F(s_L, \pi) = \sigma\left(s_R - s_L\right), \tag{3.8}$$

where π is constant, i.e., $\pi = \pi_L = \pi_R$. This corresponds to the Riemann solution of the single scalar conservation law (3.2) (see Figure 3.2). The second kind of discontinuities satisfying the RH condition (3.6)–(3.7) are stationary discontinuities with speed $\sigma = 0$ and constant fractional flow function; generically they satisfy $\pi_R \neq \pi_L$ and $s_L \neq s_R$.

Summarizing, the scanning curve and the horizontal line through state L divide the scanning region in R-regions where the solution consists of a combination of a stationary wave and scanning waves that are either rarefaction or shocks. Some simple cases are presented in Figure 3.2. A more complex Riemann solution exists when scanning curves have a maximum in Ω and there is no interaction with the imbibition and drainage curves. This RP presents multiplicity of solutions; therefore, we have to make choices to have appropriate solutions. To classify the chosen Riemann solutions, we analyze the *abcd* region in Figure 3.2. This region is sketched in Figures 3.3(a) and (b). The curve Q_dQ_i is the set of states $Q = (s, F(s, \pi))$ in Ω where the saturation s maximizes $F(s, \pi)$ for each π (see Figure 3.2).

In Figures 3.3(a) and (b), we show the regions and their associated Riemann solutions for left states L_1 and L_2 lying on opposite sides of the curve Q_dQ_i. A summary of the solutions is given in Table 3.1; the waves are ordered from lower to higher speeds. For example, the solution $L_1I_1R_1$ consists of a stationary shock L_1I_1 connecting states L_1 and I_1 (denoted by $L_1 \xrightarrow{st} I_1$), and a positive speed shock I_1R_1 connecting states I_1 and R_1 (denoted by $I_1 \xrightarrow{sh+} R_1$).

3.1.3. Intermediate flow. Denoting by I the intersection state between the π_L-scanning and the drainage (imbibition) curves, the solution between a scanning state

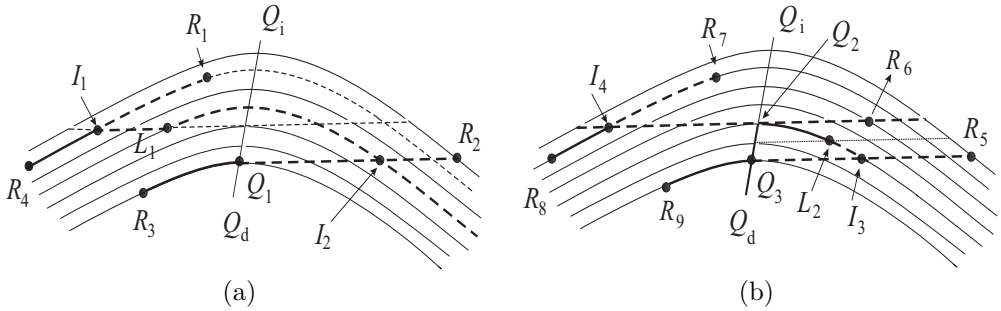


FIG. 3.3. *Regions in the scanning region abcd of Figure* 3.2.

TABLE 3.1

Riemann solutions for a region where the scanning curves have a maximum in Ω *where waves are ordered from negative to positive. Negative speed rarefaction:* $ra-$*, negative speed shock:* $sh-$*, stationary shock:* st*, positive speed rarefaction:* $ra+$*, and positive speed shock:* $sh+$*.*

Riemann solutions for L_1, Figure 3.3(a)
waves
$L_1 \xrightarrow{st} I_1 \xrightarrow{sh+} R_1$
$L_1 \xrightarrow{sh-} I_2 \xrightarrow{st} R_2$
$L_1 \xrightarrow{sh-} I_2 \xrightarrow{st} Q_1 \xrightarrow{ra+} R_3$
$L_1 \xrightarrow{st} I_1 \xrightarrow{ra+} R_4$
Riemann solutions for L_2, Figure 3.3(b)
waves
$L_2 \xrightarrow{sh-} I_3 \xrightarrow{st} R_5$
$L_2 \xrightarrow{ra-} Q_2 \xrightarrow{st} R_6$
$L_2 \xrightarrow{ra-} Q_2 \xrightarrow{st} I_4 \xrightarrow{sh+} R_7$
$L_2 \xrightarrow{ra-} Q_2 \xrightarrow{st} I_4 \xrightarrow{ra+} R_8$
$L_2 \xrightarrow{sh-} I_3 \xrightarrow{st} Q_3 \xrightarrow{ra+} R_9$

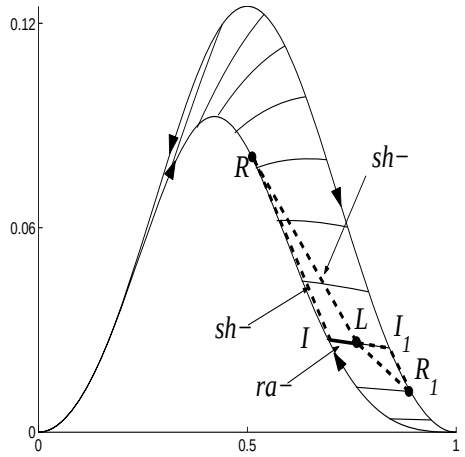
L and a drainage (imbibition) state R is constructed using a concave (convex) hull curve on the effective flux function LIR (see Figure 3.4). Therefore, the interaction of the two waves LI and IR yields a shock LR with speed

$$\sigma = \frac{F^d(s_R) - F(s_L, \pi_L)}{s_R - s_L}, \tag{3.9}$$

since the shock IR has a smaller speed than the rarefaction LI, i.e., $(F^d(s_R) - F^d(s^d(\pi_L)))/(s_R - s^d(\pi_L)) \le \partial_s F(s^d(\pi_L), \pi_L)$, and at I the drainage and scanning curves satisfy inequalities (2.12).

3.2. Riemann solutions. The construction of the Riemann solutions is simplified by describing state space in terms of (s, F) instead of (s, π). We subdivide Ω into four subregions (L-regions) defined as follows (see Figure 3.5):

$$\begin{aligned}
A &= \{(s,F) \in \Omega : s_{Q_d} \le s \le 1 \text{ and } F \le F(s, \pi_{Q_d})\},\\
B &= \{(s,F) \in \Omega : s_M \le s \le s_P \text{ and } F(s, \pi_{Q_d}) \le F \le F^d(s_M)\},\\
C &= \{(s,F) \in \Omega : 0 \le s \le 1 \text{ and } F^d(s_M) \le F\},\\
D &= \{(s,F) \in \Omega : 0 \le s \le s_M \text{ and } F \le F^d(s_M)\},
\end{aligned}$$



(a)

FIG. 3.4. *Transition from a scanning curve to a drainage or imbibition curve.*

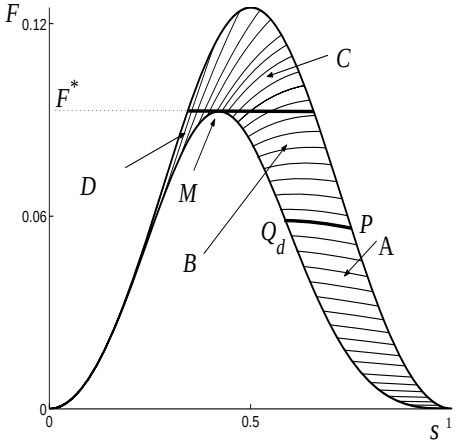
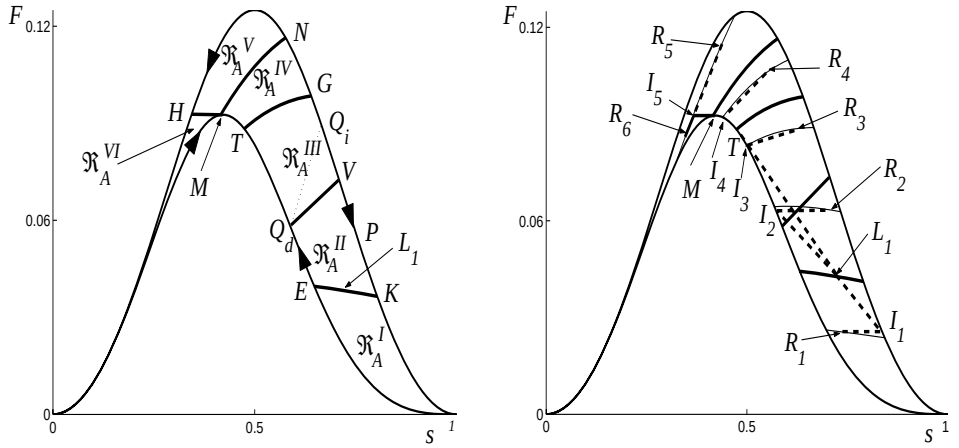


FIG. 3.5. *L-regions of* Ω.

where the imbibition state P is defined as the intersection point of the imbibition curve with the scanning curve through the states Q_d, while M is the state where s_M maximizes $F^d(s_M)$. Next we construct the solutions for L in each of the L-regions.

Case A: $L_1 = (s_L, F(s_L, \pi_L)) \in A$. In this case, there are the following six R-regions (see Figure 3.6(a)):

$$\begin{aligned}
\Re_A^I &= \{(s,F) \in \Omega : s_E \le s \text{ and } F \le F(s,\pi_L)\},\\
\Re_A^{II} &= \{(s,F) \in \text{region enclosed by } EKVQ_d\},\\
\Re_A^{III} &= \{(s,F) \in \text{region enclosed by } Q_dVGT\},\\
\Re_A^{IV} &= \{(s,F) \in \Omega : s_M \le s \le s_G \text{ and } F(s,\pi_T) \le F \le F(s,\pi_M)\},\\
\Re_A^{V} &= \{(s,F) \in \Omega : s_H \le s \le s_N \text{ and } F(s,\pi_M) \le F \text{ and } F^d(s_M) \le F\},\\
\Re_A^{VI} &= \{(s,F) \in \Omega : s \le s_M \text{ and } F \le F^d(s_M)\},
\end{aligned}$$



(a) (b)

FIG. 3.6. *Figure* (a) *(Case A): R-regions for L_1 in L-region A. Figure* (b) *(Case A): solutions for L_1 in L-region A and R in R-regions: $R_1 \in \Re_A^I$, $R_2 \in \Re_A^{II}$, $R_3 \in \Re_A^{III}$, $R_4 \in \Re_A^{IV}$, $R_5 \in \Re_A^V$, and $R_6 \in \Re_A^{VI}$.*

where the states E and K are the intersection of drainage and imbibition curves with the scanning curve through L_1. We define the state $V \in im$ and the curve Q_dV by $Q_dV := \{W = (s_W, F_W) \in \Omega : (s_W, F_W) = (s_I + 2(s_Q - s_I), F_I)\ \forall I \in dr\}$. States N and H are the intersections of the imbibition curve with the scanning and the horizontal line through M, respectively. We define the drainage state T through $\partial_s F^d(s_T) = (F^d(s_T) - F_L)/(s_T - s_L)$, and the state G as the intersection of the imbibition curve with the scanning curve though T. Notice that T and G depend on L_1. Assuming that $\partial_{ss}F(s,\pi) < 0$, the Riemann solution associated to each R-region as shown in Figure 3.6(a) is presented in Figure 3.6(b) and described below:

A.1. For $R_1 \in \Re_A^I$, the solution is $L_1 \xrightarrow{sh-} I_1 \xrightarrow{st} R_1$.

A.2. For $R_2 \in \Re_A^{II}$, the solution is $L_1 \xrightarrow{sh-} I_2 \xrightarrow{st} R_2$.

A.3. For $R_3 \in \Re_A^{III}$, the solution is $L_1 \xrightarrow{sh-} I_3 \xrightarrow{sh+} R_3$, where state I_3 is determined by the intersection of the drainage curve with the scanning curve through R_3.

A.4. For $R_4 \in \Re_A^{IV}$, the solution is $L_1 \xrightarrow{sh-} T \xrightarrow{ra-} I_4 \xrightarrow{sh+} R_4$, where I_4 is determined by the intersection of the drainage curve with the scanning curve through R_4.

A.5. For $R_5 \in \Re_A^V$, the solution is $L_1 \xrightarrow{sh-} T \xrightarrow{ra-} M \xrightarrow{st} I_5 \xrightarrow{sh+} R_5$, where state I_5 is the intersection of the horizontal line through M and the scanning curve through R_5.

A.6. For $R_6 \in \Re_A^{VI}$, the solution is $L_1 \xrightarrow{sh-} T \xrightarrow{ra-} M \xrightarrow{st} I_5 \xrightarrow{ra+} R_6$. When I_5 is on the imbibition curve, the solution will be shown in Case *C*, Figure 3.9(a)

Case B: $L_2 = (s_L, F(s_L, \pi_L)) \in B$. The R-regions are shown in Figure 3.7(a). The Riemann solutions for the R-regions $\Re_B^V$, $\Re_B^{VI}$, and $\Re_B^{VII}$ are analogous to those of solutions for the R-regions $\Re_A^{IV}$, $\Re_A^V$, and $\Re_A^{VI}$, respectively. The solutions for R-regions $\Re_B^I$, $\Re_B^{II}$, $\Re_B^{III}$, and $\Re_B^{IV}$ are shown in Figure 3.7(b) and described below:

B.1. For $R_1 \in \Re_B^I$, the solution is $L_2 \xrightarrow{sh-} I_1 \xrightarrow{st} R_1$.

B.2. For $R_2 \in \Re_B^{II}$, the solutions is $L_2 \xrightarrow{sh-} I_2 \xrightarrow{st} Q \xrightarrow{ra+} R_2$.

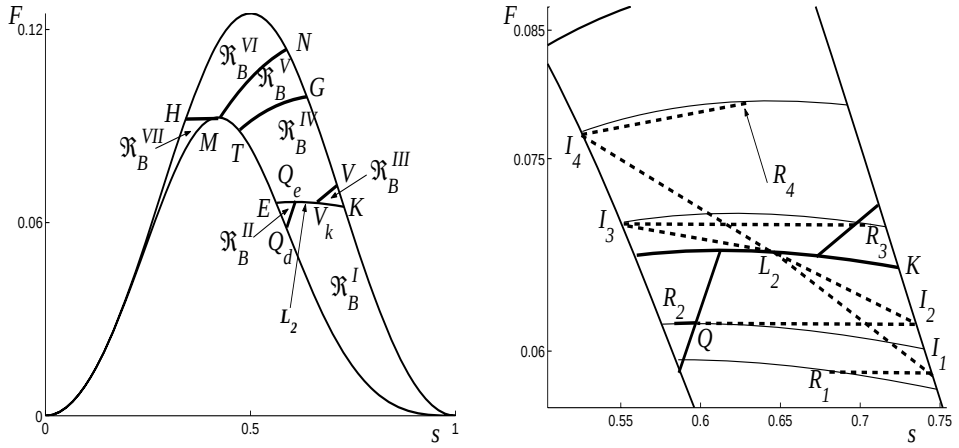


FIG. 3.7. *Figure* (a) *(Case B): R-regions for L_2 in the L-region B. Figure* (b) *(Case B): L_2 in L-region B and $R_1 \in \Re_B^I$, $R_2 \in \Re_B^{II}$, $R_2 \in \Re_B^{III}$, and $R_4 \in \Re_B^{IV}$.*

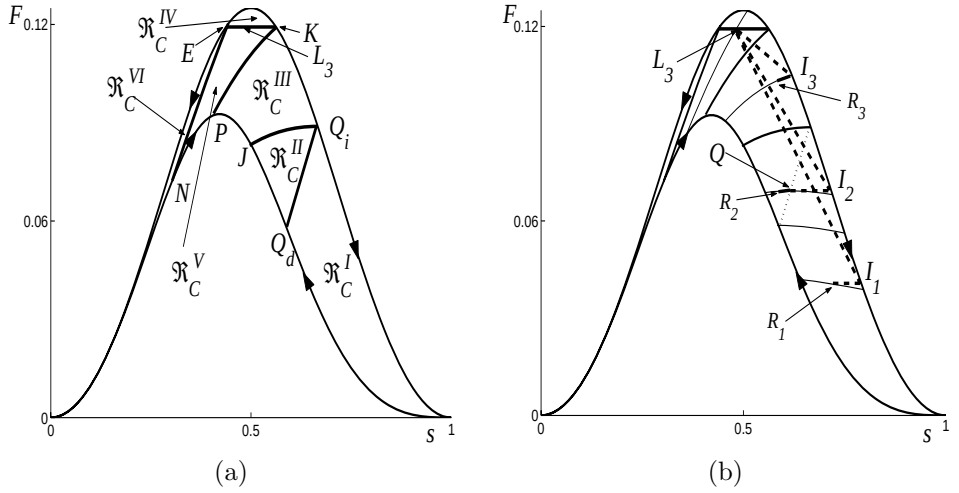


FIG. 3.8. *Figure* (a) *(Case C): R-regions for L_3 in L-region C. Figure* (b) *(Case C): solutions for L_3 in L-region C and R in R-regions: $R_1 \in \Re_C^I$, $R_2 \in \Re_C^{II}$, and $R_3 \in \Re_C^{III}$.*

B.3. For $R_3 \in \Re_B^{III}$, the solution is $L_2 \xrightarrow{sh-} I_3 \xrightarrow{st} R_3$.

B.4. For $R_4 \in \Re_B^{IV}$, the solution is $L_2 \xrightarrow{sh-} I_4 \xrightarrow{sh+} R_4$.

Case C: $L_3 = (s_L, F(s_L, \pi_L)) \in C$. The R-regions are shown in Figure 3.8(a). The Riemann solution belongs to one of the following cases:

C.1. For $R_1 \in \Re_C^I$ (Figure 3.8(b)), the solution is $L_3 \xrightarrow{sh-} I_1 \xrightarrow{st} R_1$.

C.2. For $R_2 \in \Re_C^{II}$, the solution is analogous to that of Case B (subcase B.2). In the solution shown, I_2 cannot be connected to R_2 only by a stationary shock because the stationary shock intersects the scanning curve with π_R. Therefore, the solution is $L_3 \xrightarrow{sh-} I_2 \xrightarrow{st} Q \xrightarrow{ra+} R_2$.

C.3. For $R_3 \in \Re_C^{III}$, the solution is $L_3 \xrightarrow{sh-} I_3 \xrightarrow{ra+} R_3$.

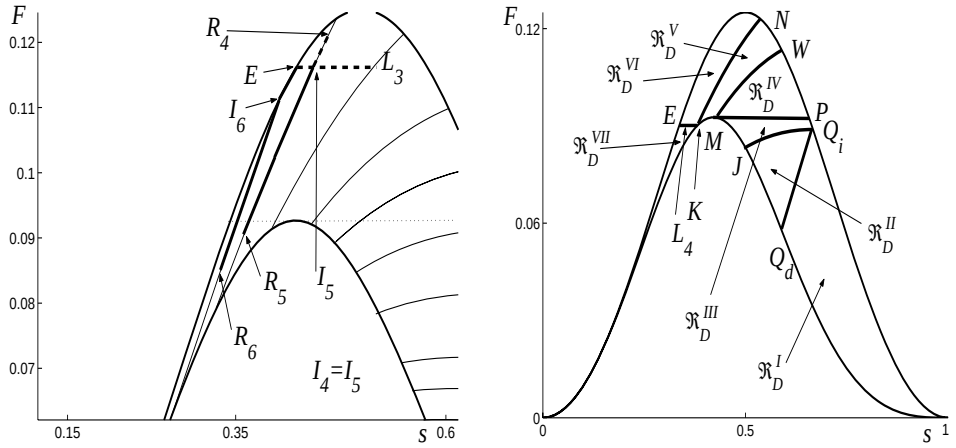


FIG. 3.9. *Figure* (a) *(Case C): solutions for* $L_3 \in C$ *and* R *in* R*-regions:* $R_5 \in \Re_C^V$ *and* $R_6 \in \Re_C^{VI}$. *Figure* (b) *(Case D): R-regions for* $L_4 \in D$.

C.4. For $R_4 \in \Re_C^{IV}$ (Figure 3.9(a)), the solution is $L_3 \xrightarrow{st} I_5 \xrightarrow{sh+} R_4$.

C.4. For $R_5 \in \Re_C^V$, the solution is $L_3 \xrightarrow{st} I_5 \xrightarrow{ra+} R_5$.

C.5. For $R_6 \in \Re_C^{VI}$, L_3 cannot be connected to the scanning curve through R_6 only by a stationary shock, because the latter intersects the imbibition curve (state E). For L_3 and R_6 the solution is $L_3 \xrightarrow{st} E \xrightarrow{ra+} I_6 \xrightarrow{ra+} R_6$.

Case D: $L_4 = (s_L, F(s_L, \pi_L)) \in D$. The R-regions are shown in Figure 3.9(b). The Riemann solutions belong to one of the following cases:

D.1. For $R_1 \in \Re_D^I$ (Figure 3.10(a)), the solution is $L_4 \xrightarrow{st} K \xrightarrow{sh-} I_2 \xrightarrow{st} R_1$.

D.2. For $R_2 \in \Re_D^{II}$, the solution is $L_4 \xrightarrow{st} K \xrightarrow{sh-} I_1 \xrightarrow{st} Q \xrightarrow{ra} R_2$.

D.3. For $R_3 \in \Re_D^{III}$, the solution is $L_4 \xrightarrow{st} K \xrightarrow{sh+} I_3 \xrightarrow{ra+} R_3$. Notice that segment KI_3 is tangent to the *dr* curve at I_3 to the scanning curve through R_3.

D.4. For $R_4 \in \Re_D^{IV}$, the solution is $L_4 \xrightarrow{st} K \xrightarrow{sh+} I_4 \xrightarrow{sh+} R_4$.

D.5. For $R_5 \in \Re_D^{IV}$ (Figure 3.10(b)), the solution is $L_4 \xrightarrow{st} K \xrightarrow{sh+} I_5 \xrightarrow{sh+} R_5$.

D.6. For $R_6 \in \Re_D^{VI}$, the solution is $L_4 \xrightarrow{st} I_6 \xrightarrow{sh+} R_6$.

D.7. For $R_7 \in \Re_D^{VII}$, the solution is $L_4 \xrightarrow{st} I_7 \xrightarrow{ra+} R_7$. If the stationary shock intersects the *im* curve, the solution is analogous to that of Case C (subcase C.5).

3.3. Uniqueness criteria. Without appropriate restrictions for given L and R states, the Riemann problem might have multiple solutions. We are interested in solutions satisfying (1) the Oleinik condition in Ω and (2) the orientation of the drainage and imbibition curves. However, these conditions are insufficient to guarantee uniqueness of solution. A solution can be obtained if we further require that (3) the solution must be $\mathcal{L}_1^{loc}$ continuous with respect to changes in L and R. A Riemann solution is well posed if it satisfies conditions (1), (2), and (3) mentioned above. The choices presented in section 3.2 give rise to well-posed global solutions; however, some of them are not evident; so we discuss these cases here.

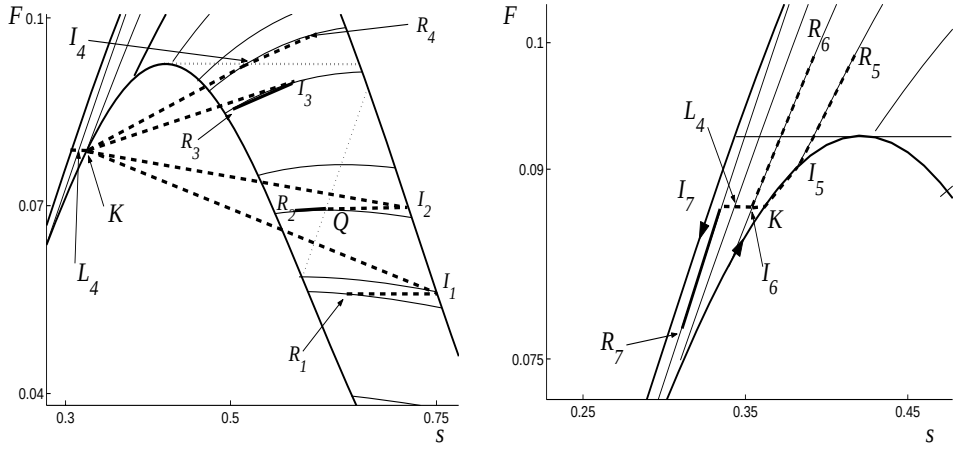


FIG. 3.10. *Figure* (a) *(Case D): L_4 in L-region D and R in R-regions $R_3 \in \Re^{III}$, $R_4 \in \Re^{IV}$, $R_5 \in \Re^{V}$, and $R_6 \in \Re^{VI}$. Figure* (b) *(Case D): $L_4 \in D$ and R in R-regions $R_1 \in \Re^{I}$ and $R_2 \in \Re^{II}$.*

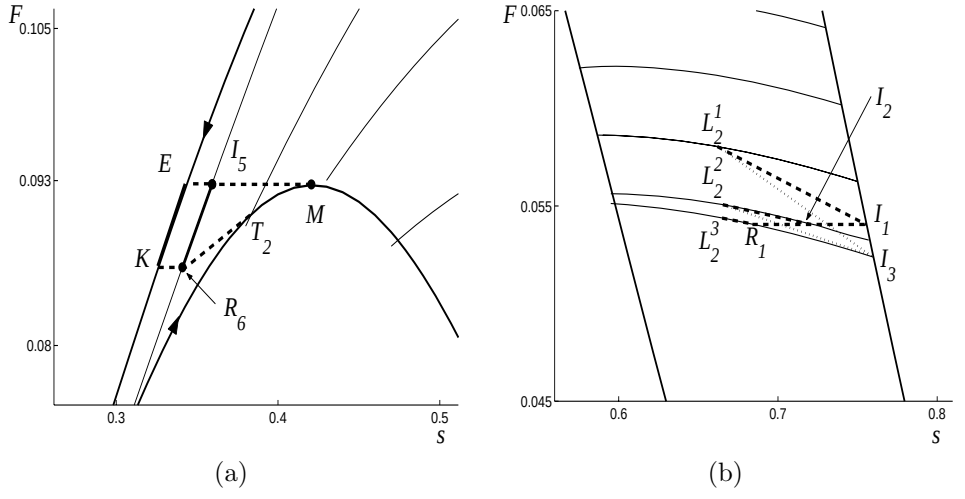
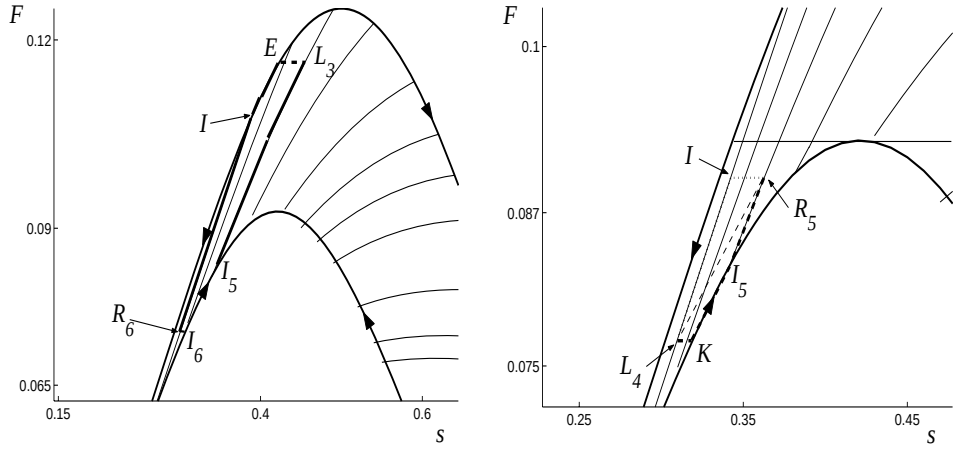


FIG. 3.11. *Figure* (a) *(Case A): $R_6 \in \Re_A^{VI}$. Figure* (b) *(Case B): $L_2^1 \in B$ and $R_1 \in \Re_B^I$.*

Case A: $L_1 \in A$ and $R_6 \in \Re_A^{VI}$. As mentioned in section 3.2, the solution for this case is represented by the segment $L_1TMI_5R_6$ as shown in Figure 3.6(b). We show that other tentative solutions are impossible. In Figure 3.11(a) we analyze the connection between states M and R_6. Consider the curve MT_2R_6: the rarefaction MT_2 and shock T_2R_6. This solution is inadmissible because the rarefaction MT_2 violates the physical orientation of the drainage curve. Another tentative solution such as $MEKR_6$ is also inadmissible because the interaction of the waves EK and KR_6 yields the shock ER_6, which violates the scanning region RH condition (3.6), (3.7). Therefore, between states M and R_6, the sequence $LTMIR_6$ is the only admissible solution that we were able to find.

Case B: $L_2^1 \in B$ and $R_1 \in \Re_B^I$ (see Figure 3.7(b)). We choose the Riemann solution $L_2^1I_1R_1$. We can also consider $L_2^1I_3R_1$ as another solution (see Figure 3.11(b)).



(a) (b)

FIG. 3.12. *Figure* (a) *(Case C): $L_3 \in C$ and $R_6 \in \Re_C^{VI}$. Figure* (b) *(Case D): $L_4 \in D$ and $R_5 \in \Re_D^V$.*

In principle $L_2^1 I_1 R_1$ and $L_2^1 I_3 R_1$ seem possible. However, only the solution $L_2^1 I_1 R_1$ depends continuously on changes of the left and right states. To analyze these solutions we change the state L_2^1. In this case the solution $L_2^1 I_3 R_1$ generates a sequence of solutions

$$L_2^1 I_3 R_1 \to L_2^2 I_3 R_1 \to L_2^3 I_3 R_1, \tag{3.10}$$

which converges to the wrong solution $L_2^3 I_3 R_1$. Notice that when $L_2^1 = L_2^3$ the correct solution is a shock $L_2^3 R_1$. Additionally, the solution $L_2^1 I_1 R_1$ generates a sequence of solutions

$$L_2^1 I_1 R_1 \to L_2^2 I_2 R_1 \to L_2^3 R_1 \tag{3.11}$$

converging to the correct solution $L_2^3 R_1$ without producing $\mathcal{L}_1^{loc}$ discontinuities in the Riemann solution. Therefore, solutions other than $L_2^1 I_1 R_1$ do not ensure the continuity of the solution when the states L and R are perturbed.

Case C: $L_3 \in C$ and $R_6 \in \Re_C^{VI}$. In this case the orientation of the drainage and imbibition curves ensures the uniqueness and $\mathcal{L}_1^{loc}$ continuity of the Riemann solution. For example, in Figure 3.12(a), the solution for left and right states L_3 and R_6 is $L_3 E I R_6$; notice that another tentative solution such as $L_3 I_5 I_6 R_6$ is inadmissible because the rarefaction $I_5 I_6$ violates the drainage curve orientation.

Case D: $L_4 \in D$ and $R_5 \in \Re_D^V$. From Figure 3.12(b), the solution chosen is $L_4 K I_5 R_5$. A tentative solution such as $L_4 I R_5$ is not possible because its yields a shock $L_4 R_5$ which violates the scanning region Oleinik condition.

4. The corrected Godunov method. We discretize the z-t plane by choosing a mesh width $h := \Delta z = 1/N_z$ and a time step $k = \Delta t$, and we define the discrete grid points (z_j, t_n) by

$$z_j = jh + h/2,\ z_{j\pm 1/2} = z_j \pm h/2,\ \ j = 0, 1, 2, \ldots, N_z - 1, \tag{4.1}$$

$$t_n = nk,\ \ n = 0, 1, 2, \ldots . \tag{4.2}$$

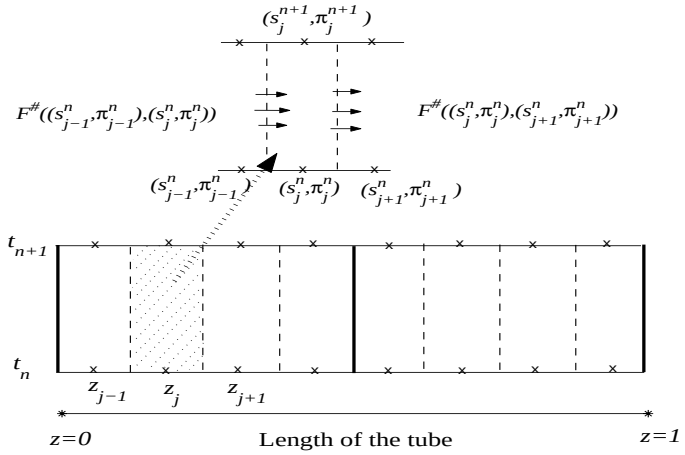


FIG. 4.1. *Numerical fluxes using local Riemann solution.*

Each time step of the numerical method consists of two stages. In the first stage (*predictor*) equations (3.1) and (3.2) are solved numerically by the Godunov method [11]. The second stage (*corrector*) is the correction of the parameter π, which is necessary when the predicted state lies outside the admissible region Ω. In such a case, the corrected state is obtained from the predicted state by a projection (with fixed s) on $\partial\Omega$ of the predictor state. The saturation is maintained constant in order to preserve conservation of mass, which is guaranteed in the prediction step because it uses Godunov method.

Predictor. Denoting the cell state at (z_j, t_n) by (s_j^n, π_j^n), we then have that the conservative Godunov method on a cell $[z_{j-1/2}, z_{j+1/2}] \times [t_n, t_{n+1}]$ for the saturation value is expressed as

$$s_j^{n+1} = s_j^n - \frac{k}{h}[F^{\#}((s_j^n,\pi_j^n),(s_{j+1}^n,\pi_{j+1}^n)) - F^{\#}((s_{j-1}^n,\pi_{j-1}^n),(s_j^n,\pi_j^n))], \tag{4.3}$$

where $F^{\#}((s_j^n,\pi_j^n),(s_{j+1}^n,\pi_{j+1}^n))$ and $F^{\#}((s_{j-1}^n,\pi_{j-1}^n),(s_j^n,\pi_j^n))$ are the numerical fluxes at the right and left boundaries of the cell $[z_{j-1/2}, z_{j+1/2}] \times [t_n, t_{n+1}]$, respectively. We impose zero numerical flux boundary conditions at $z = 0$ and $z = 1$; see Figure 4.1.

Notice that the left and right numerical fluxes of each cell are constant along the left and right boundaries, respectively. This is so because these boundaries coincide with zero speed characteristics at $z_{j-1/2}$ and $z_{j+1/2}$, respectively. Consequently, for each pair of states (s, π), the numerical fluxes can be determined directly from the Riemann solution presented in section 3.2.

For example, consider the left state $L = (s_L, F_L)$ and right state $R = (s_R, F_R)$ at time t_n (see Figure 4.1), with the Riemann solution represented in Figure 5.1(b). Then the numerical flux $F^{\#}((s_L^n,\pi_L^n),(s_R^n,\pi_R^n))$ is exactly the flux F_M^d specified by the state M or I; notice that either choice produces the same numerical flux as shown in Figure 5.1(b). As another example, consider the left state L and right state R, and suppose that the Riemann solution consists of a positive speed shock represented in Figure 3.10(b) by $L_4KI_5R_5$. Consequently, the numerical flux $F^{\#}((s_L^n,\pi_L^n),(s_R^n,\pi_R^n))$ is $F(s_L, \pi_L)$ at state L_4. In this way, the numerical flux for each pair of left and right states is chosen by using the solutions presented in section 3.2.

Corrector. Once (4.3) for the saturation is satisfied, we update π_j^n. We de-

fine the external imbibition and drainage regions as $\Omega^i := \{(s,F) \in \Re^2 : 0 \le s \le 1 \text{ and } F^i(s) \le F\}$ and $\Omega^d := \{(s,F) \in \Re^2 : 0 \le s \le 1 \text{ and } F \le F^d(s)\}$, respectively. To obtain π_j^{n+1}, we choose $\pi_j^{n+1} = \pi_j^n$ if (s_j^{n+1}, π_j^n) lies in the scanning region Ω, and $\pi_j^{n+1} = \pi^i(s_j^{n+1})$ or $\pi_j^{n+1} = \pi^d(s_j^{n+1})$ if $(s_j^{n+1}, F(s_j^{n+1}, \pi_j^n))$ lies on the external imbibition or drainage region, respectively. This strategy for updating π was proposed in [18] and guarantees mass conservation of each phase. For the computational results presented in section 5, we adopt the global CFL condition as $|\frac{k}{h} v_{\max}| \le 1$, where

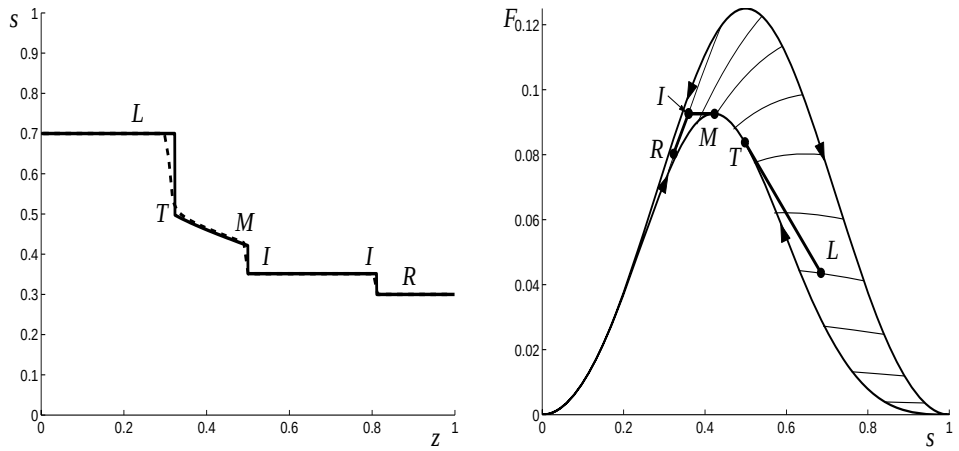
$$v_{\max} = \max\left(\max_{(s,\pi)\in\Omega} \mid \partial_s F(s,\pi) \mid, \max_{(s,F)\in j} \mid \partial_s F^j(s) \mid\right), \; j = i, d. \tag{4.4}$$

We note, however, that a sharper time-dependent CFL condition can be derived as the maximum velocity of all local Riemann solutions in each time step.

5. Computational results.

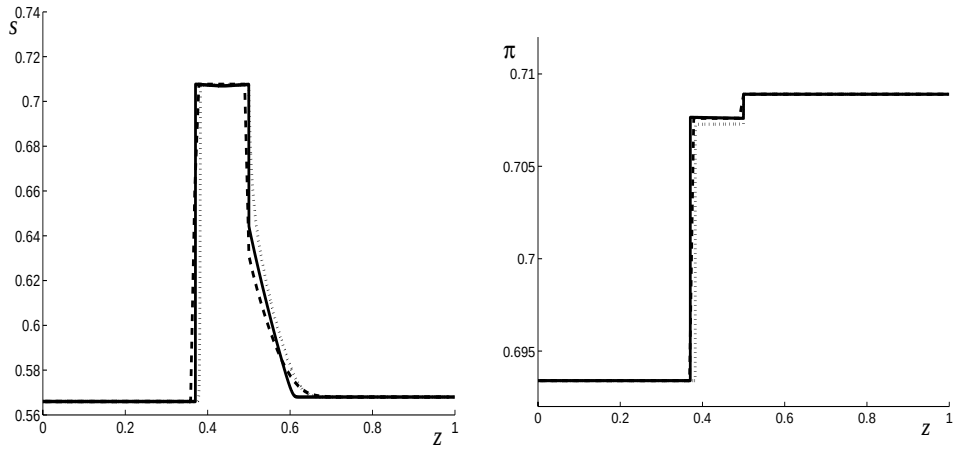
Example 1: Comparison between the numerical and the analytical solutions. The aim of this comparison is to demonstrate that the numerical method can capture accurately all the features of the analytical solution. We consider the numerical and analytical dimensionless solutions associated to the Riemann problem with initial discontinuity at $z = 0.50$ separating states $(s_l, \pi_l) = (0.7, 0.8)$ and $(s_r, \pi_r) = (0.3, 0.4)$. Both states lie inside the scanning region. The following parameters were used in the simulations $\mu_w = 1\ cp$, $\mu_o = 0.9\ cp$, and for this case the global CFL condition is 0.3849. The dimensionless simulated time was $t = 2.4\ 10^3$. The analytical solution (represented in Figure 5.1(b)) is shown as a solid line in Figure 5.1(a), and the numerical solution obtained with the proposed scheme using $N_z = 100$ as a dashed line in the same figure.

The numerical and analytical profiles at dimensionless time are presented in Figure 5.1. The saturation profile consists from left to right of a scanning to drainage shock wave LT, followed by a rarefaction wave TM, a stationary wave MI, and a rarefaction wave IR. The small discrepancy around the shock LT in Figure 5.1(a) is due to the correction scheme for π. However, the discrepancy potentially occurs at only one mesh point.



(a) (b)

FIG. 5.1. *Comparison between the analytical and the numerical method. Dashed line corresponds to the numerical solution, and solid corresponds to analytical solution.*



(a) (b)

FIG. 5.2. *The solid curve corresponds to the analytical solution, the dashed curve to the corrected Godunov method with* $N_z = 50$, *and the dotted curve to the corrected Lax–Friedrichs scheme with* $N_z = 4000$. (a) *Saturation values and* (b) π *values.*

Example 2: Comparison between corrected Godunov and corrected Lax–Friedrichs schemes. We consider the Riemann problem with initial discontinuity at $z = 0.50$ separating states $(s_l, \pi_l) = (0.57, 0.69)$ and $(s_r, \pi_r) = (0.57, 0.71)$. Figure 5.2(a) and (b) show a comparison among the proposed corrected Godunov scheme with $N_z = 40$, the corrected Lax–Friedrichs with $N_z = 4000$, and the analytical solution. The global CFL restriction is followed. To obtain an accurate approximation for the analytical solution, the corrected Godunov scheme requires a mesh size $N_z = 40$, while the corrected Lax–Friedrichs requires $N_z = 4000$. Hence, to satisfy the global CFL condition (4.4), the corrected Godunov scheme requires substantially fewer time steps than the corrected Lax–Friedrichs, so that the simulation using the Godunov scheme takes 500 times less CPU time than using the Lax–Friedrichs scheme.

The classical Godunov and Lax–Friedrichs schemes do not work without the corrector step. We note also that the design of numerical schemes for the problems considered here is not trivial. Because of hysteresis, there are stationary waves, the orientation of the curves (imbibition and drainage) has to be taken into account, and the restriction on the admissibility scanning region Ω must be followed.

Example 3: Comparison between solutions with and without hysteresis. The strong influence of the hysteresis effect in the saturation profiles can be seen by comparing the profiles obtained numerically with and without hysteretic relative permeabilities. We choose the drainage curve opposite the imbibition curve as the nonhysteretic relative permeability curve since the corresponding nonhysteretic solution agrees most with the hysteretic solution.

We introduce in the simulation the top and bottom boundaries of the tube. We perform the simulation with $N_z = 50$, $\mu_w = 1$ *cp*, and $\mu_o = 0.9$ *cp*. The density of the wetting phase is $\rho_w = 1$ g/cm^3 and the density of the nonwetting phase is taken as $\rho_o = 0.844$ g/cm^3. We consider the porous media having a permeability of 11.84 Darcys and porosity of 0.389. The tube length is 86 cm. Additionally, $s_L = 0.9$ and $s_R = 0.1$.

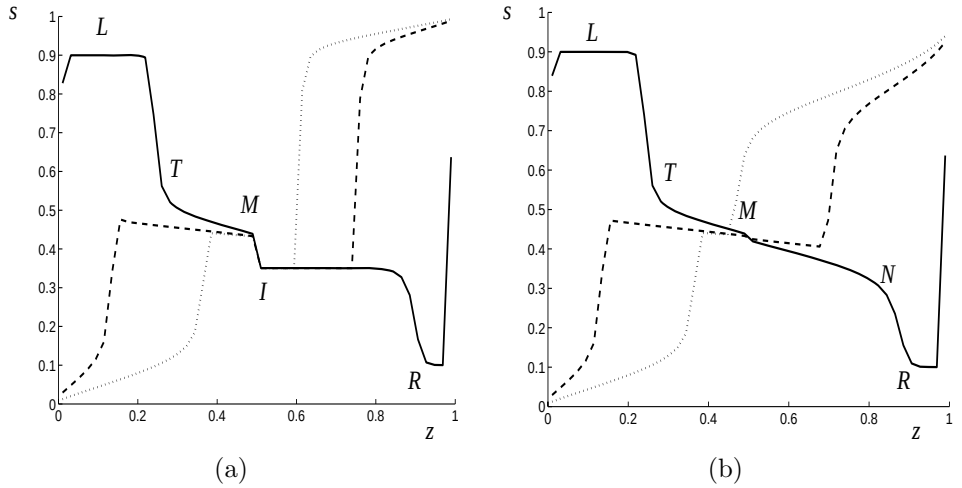


FIG. 5.3. *Profile $s(z)$ at several times:* (a) *hysteretic solution,* (b) *nonhysteretic solution. Saturations are represented as a solid curve at* 5.87 h*, as a dashed curve at* 14.59 h*, and as a dotted curve at* 19.38 h*. Compare with Figure* 5.4.

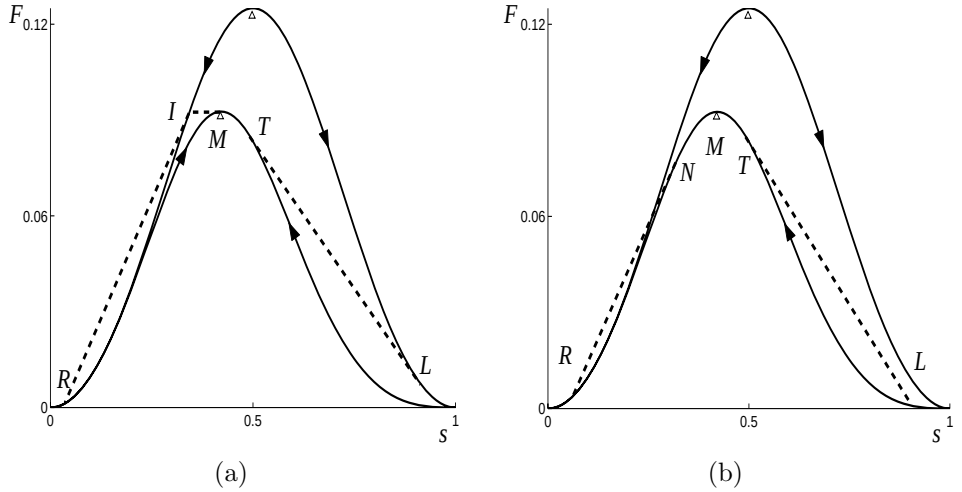


FIG. 5.4. *Solutions in (s, F) space* (a) *with hysteresis,* (b) *without hysteresis at imbibition and drainage curves.*

The hysteretic and nonhysteretic simulation results differ significantly; see Figure 5.3. Observing Figure 5.3(a), three sections are clearly identified. We can see that in the top zone (around $z = 0$), wetting phase saturation decreases with time, i.e., a drainage process. At the bottom zone (around $z = 1$), the saturation increases with time, i.e., an imbibition process. The most relevant discrepancy between hysteretic and nonhysteretic solutions occurs at the middle zone ($z \in [0.3, 0.7]$), where both drainage and imbibition take place. In this zone the hysteretic profiles have a sharper decline MI, which is not captured in Figure 5.3(b) without hysteresis.

The discrepancy between the hysteretic and nonhysteretic solutions can be under-

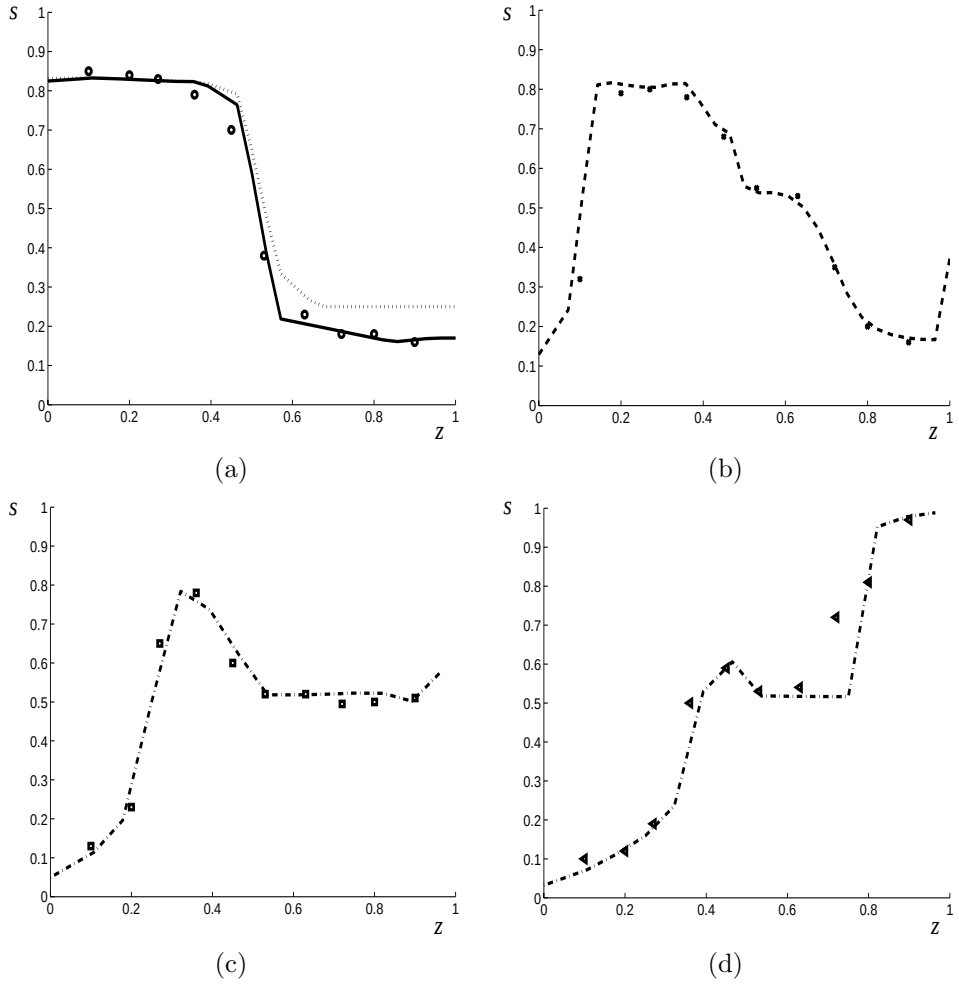


FIG. 5.5. *Comparison between saturation profiles obtained by the numerical simulation and the laboratory data (marked points) several times (*(a) 0 *min.,* (b) 19 *min.,* (c) 60 *min.,* (d) 38.6 *hours). In* (a)*, the dotted line is the initial value of* π *used for the simulation.*

stood by comparing the Riemann solution for hysteretic and nonhysteretic fractional flow functions. With hysteresis (Figure 5.4(a)), the solution combining imbibition and drainage necessarily goes through the scanning region. For instance, consider the case when $s_L > s_R$. Due to the imbibition and drainage curve orientation, the solution is $LTMIR$, a negative speed shock LT, a drainage rarefaction TM, a stationary shock wave MI connecting the drainage and the imbibition curves, and a positive speed imbibition shock IR. On the other hand, without hysteresis (Figure 5.4(b)), the solution is $LTNR$, a negative speed shock LT, a rarefaction wave TN, and a positive speed shock NR.

Example 4: Comparison with experimental work data. To validate the proposed model we compare the numerical gas-water saturation profile with saturation profiles found in [7] and [20]. Following [7], we consider a sand pack with a permeability of 11.84 Darcys and porosity of 0.389. The tube length is 86 cm. Additionally,

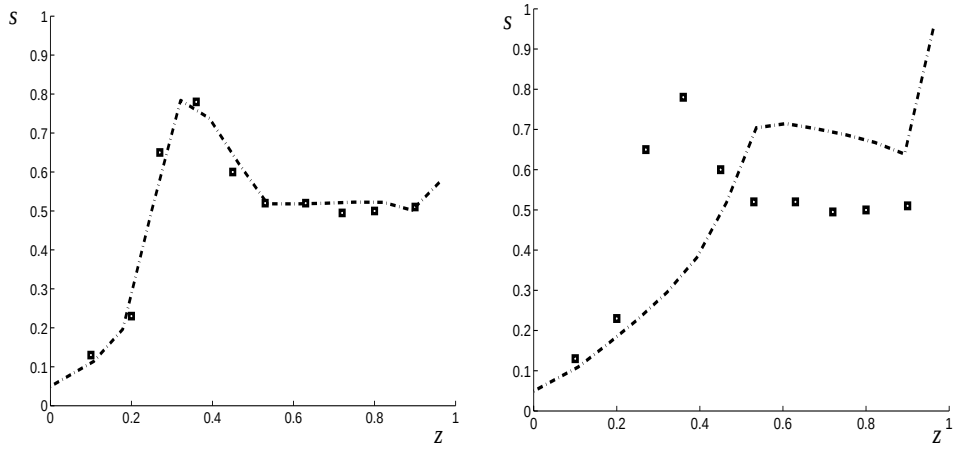


FIG. 5.6. *Comparison between saturation profiles with hysteresis* (a) *and without hysteresis* (b) *obtained by the numerical simulation and the laboratory data (marked points) for* 60 *min.*

we use the following values for the parameters: $\alpha = 0.5$, $\gamma = 0.5$, $\mu_w = 0.8550$ *cp*, and $\mu_o = 0.0185$ *cp*. The density of the wetting phase (water) is $\rho_w = 1$ g/cm^3 and the density of the nonwetting phase (air) is considered negligible compared to ρ_w. These values were taken from [7].

The procedure consists in injecting a measured amount of radioactive water into the tube containing air in its pores, closing it off at both ends, and allowing the two phases to reach the equilibrium distribution. This saturation distribution is measured, providing the initial distribution in the simulation. This distribution is a spatial transition between air in the presence of residual water to water in the presence of residual air. This water saturation distribution results from the balance between pressure gradient and the buoyancy force.

Usage of the correct initial distribution of π is crucial to reproducing the results presented in [7]. To obtain this initial distribution from the procedure described in [7], we consider the injection of a certain quantity of water from the top of an air-saturated tube. Consequently, the lower half of the tube corresponds to an imbibition process and the upper half corresponds to a drainage process. The values of π are chosen accordingly. This choice of initial values for π are shown by the dotted line in Figure 5.5(a).

The tube is then inverted and measurements of water saturation are made during the segregation process at ten different locations along the tube. Figures 5.5(a)–(d) confirm the excellent agreement between the laboratory data and the simulated saturation profiles at different times, while Figures 5.6(a) and (b) show the importance of including the hysteresis effects in the model. There is agreement in Figure 5.6(a) with hysteresis and disagreement in Figure 5.6(b) without hysteresis.

6. Concluding remarks. We present Riemann solutions for each left and right state for a hysteretic counterflow segregation problem as well as criteria to guarantee well-posedness to the solution. Based on the Riemann solutions, we propose a corrected Godunov scheme that updates both the saturation and the hysteretic parameter. This scheme conserves mass locally. Numerically, we show that the solution

obtained by the proposed method agrees with the analytical solution. We also validate the numerical scheme by comparing simulations with laboratory experimental data. We show numerically that the inclusion of hysteresis effects in the relative permeability suffices for simulations for accurate simulations.

REFERENCES

[1] J. E. F. Altoé, *Modelagem Analítica do Processo de Migração Secundária de Hidrocarbonetos*, Master of Sciences Thesis, Universidade Estadual do Norte Fluminense, LENEP, Macaé, Rio de Janeiro, Brazil, 2002.

[2] K. Aziz and A. Settari, *Petroleum Reservoir Simulation*, Applied Science, London, 1979.

[3] J. Bear, *Dynamics of Fluids in Porous Media*, Elsevier, New York, 1972.

[4] P. Bedrikovetsky, *Mathematical Theory of Oil and Gas Recovery*, Kluwer Academic, London, 1993

[5] P. Bedrikovetsky, D. Marchesin, and P. R. Ballin, *Mathematical theory for two-phase displacement with hysteresis (with application to WAG injection)*, in Proceedings of the 5th European Conference on the Mathematics of Oil Recovery, Leoben, Austria, 1996.

[6] E. M. Braun and R. F. Holland, *Relative permeability hysteresis: Laboratory measurements and a conceptual model*, SPE Reservoir Engineering, 10 (1995), pp. 222–228 (paper 28615).

[7] J. E. Briggs and D. L. Katz, *Drainage of Water from Sand in Developing Aquifer Storage*, in Proceedings of the Fall Meeting of the Society of Petroleum Engineers of AIME, Dallas, TX, Society of Petroleum Engineers of AIME, 1966 (paper 1501-MS).

[8] J. Colonna, F. Brissaud, and J. L. Millet, *Evolution of capillarity and relative permeability hysteresis*, 253 (1972), pp. 28–38 (paper 2941-PA).

[9] R. E. Ewing, *The Mathematics of Reservoir Simulation*, Frontiers Appl. Math. 1, SIAM, Philadelphia, 1984.

[10] K. M. Furati, *Effects of relative permeability history dependence on two-phase flow in porous media*, Transp. Porous Media, 28 (1997), pp. 181–203.

[11] S. K. Godunov, *A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics*, Mat. Sb., 47 (1959), pp. 271–290.

[12] J. E. Killough, *Reservoir simulation with history-dependent saturation functions*, SPE J., 1976, pp. 37–48 (paper 5106-PA).

[13] A. Kjosavik, J. K. Ringen, and S. M. Skjaeveland, *Relative permeability correlation for mixed-wet reservoirs*, SPE J., 7 (2002), pp. 49–58 (paper 77328-PA).

[14] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser, Berlin, 1992.

[15] H. B. Medeiros, D. Marchesin, and P. L. Paes Leme, *Hysteresis in two-phase flow: A simple mathematical model*, Comput. Appl. Math., 17 (1998), pp. 81–99.

[16] O. A. Oleinik, *Uniqueness and stability of the generalized solution of the Cauchy problem for a quasilinear equation*, Uspekhi Mat. Nauk. 14 (1959), pp. 165–170. English translation in Amer. Math. Soc. Transl. Ser. 2, 33, AMS, Providence, RI, 1964, pp. 285–290.

[17] W. W. Owens and D. L. Archer, *The effect of rock wettability on oil-water relative permeability relationships*, J. Petroleum Tech. AIME, 1971, pp. 873–878 (paper 3034-PA).

[18] B. Plorh, D. Marchesin, P. Bedrikovetsky, and P. Krause, *Modeling hysteresis in porous media flow via relaxation*, Comput. Geosci., 5 (2001), pp. 225–256.

[19] N. Shahidzadeh-Bonn, E. Bertrand, J. P. Dauplait, J. C. Vié Borgotti, and D. Bonn, *Gravity drainage in porous media: The effect of wetting*, J. Petroleum Science & Engineering, 1085 (2003), pp. 1–8.

[20] E. E. Templeton, R. F. Nielsen, and C. D. Stahl, *A study of gravity counterflow segregation*, SPE J., 1962, pp. 186–193 (paper 186-PA).

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1533–1562

A SEMIANALYTICAL THERMAL STRESS MODEL FOR THE CZOCHRALSKI GROWTH OF TYPE III-V COMPOUNDS*

C. SEAN BOHUN†, IAN FRIGAARD‡, HUAXIONG HUANG§, AND SHUQING LIANG§

Abstract. In this paper we describe a semianalytical approach to computing the temperature and thermal stress inside a III-V compound grown with the Czochralski technique. An analysis of the growing conditions indicates that the crystal growth occurs on the conductive time scale. A perturbation method for the temperature field is developed for an arbitrary crystal profile using the Biot number as a (small) expansion parameter. The zeroth order solution is one-dimensional in the axial direction. Explicit solutions are obtained for a cylindrical and a conical crystal. Under typical growth conditions, a parabolic temperature profile in the radial direction is shown to arise naturally as the first order correction. As a result, the thermal stress is obtained explicitly and its magnitude is shown to depend on the zeroth order temperature and Biot number. Both the axial temperature gradient and crystal profile are shown to be important for controlling thermal stress and defect density. Some issues relevant to growth conditions are also discussed.

Key words. crystal growth, asymptotic expansion, moving interface, thermal stress, dislocation density, finite difference method, Czochralski technique

AMS subject classifications. 74A10, 74F05, 74H10, 80A22, 82D25, 82D37, 65M06

DOI. 10.1137/S0036139903436455

1. Introduction.

Directional solidification methods are widely used for growing large industrial sized crystals. Among them, the Czochralski (Cz) method is the most popular technique for growing crystals used by the semiconductor and related industries. By dipping a small seed crystal into a pool of molten material in the crucible and carefully controlling the heat balance inside the grower, a large crystal can be grown by pulling the crystal away from the melt in a slow and steady fashion. The pulling rod and the crucible are normally rotated in opposite directions during the growth period. Delicate control is often needed to maintain the crystal quality, and a slight change of the growth conditions may result in defect formation inside the crystal. With care, a single crystal with low defect density can be obtained routinely when the size of the crystal does not exceed a critical value. For a more detailed account of the Cz and other techniques, we refer the readers to the extremely informative handbooks by Hurle [16, 17].

Due to the complex nature of the thermal, structural, and dynamic coupling of the molten material, the crystal, the crucible, the gas chamber, and other parts of the grower, considerable efforts have been devoted to laboratory experiments and to modeling and simulation of the growth environment over the past several decades. As a result, there exists an extensive literature, mostly in engineering fields. These studies

*Received by the editors October 23, 2003; accepted for publication (in revised form) March 15, 2006; published electronically May 26, 2006. This work was supported by Firebird Semiconductors Ltd., the Mathematics of Information Technology and Complex Systems (MITACS), a Network of Centers of Excellence and the Nature Sciences and Engineering Research Council (NSERC) of Canada, and BCASI.

http://www.siam.org/journals/siap/66-5/43645.html

†Department of Mathematics, Pennsylvania State University, Mont Alto, PA 17237 (csb15@psu.edu).

‡Department of Mathematics, University of British Columbia, Vancouver V6T 1Z2, BC, Canada (frigaard@math.ubc.ca).

§Department of Mathematics and Statistics, York University, Toronto M3J 1P3, ON, Canada (hhuang@yorku.ca, sqliang@mathstat.yorku.ca).

cover a wide spectrum of areas, from decoupled one- or two-dimensional simulations to fully coupled three-dimensional computations; see, e.g., [4, 5, 16, 17, 18, 26, 27, 29]. Most of the studies rely heavily on computer simulation since the fully coupled system cannot be solved otherwise. These investigations have generated useful information including temperature distribution, crystal-melt interface shape, and melt flow patterns inside the crucible. By comparison, much less attention has been paid to the coupling of defect modeling and field variables even though significant progress has been made in identifying main factors that determine the formation of defects [31].

In this paper, we present a semianalytical approach for studying the temperature field inside the crystal and the related thermal stress. It is believed that defect formation can be related to an excessive thermal stress above some critical value; see, e.g., [1, 11, 13, 19, 29, 33, 34] and the references therein. Therefore, analysis of the growth factors that determine the stress level will be extremely useful for crystal growers. The stress analysis requires that a particular crystal structure be specified, and we have chosen the ZnS structure shared by the type III-V binary semiconductors. Even though the basic mathematical structure remains the same for any III-V compound, we will focus on indium antimonide (InSb) for the rest of the paper. InSb has the narrowest bandgap and highest temperature mobility of the III-V binary compound semiconductors. Because of these properties, InSb is widely used in both magnetic field detectors and infrared sensors. A review of these properties can be found in Micklethwaite and Johnson [24].

The primary reason for focusing on InSb is that it is exceedingly difficult to grow with the Cz technique mainly because of its small critically resolved shear stress (CRSS). It is known experimentally that attempting to grow InSb in a cylindrical profile with the Cz technique produces crystals with an unacceptable defect density, contrary to the growth of more common crystals such as silicon, where low defect density crystals can be grown in a cylindrical shape. Thus it is often an art to find the most suitable profile of the solidifying crystal by carefully varying the furnace temperature and the rate that the crystal is extracted [23]. To determine the influence of various resulting crystal profiles (i.e., axial variations of the lateral surface or crystal shapes) on the stress experienced within the crystal, we assume that the profile of the crystal is an arbitrary function of the axial displacement while allowing the solid-liquid interface to be driven by a Stefan condition and a compatibility condition at the solid-liquid-gas triple point.

By examining the physical process and parameter values of the growth environment closely, we are able to identify the main features associated with InSb crystals. In particular, if the heat flux from the melt is uniform across the crystal-melt interface, the temperature field will be dominated by the lateral flux through the crystal-gas surface, characterized by a nondimensional Biot number. The value of the Biot number is small under the growth conditions for InSb crystals, suggesting an asymptotic expansion of the solution in terms of this parameter. Much of the asymptotic framework discussed in this paper has appeared elsewhere in the literature [3, 14, 20, 35, 36, 37, 38, 39]. For example, Kuiken and Roksnoer [20] assumed a pseudosteady solid-liquid interface to obtain an accurate temperature distribution of a Si crystal grown with the floating-zone technique. Their solution takes the form of an expansion in terms of the Peclet and Nusselt numbers of the crystal, giving a solution valid for slender crystals grown in conductive heat transfer environments. By specifying an externally defined solid-liquid interface shape these authors avoided using a Stefan condition to evolve the interface. An asymptotic analysis that con-

sidered the melt was undertaken by Brattkus and Davis [3], where the geometry allowed an expansion in terms of the aspect ratio of the solidification cell. Young and Chait [36] considered a system driven by surface tension and more recently Young and Heminger [37, 38] have utilized a small aspect ratio to study the growth of single crystal fibers.

This paper blends the asymptotic expansion with the plane strain approximation to examine thermal stress inside the crystals. It contrasts significantly with the above references by assuming a radially independent heat flux from the melt that avoids a boundary layer analysis around the solidification front, greatly simplifying the asymptotics. Any angular dependence is minimized by the rotation of the seed and crucible, and the experimental evidence of an almost flat interface for Cz grown InSb crystals [23] suggests that the flux from the melt is likely to be largely independent of the radius. This is also supported in the literature [8, 9]. On the other hand, if the heat flux is radially dependent then the same asymptotic framework applies, but there will be a boundary layer solution similar to that in [3, 36] or [20] on the crystal side to match the asymptotic solutions.

While the details of the motion of the melt are ignored, the crystal-melt and crucible-melt heat transfer coefficients are estimated from the Ekman layer and natural convection submodel, respectively. As the crystal grows two situations are investigated. In the first, the flux is constant so that the model is more accurately a description of the solid in a directional solidification technique. In the second, the flux is calculated through an essentially zero-dimensional model for the melt. This enables the heat flux from the melt to be influenced by some of the crucible and operating parameters of the Cz process.

In the fully unsteady case, the asymptotic expansion results in a system of one-dimensional equations and the thermal stress can be obtained explicitly in an analytical form, under the plane strain assumption. In the pseudosteady limit this reduces to the classical result that the stress is proportional to the concavity of the temperature field [22, 34]. This also extends the work of [19], where stress was obtained for a cylindrical crystal with a flat crystal-melt interface. The pseudosteady solutions are contrasted with the unsteady solutions for cylindrical and conical crystals justifying the pseudosteady approximation for the growth parameters of InSb used in this paper. A detailed description of the pseudosteady approximation with respect to Cz growth can be found in the paper by Derby and Brown [10]. Other examples of the use of the pseudosteady approximation can be found elsewhere [14, 35].

Compared to most of the previous work using asymptotic or numerical methods, this study moves a step further by coupling stress calculation with the asymptotic field temperature solution and deriving an explicit form for the stress. Furthermore, formulated in a nondimensional form, the dependence of the stress level on the Biot number is useful for crystal growers when larger crystals are grown. Since the Biot number is proportional to the product of the heat transfer coefficient and the mean crystal radius, it is obvious that one should try to reduce the heat flux via the lateral surface when a crystal of larger radius is grown. In addition, the explicit nature of the stress solution enables us to identify the effect of crystal profile (shape) as well as the crystal size (radius) on the stress. More importantly, obtaining an explicit formulation for the stress allows us to apply other techniques such as optimal control methodologies to efficiently search for better growth conditions. A simplified model for the heat exchange between the gas flow and the crystal is used to clarify the presentation. However, the asymptotic solution developed here is still valid if a more

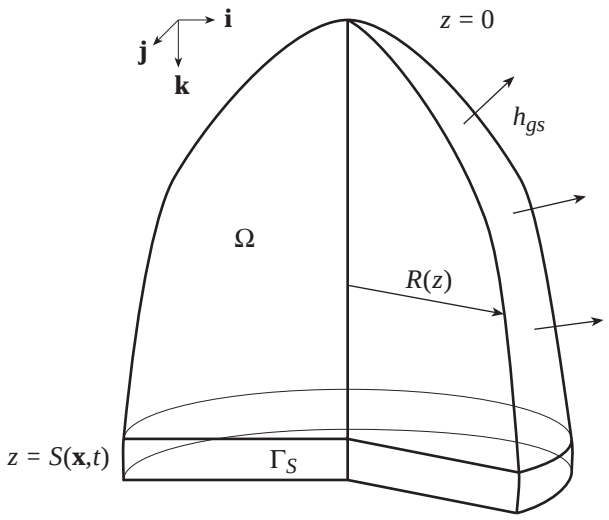


FIG. 1. *Shown is a typical crystal at some time t during a growth run with a newly solidified portion at* $z = S(\mathbf{x}, t)$. *The coordinate system is chosen so that the top of the crystal remains at* $z = 0$ *and the solidification front grows downward in the positive z direction. The radial profile is given by* $R(z)$ *and the crystal length is* $S(\mathbf{x}, t)$. *Finally, the heat transfer coefficient* h_{gs} *may be a function of the axial position z.*

realistic model for the gas is incorporated. More detailed discussion related to the growth conditions for InSb will be given in sections 2 and 5.

The rest of the paper is organized as follows. In section 2, we will present the mathematical model and dimensional analysis. Asymptotic solutions are given in section 3. Thermal stress is discussed in section 4. In section 5, results are presented for both pseudosteady and unsteady cases. We conclude the paper with a brief summary and discussion on future directions in section 6.

2. Mathematical model and dimensional analysis. The basic assumptions made in this study are (1) the crystal is axis-symmetric; (2) the heat exchange between the crystal and gas along the lateral surface of the crystal is a constant; (3) the heat exchange between the crystal and melt along the crystal-melt interface is uniform; (4) the mean crystal radius is small compared to its length; (5) thermal stress is elastic and computed under a plane strain assumption. Some of the assumptions are made to simplify the derivation, and others are made based on previous study of similar problems or observations made by us and engineers we have been collaborating with. We will revisit some of the assumptions in section 6.

Figure 1 illustrates the geometry of a typical crystal. The coordinate system is fixed to the top of the growing crystal at $z = 0$, the final length of the crystal is denoted Z, and the crystal radius is denoted $R(z)$. The growth starts with a seed crystal with radius of order $R_0 = 0.5$ cm and length $Z_0 = 3$ cm. The crystal grows outward in a slowly developing cone, eventually reaching a radius $R(Z) \simeq 5$ cm after a length $Z \simeq 30$ cm. A crystal can take 10–20 hours to grow. Thus, at the outset we make two observations. First, the crystal growth is characterized by a large aspect ratio. Second, it is evident that any transients in the system, unless caused by rapidly changing boundary conditions, are very slow. These two features will be used to derive our eventual model.

Within the crystal Ω, the temperature $T(\mathbf{x}, t)$ satisfies the heat equation

$$\rho_s c_s \frac{\partial T}{\partial t} = k_s \sum_j \frac{\partial^2 T}{\partial x_j^2}, \qquad \mathbf{x} \in \Omega,\ t > 0, \tag{1}$$

where ρ_s, c_s, and k_s, respectively, are the density, specific heat, and thermal conductivity of the crystal solid phase. The lateral surface of the crystal is denoted Γ_g and is subjected to cooling from the circulating chamber gases and from radiative heat losses. Although radiation is not insignificant, for simplicity we model both effects through a simple Newtonian cooling law

$$-k_s \frac{\partial T}{\partial \mathbf{n}} = h_{gs}(T - T_g), \qquad \mathbf{x} \in \Gamma_g. \tag{2}$$

Here we assume that the heat transfer coefficient h_{gs} incorporates both convective and radiative heat transfer (the latter via linearization). The top of the crystal is fixed at $z = 0$, where we also invoke a Newtonian cooling law

$$k_s \frac{\partial T}{\partial z} = h_{ch}(T - T_{ch}), \tag{3}$$

in the case that the radius at $z = 0$ is assumed to be nonzero. Here h_{ch} represents the heat transfer coefficient for the seed-chuck connection and T_{ch} is the chuck temperature.

The crystal-melt interface is denoted Γ_S and is where $T = T_m$, the melting temperature. The interface of the phase transition is thus implicitly defined from the temperature field. Explicitly we denote the melting isotherm by

$$z - S(\mathbf{x}, t) = 0, \qquad \mathbf{x} \in \Gamma_S. \tag{4}$$

The motion of the interface of the phase transition is governed by the Stefan condition

$$\rho_s L\, |\vec{v}_n| = k_s \left.\frac{\partial T}{\partial \mathbf{n}}\right|_{z \to S^-} - q_{l,n}, \tag{5}$$

where $|\vec{v}_n|$ is the speed at which the interface moves in the direction of the outward unit normal $\mathbf{n}$, L is the latent heat, and $q_{l,n}$ is the heat flux from the melt normal to the interface.

Figure 2 illustrates the triple point (TP) where the solid, liquid, and gas come into contact and the solid-liquid interface moves at a velocity $\vec{v}_n = v_n \mathbf{n}$. If $\partial S/\partial t$ denotes the speed of the interface in the $\mathbf{k}$ direction, then

$$|\vec{v}_n| = v_n = \frac{\partial S}{\partial t} \mathbf{k} \circ \mathbf{n}. \tag{6}$$

Still referring to Figure 2, the profile (shape) of the crystal $R(z)$ is determined by the motion of the TP given by

$$\left.\frac{\partial R}{\partial t}\right|_{z=S} = \tan(\theta - \theta_c) \left.\frac{\partial S}{\partial t}\right|_{r=R}, \tag{7}$$

where θ_c is the contact angle formed by the wetting fluid (melt) and the crystal and θ is the angle formed by the meniscus with the vertical z-axis. This expression simply

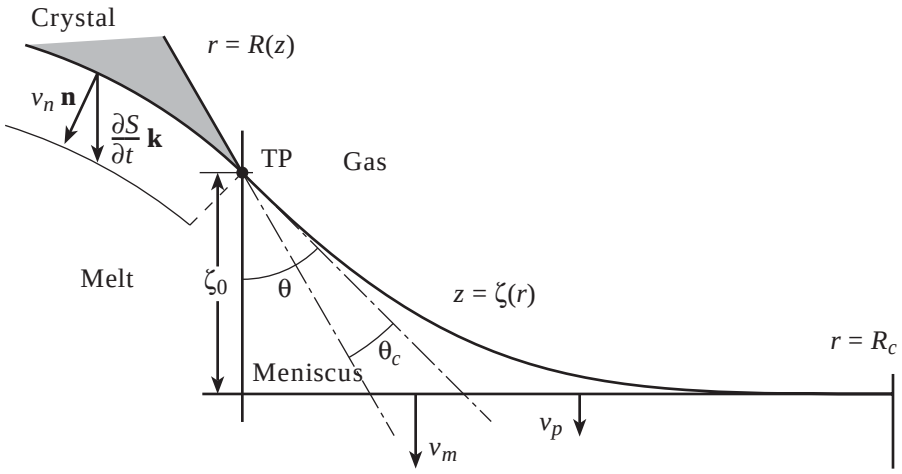


FIG. 2. *Schematic diagram of the meniscus $z = \zeta(r)$ with capillary height ζ_0, defined on $R(S) \le r \le R_c$, where $R(S)$ is the radius of the crystal at the interface, θ_c is the contact angle, and R_c is the radius of the crucible. The motion of the TP is determined by the advancing interface S, the speed at which the melt falls v_m, and the pull rate applied to the crucible v_p.*

states that the crystal prefers to grow in the direction defined by the contact angle. The motion of the advancing interface S, the advancement due to the pulling rate, and the motion due to the loss of melt determine the vertical position of the triple junction ζ_0 via

$$\frac{d\zeta_0}{dt} = v_p + v_m - \frac{\partial S}{\partial t}, \tag{8}$$

where v_m is the rate at which the melt-gas surface drops given by

$$v_m = \frac{\rho_s R^2}{\rho_l R_c^2} \frac{\partial S}{\partial t} \tag{9}$$

from the law of mass conservation and v_p is the pulling rate at which the crucible is dropped to ensure that the crystal-melt interface remains at the surface of the liquid.

We note that properly it is necessary to close the model by relating growth in S to that in R, i.e., solving (7). To do this we must model the crystal withdrawal from the crucible, the formation of the meniscus, and the coupling of S and R. It has been shown in [32] that the growth angle is related to the capillary height ζ_0 for large Bond number growth. In principle, crystals with desirable shapes can be grown by adjusting the pulling rate, which determines the meniscus angle θ in (7). Therefore, if we are not interested in the dynamics, we can impose a geometry $R(z)$ on the model. This approach has the advantage of allowing us to investigate the thermal fields and associated stresses that develop for a particular observed shape.

Note that for cylindrical crystals the capillary height does not change throughout the growth cycle. Therefore (8) can be simplified as

$$v_p = \frac{\partial S}{\partial t} - v_m. \tag{10}$$

TABLE 1
A summary of the thermophysical and typical growth parameters of InSb.

Data	Symbol	Value
Growing properties		
Mean crystal radius	$\overline{R}$	0.03 m
Final crystal length	Z	0.30 m
Characteristic growth rate	$\overline{v}$	6.9×10^{-6} m/s
Ambient gas temperature	T_g	600 K
Solid properties at $T = T_m$		
Melting temperature	T_m	798.4 K
Density	ρ_s	5.64×10^3 kg/m^3
Thermal conductivity	k_s	4.57 W/m K
Heat capacity	$\rho_s c_s$	1.5×10^6 J/m^3 K
Latent heat of fusion	L	2.3×10^5 J/kg
Heat transfer coefficients		
Crystal-gas	h_{gs}	$1 - 4$ W/m^2 K

2.1. Typical scales in InSb crystal growth. Although it is possible to treat the three-dimensional case above, it is somewhat unwieldy, and hence we instead attempt to simplify the model first. Table 1 specifies a typical set of thermophysical and process data. Consider a time t, after any initial growth transient, when the crystal has length S onto which a thin layer of crystal of radius $\overline{R}$ has just solidified. Utilizing (1) and the characteristic values in Table 1, the conduction of heat across a crystal cross-section and the time taken to grow a length $\overline{R}$ of crystal have the following time scales:

$$t_{\text{cond}} \simeq \frac{\overline{R}^2 \rho_s c_s}{k_s} = 3.0 \times 10^2 \text{ s}, \qquad t_{\text{grow}} \simeq \frac{\overline{R}}{\overline{v}} = 1.7 \times 10^4 \text{ s}.$$

Thus, the conductive time scale is typically much shorter than that for growth (i.e., over similar length scales). The growth time scale for the entire crystal is still longer and given by $t \simeq Z/\overline{v}$. It is over this latter growth time scale that significant changes in either the radius or area occur related to significant changes in the cooling capacity and heat capacity, respectively.

Therefore, apart from imposed rapid changes in the growth (e.g., at the start of the process and at the end as the crystal is withdrawn from the melt), all other thermal changes are slow and occur on the growth time scale. Since there is no process change that occurs on the conductive time scale, the process is likely to be pseudosteady on the growth related time scale.

Turning now to the thermal gradients, the magnitude of the radial variation in the temperature is maximized at the lateral surface where the crystal comes into contact with the surrounding gas. From (2) and Table 1,

$$\left|\frac{\partial T}{\partial n}\right|_{\Gamma_g} \leq \frac{h_{gs}}{k_s}(T_m - T_g) \simeq 175 \text{ K/m}.$$

The magnitude of the axial temperature gradient is maximal at the interface of the phase transition where the Stefan condition (5)–(6) is satisfied. Assuming a nearly flat interface, which will be justified later, an estimate for $T_z|_{S^-}$ is obtained by neglecting the heat flux in the liquid phase

$$\left|\frac{\partial T}{\partial z}\right|_{S^-} \simeq \frac{\rho_s L \overline{v}}{k_s} = 850 \text{ K/m}.$$

The sides of the crystal are predominantly vertical and the crystal-melt interface predominantly horizontal. Thus, we see that the vertical gradients dominate, at least in some neighborhood of the crystal-melt interface. However, we must note that the vertical gradients arise mostly due to heat loss to the cooling gases, which occurs in the radial direction. Since the cooling influences are weak, this implies that a long crystal is needed to get a significant temperature drop along the crystal length and suggests we will need to scale the axial and radial directions differently.

2.2. Nondimensionalization. The above discussion motivates our scaling below. For simplicity, we start by assuming an axisymmetric model, although the crystal cross-section is not in fact circular. The other assumptions that we make here, for simplicity only, are that the heat transfer coefficient h_{gs} and the gas temperature T_g are constant. In reality there will be local variations along the crystal surface, but in any case these require a more detailed analysis of the gas flows in order to be properly evaluated.

We define the Biot number by

$$\epsilon = \frac{h_{gs}\overline{R}}{k_s}, \tag{11}$$

and using the parameter values in Table 1, we find $\epsilon \lesssim 0.026 \ll 1$. We seek an asymptotic expansion in terms of ϵ. With this in mind we adopt the following scalings:

$$r = \overline{R}\hat{r}, \qquad R(z) = \overline{R}\hat{R}(\hat{z}), \quad \epsilon^{1/2}z = \overline{R}\hat{z}, \qquad \epsilon^{1/2}S(r,t) = \overline{R}\hat{S}(\hat{r},\hat{t}),$$

$$\Delta T = T_m - T_g, \qquad \mathrm{St} = \frac{L}{c_s\Delta T}, \qquad T = T_g + \Delta T\Theta, \qquad t = \frac{\mathrm{St}\,\overline{R}^2\rho_s c_s}{k_s\epsilon}\hat{t}.$$

Here variables with hats ($\hat{}$) are the nondimensional ones. In terms of these variables the heat equation in the crystal (1) becomes

$$\frac{\epsilon}{\mathrm{St}}\Theta_t = \frac{1}{r}(r\Theta_r)_r + \epsilon\Theta_{zz}, \qquad \mathbf{x}\in\Omega,\ t>0, \tag{12a}$$

with boundary conditions (2)–(4) becoming

$$-\Theta_r + \epsilon\Theta_z R'(z) = \epsilon\left[1+\epsilon(R'(z))^2\right]^{1/2}\Theta, \qquad \mathbf{x}\in\Gamma_g, \tag{12b}$$

$$\Theta_z(0,t) = \delta(\Theta(0,t) - \Theta_{ch}), \tag{12c}$$

$$\Theta = 1, \qquad \mathbf{x}\in\Gamma_S, \tag{12d}$$

where $\delta = \epsilon^{1/2}h_{ch}/h_{gs}$. The hats have been dropped for brevity. The crystal-melt interface advances according to the Stefan condition (5)–(6) which in nondimensional coordinates becomes

$$\Theta_z - \frac{1}{\epsilon}S_r\Theta_r = (\gamma + S_t), \qquad \gamma = \frac{q_l\overline{R}}{\epsilon^{1/2}k_s\Delta T}, \tag{12e}$$

where q_l and γ are the dimensional and nondimensional heat fluxes in the liquid across the crystal-melt interface in the axial direction. Note that we have chosen the rate of solidification to define the characteristic time scale. The Stefan number St gives the ratio of this characteristic solidification time scale to the time scale associated with conductive heat loss through the crystal side surface. Based on the parameter values in Table 1, we have $\mathrm{St} \simeq 4.3$, suggesting that the conductive scale is small and the temperature inside the crystal is steady on the growth time scale.

2.3. Growth conditions. Under general growth conditions the process may be pseudosteady. However, near the end of the process, transient influences may become important. To investigate both possibilities two situations are considered.

1. The growth of the crystal is characterized by an externally chosen value of q_l (or the nondimensional flux γ), constant for the duration of the simulation.
2. Using the temperature of the crucible Θ_c as a control parameter, q_l is determined implicitly by an effective heat transfer coefficient of the crucible to the crystal through the melt.

In the second scenario a simple model is used to couple the heat fluxes inside the grower based on the fact that the system is almost at thermal equilibrium. The nondimensional liquid temperature satisfies

(13a)
$$\frac{\phi}{\lambda\,\mathrm{St}}\frac{d}{dt}\left[\left(\frac{T_g}{\Delta T}+\Theta_l\right)V_l\right]=-\mu\lambda^2\frac{h_{sl}}{h_{gs}}A(\Theta_l-1)-\mu\frac{h_{gl}}{h_{gs}}(A_c-\lambda^2A)\Theta_l+\frac{h_{cl}}{h_{gs}}A_l(\Theta_c-\Theta_l)$$

with $\Theta_l(0)=1$ and $\Theta_c(0)$ chosen so that $\Theta_l'(0)=0$. The detailed derivation is given in the appendix, and from expression (47c) the γ in (12e) becomes

(13b)
$$\gamma=\epsilon^{1/2}\frac{h_{sl}}{h_{gs}}(\Theta_l-1).$$

3. Perturbation solution. We now seek to approximate the scaled model in section 2.2 via a straightforward perturbation expansion. In turn, this perturbation model will form the basis for a numerical solution. Since St is O(1) under the current growth conditions it is retained as a parameter. Equations (12a) and (12b) strongly suggest that the temperature Θ is independent of r to leading order. If true, then the crystal-melt interface S is also independent of r to leading order, and we see that this is consistent in (12e) with the growth being driven primarily by the vertical gradients. These observations motivate the following approximations:

(14)
$$\begin{aligned}\Theta&\sim\Theta_0(z,t)+\epsilon\Theta_1(r,z,t)+\epsilon^2\Theta_2(r,z,t)+\cdots,\\ S&\sim S_0(t)+\epsilon S_1(r,t)+\epsilon^2S_2(r,t)+\cdots.\end{aligned}$$

We substitute them into the scaled model, expand in powers of ϵ, simplify, and collect terms. The resulting field equations to first order are

(15a)
$$\frac{1}{\mathrm{St}}\Theta_{0,t}-\Theta_{0,zz}=\frac{1}{r}\frac{\partial}{\partial r}(r\Theta_{1,r}),\qquad \mathbf{x}\in\Omega,\ t>0,$$

(15b)
$$\frac{1}{\mathrm{St}}\Theta_{1,t}-\Theta_{1,zz}=\frac{1}{r}\frac{\partial}{\partial r}(r\Theta_{2,r}),\qquad \mathbf{x}\in\Omega,\ t>0,$$

where the boundary condition on the lateral surface becomes

(16a)
$$(\Theta_{1,r}-R'\Theta_{0,z}+\Theta_0)(R(z),z,t)=0,$$

(16b)
$$\left(\Theta_{2,r}-R'\Theta_{1,z}+\frac{1}{2}R'^2\Theta_0+\Theta_1\right)(R(z),z,t)=0.$$

Continuing this procedure for the remaining conditions, at the top of the crystal one has

$$\begin{aligned}\Theta_{0,z}(0,t)&=\delta(\Theta_0(0,t)-\Theta_{ch}),\\ \Theta_{1,z}(r,0,t)&=\delta\Theta_1(r,0,t),\end{aligned}$$

and at the solid-liquid interface

$$\Theta_0(S_0(t),t) = 1, \tag{17a}$$

$$(S_1\Theta_{0,z} + \Theta_1)(r, S_0(t), t) = 0. \tag{17b}$$

Finally, the evolution of the interface is governed by

$$S_0'(t) = \Theta_{0,z}(S_0(t),t) - \gamma, \qquad S_0(0) = Z_0, \tag{18a}$$

$$S_{1,t}(r,t) = \left(\Theta_{1,z} + S_1\Theta_{0,zz} + \frac{\Theta_{1,r}^2}{\Theta_{0,z}}\right)(r, S_0(t), t), \qquad S_1(r,0) = 0, \tag{18b}$$

where we have used (17b) to eliminate the $S_{1,r}$ term. We note that by expanding the solid-liquid interface into the same asymptotic series and deriving the above equations, we have implied that the interface will adjust its shape to avoid a temperature singularity at the solid-liquid-gas triple point. We also note that Z_0 is the nondimensional length of the seed. In addition there will be symmetry conditions at $r = 0$ for Θ_k, S_k, $k = 0, 1$.

3.1. Resolution of the zeroth order model. Integrating (15a) once and imposing the symmetry condition $\Theta_{1,r} = 0$ at $r = 0$, we have

$$\frac{r}{2}\left(\frac{1}{\mathrm{St}}\Theta_{0,t} - \Theta_{0,zz}\right) = \Theta_{1,r}.$$

Applying (16a) at $r = R$ gives the zeroth order problem

$$\frac{1}{\mathrm{St}}\Theta_{0,t} - \Theta_{0,zz} = \frac{2}{R}\left(R'\Theta_{0,z} - \Theta_0\right), \qquad 0 < z < S_0(t),\ t > 0, \tag{19a}$$

$$\Theta_{0,z}(0,t) = \delta(\Theta_0(0,t) - \Theta_{ch}), \qquad t \geq 0, \tag{19b}$$

$$\Theta_0(S_0(t),t) = 1, \qquad t \geq 0, \tag{19c}$$

$$S_0'(t) = \Theta_{0,z}(S_0(t),t) - \gamma, \qquad S_0(0) = Z_0,\ t > 0, \tag{19d}$$

with an initial condition $\Theta_0(z,0) = f(z) \leq 1$ compatible with the boundary conditions. Provided that $R \in C^1([0, S_0])$, the Stefan problem will have a unique solution. For details see Friedman [12].

Equation (19a) is parabolic and involves only the heat fluxes along the length of the crystal. With the chosen expansion we see that at zeroth order the temperature field has no radial dependence. In addition, we can see that the thermal gradients, as discussed previously, are caused by cooling effects at the surface. In section 5.1 we solve the time dependent system (19) on $0 < z < S_0(t)$ for a suitable set of initial conditions. Also notice expression (19d) illustrates that the chosen time scale balances the growth. The appearance of $\mathrm{St} > 1$ in (19a) suggests that thermal transients in the bulk of the crystal are not as important as the growth transient. This is explored further in section 5.1. The limit as $\mathrm{St} \to \infty$ leads naturally to a pseudosteady leading order model in which time dependency enters the thermal model only through the growth; i.e., we also solve as the pseudosteady limit

$$\Theta_{0,zz} + \frac{2}{R}\left(R'\Theta_{0,z} - \Theta_0\right) = 0, \qquad 0 < z < S_0(t),\ t > 0, \tag{20}$$

with (19b)–(19d). Expression (20) together with (19b)–(19d) is analogous to a Hele–Shaw problem, and it is well known that in one dimension solutions of the Stefan

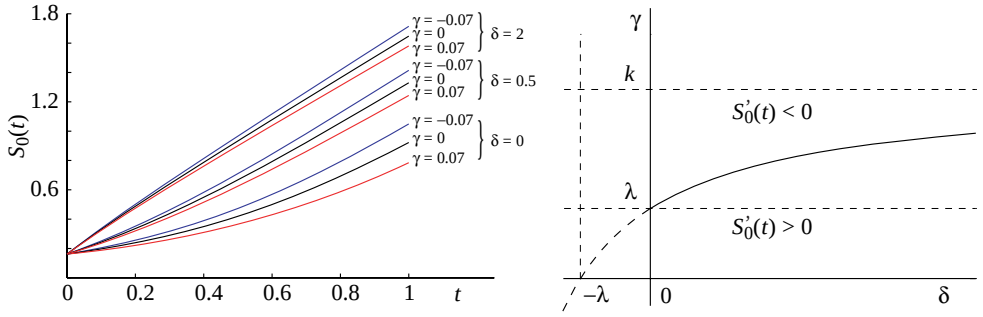


FIG. 3. *To the left is the evolution of the length of a cylindrical crystal grown using the pseudosteady approximation according to* (22). *For the simulations,* $\Theta_{ch} = 0$, $R_0 = 1$, $Z_0 = 0.162$ *(3 cm), and* $\epsilon = 0.026$ *(*$h_{gs} = 4\,W/m^2K$*). The position* $S_0 = 1.8$ *at* $t = 1$ *corresponds to a crystal length of 33 cm grown in 13.6 hours, typical for Cz grown InSb* [24]. *The curve on the right is* $S_0'(t) = 0$, *which characterizes the balance of the heat flux from the melt with the loss of heat at the chuck. The quantity* $\lambda = k \tanh kS_0$ *increases with* S_0.

problem (19) converge to solutions of Hele–Shaw as St $\to \infty$. However, the convergence is not uniform and the intermediate asymptotic behaviors are different [30].

We start by exploring two special cases for which an analytic solution may be computed to the pseudosteady model.

3.1.1. Constant radius crystals. In this case we take $R(z) = R_0$, and (20) becomes simply

$$\Theta_{0,zz} - \frac{2}{R_0}\Theta_0 = 0, \qquad 0 < z < S_0(t),\ t > 0,$$

with boundary conditions (19b) and (19c). Solving for Θ_0 gives

$$\Theta_0(z,t) = \frac{k\cosh kz + \delta \sinh kz + \delta\Theta_{ch}\sinh k(S_0 - z)}{k\cosh kS_0 + \delta \sinh kS_0}, \quad k^2 = \frac{2}{R_0}. \tag{21}$$

The crystal grows at a rate governed by the Stefan condition (19d)

$$S_0'(t) = k\frac{k\sinh kS_0 + \delta\cosh kS_0 - \delta\Theta_{ch}}{k\cosh kS_0 + \delta\sinh kS_0} - \gamma, \quad S_0(0) = Z_0. \tag{22}$$

The left-hand side of Figure 3 shows the time dependence of the position of the interface and therefore the length of the crystal as a function of time for various combinations of δ and γ. The right-hand side details the balance between these two parameters. The initial position of the curve is determined by the length of the seed Z_0. If the γ, δ pair is chosen above the curve $S_0'(t) = 0$ $\big(\gamma > \gamma_{\max}(\delta) = k^2(\lambda_0 + \delta)/(k^2 + \delta\lambda_0), \lambda_0 = k\tanh kZ_0, \Theta_{ch} = 0\big)$, the seed melts back. If we are below the curve, then $S_0(t)$ increases without bound and the curve asymptotically approaches $\gamma = k$. For small S_0, (22) gives $S_0' = k^2S_0 - \gamma + \delta(1-\Theta_{ch})(1-\delta S_0) + \mathcal{O}(S_0^2)$ and for large S_0, the growth rate is asymptotically $S_0' = k - \gamma$.

3.1.2. Conical crystals. One source of ambiguity in the constant radius model above is the need to specify the chuck temperature and heat transfer coefficient. In the case of a conical crystal, which is closer to reality, this ambiguity is less prominent. We assume $R(z) = R_0 + \alpha z$, where $\arctan\alpha \simeq \mathcal{O}(1)$ is one-half the opening angle of

the crystal when using nondimensional units. This assumption is predicated on the condition that the dimensional version of $\alpha \simeq \mathcal{O}(\sqrt{\epsilon})$, and it is easily verified using the data in Table 1. Substituting for $R(z)$ we solve

$$\Theta_{0,\eta\eta} + \frac{2}{\eta}(\Theta_{0,\eta} - \Theta_0) = 0, \qquad R_0 < \alpha^2\eta < R_0 + \alpha S_0,\ t > 0, \tag{23a}$$

$$\Theta_{0,\eta} = \alpha\delta(\Theta_0 - \Theta_{ch}), \qquad \alpha^2\eta = R_0,\ t \geq 0, \tag{23b}$$

$$\Theta_0 = 1, \qquad \alpha^2\eta = R_0 + \alpha S_0,\ t \geq 0, \tag{23c}$$

where $\alpha^2\eta(z) = R_0 + \alpha z$. The resulting solution takes the form of linear combinations of modified Bessel functions

$$\Theta_0(\eta, t) = \frac{W_{fg}(\eta, \eta(0)) - \alpha\delta\left[V_{fg}(\eta, \eta(0)) - \Theta_{ch}V_{fg}(\eta, \eta(S_0))\right]}{W_{fg}(\eta(S_0), \eta(0)) - \alpha\delta V_{fg}(\eta(S_0), \eta(0))}, \tag{24}$$

where

$$W_{fg}(x,y) = \begin{vmatrix} f(x) & g(x) \\ f'(y) & g'(y) \end{vmatrix}, \quad V_{fg}(x,y) = \begin{vmatrix} f(x) & g(x) \\ f(y) & g(y) \end{vmatrix},$$
$$f(\eta) = \frac{I_1(\sqrt{8\eta})}{\sqrt{\eta}}, \qquad g(\eta) = \frac{K_1(\sqrt{8\eta})}{\sqrt{\eta}}.$$

The corresponding expression for the growth rate is

$$S_0'(t) = \frac{1}{\alpha}\frac{V_{f'g'}(\eta(S_0), \eta(0)) + \alpha\delta\left[W_{fg}(\eta, \eta(0)) - \Theta_{ch}W_{fg}(\eta, \eta(S_0))\right]}{W_{fg}(\eta(0), \eta(S_0)) - \alpha\delta W_{fg}(\eta(S_0), \eta(S_0))} - \gamma, \quad S_0(0) = Z_0.$$

Two limiting cases are considered. To compare with the cylindrical case one sets $R_0 = 1$ and expands (24) in a power series of α yielding

$$\Theta_0(z,t) = \frac{\cosh\sqrt{2}z}{\cosh\sqrt{2}S_0}\left\{1 - \frac{\alpha}{8}\left[6(z - S_0) + \sqrt{8}(z^2 - U)\tanh\sqrt{2}z \right.\right.$$
$$\left.\left. - \sqrt{8}(S_0^2 - U)\tanh\sqrt{2}S_0\right]\right\} + \mathcal{O}(\alpha^2) \tag{25a}$$

with

$$U = \frac{3}{2} + 2\delta\left(1 - \Theta_{ch}\cosh\sqrt{2}S_0\right). \tag{25b}$$

Expressions (25) should be compared to (21). Since for a cone $R_0 \ll 1$, a simple form of (24) can be obtained by expanding the solution in R_0 as

$$\Theta_0(z,t) = \sqrt{\frac{S_0}{z}}\frac{I_1(\sqrt{8z/\alpha})}{I_1(\sqrt{8S_0/\alpha})}\left[1 + \frac{R_0}{\alpha}\left(\sqrt{\frac{2}{\alpha z}}\frac{I_0(\sqrt{8z/\alpha})}{I_1(\sqrt{8z/\alpha})}\right.\right.$$
$$\left.\left. - \sqrt{\frac{2}{\alpha S_0}}\frac{I_0(\sqrt{8S_0/\alpha})}{I_1(\sqrt{8S_0/\alpha})} - \frac{1}{z} + \frac{1}{S_0}\right)\right] + \mathcal{O}(R_0^2). \tag{26}$$

The solution for conical crystals is more complicated, and we will defer the discussion to section 5.

3.1.3. Comments. This model for a one-dimensional temperature variation in the axial direction is not new. For example, it has been used in [32] but without formal justifications. What we have done here, by deriving it using asymptotic expansion, is to allow the reader to realize the applicability and restrictions of the model.

3.2. Radial variations: Resolution of the first order model. Having solved the zeroth order model, to give Θ_0 and S_0, we can resolve the radial variations in temperature, which occur at first order in Θ_1, and also consider the shape of the crystal-melt interface as it evolves, through S_1. From resolution of the zeroth order model we have

$$\Theta_{1,r} = \frac{r}{2}\left(\frac{1}{\mathrm{St}}\Theta_{0,t} - \Theta_{0,zz}\right),$$

and integrating with respect to r we have

$$\Theta_1(r,z,t) = \Theta_1(0,z,t) + \frac{r^2}{4}\left(\frac{1}{\mathrm{St}}\Theta_{0,t} - \Theta_{0,zz}\right),$$

or

$$\Theta_1(r,z,t) = \Theta_1^0(z,t) + r^2\Theta_1^1(z,t), \tag{27}$$

where $\Theta_1^0(z,t) = \Theta_1(0,z,t)$ and using (19a)

$$\Theta_1^1(z,t) = \frac{1}{2R}(R'\Theta_{0,z} - \Theta_0). \tag{28}$$

The function $\Theta_1^1(z,t)$ is known from the data and the zeroth order solution. By adopting the same procedure as for the zeroth order model we can find $\Theta_1^0(z,t)$, i.e., integrating (15b) with respect to r and using the boundary condition at $r = R$ to eliminate $\Theta_{2,r}$. We derive

$$\frac{1}{\mathrm{St}}\Theta_{1,t}^0 - \Theta_{1,zz}^0 = \frac{2}{R}(R'\Theta_{1,z}^0 - \Theta_1^0) + F_1, \qquad 0 < z < S_0(t),\ t > 0, \tag{29a}$$

$$\Theta_{1,z}^0(0,t) = \delta\Theta_1^0(0,t), \qquad t \geq 0, \tag{29b}$$

$$\Theta_1^0(S_0(t),t) = -S_1(0,t)\Theta_{0,z}(S_0(t),t), \qquad t \geq 0, \tag{29c}$$

$$S_1'(r,t) = \left(\Theta_{1,z}^0 + S_1\Theta_{0,zz} + r^2F_2\right)(S_0(t),t), \qquad S_1(r,0) = 0,\ t > 0, \tag{29d}$$

where

(29e)

$$F_1 = -\frac{R^2}{2}\left(\frac{1}{\mathrm{St}}\Theta_{1,t}^1 - \Theta_{1,zz}^1\right) + 2R(R'\Theta_{1,z}^1 - \Theta_1^1) - \frac{R'^2\Theta_0}{R}, \quad F_2 = \Theta_{1,z}^1 + \frac{4(\Theta_1^1)^2}{\Theta_{0,z}},$$

and r appears as a parameter in (29d).

This first order problem (29) has the same structure as the zeroth order problem (19) but is inhomogeneous; i.e., the zeroth order solution provides the forcing (or heating). A further key difference is in the coupling with the crystal-melt interface position S_1. Equation (29c) provides the lower boundary condition for Θ_1 and S_1 advances through (29d), which is consequently a first order quasi-linear partial differential equation.

In general, the coupled system (29) must be solved numerically. For the pseudosteady case, the formula can be simplified as follows. From the definition of Θ_1^1 and using the pseudosteady condition $\Theta_{0,zz} = -4\Theta_1^1$, expressions (29a) and (29d) reduce to

(30)
$$\Theta^0_{1,zz} + \frac{2}{R}(R'\Theta^0_{1,z} - \Theta^0_1) = \frac{1}{4}(2R' + 5R'R'' - RR''')\Theta_{0,z} - \frac{1}{4R}(2R - 4R'^2 + 5RR'')\Theta_0$$

and

(31)
$$S_1'(r,t) = \Theta^0_{1,z} + \frac{2}{R}(1 - R'\Theta_{0,z})S_1 + \frac{r^2}{2R^2}\left(-3R' + (RR'' + R'^2 + R)\Theta_{0,z} + \frac{2}{\Theta_{0,z}}\right),$$

where the right-hand side of (31) is evaluated at $z = S_0(t)$ and r appears as a parameter.

3.2.1. Constant radius crystals. Since $R(z) = R_0$, expression (30) reduces to $\Theta^0_{1,zz} - k^2\Theta^0_1 = -\Theta_0/2$ with $k^2 = 2/R_0$. Solving for Θ^0_1 and using (27)–(28) one finds

$$\Theta_1(r,z,t) = \frac{\cosh kz}{\cosh kS_0}\left[-\Gamma(S_0) - S_1^0\Theta_{0,z}(S_0)\right] + \Gamma(z) - \frac{1}{4}k^2r^2\Theta_0(z)$$

with $S_1^0 = S_1(0,z)$ and

$$\Gamma(z) = \frac{1}{2k}\int_0^z \Theta_0(\xi)\sinh k(\xi - z)\,d\xi.$$

When $\delta = 0$, $\Gamma(z) = -z\sinh kz/4k\cosh kS_0$ yields

$$\Theta_1(r,z,t) = \frac{1}{4k}\frac{\cosh kz}{\cosh kS_0}\left(S_0\tanh kS_0 - z\tanh kz - 4k^2S_1^0\tanh kS_0 - k^3r^2\right).$$

S_1 can be obtained by (29d).

3.2.2. Conical crystals. Since $R(z) = R_0 + \alpha z$, and from (20) $\Theta_1^1 = (\alpha\Theta_{0,z} - \Theta_0)/2R = -\Theta_{0,zz}$, we find

$$\Theta^1_{1,z} = -\left(\frac{3\alpha^2}{2R^2} + \frac{1}{2R}\right)\Theta_{0,z} + \frac{3\alpha}{2R^2}\Theta_0, \quad \Theta^1_{1,zz} = \left(\frac{6\alpha^3}{R^3} + \frac{3\alpha}{R^2}\right)\Theta_{0,z} - \left(\frac{6\alpha^2}{R^3} + \frac{1}{R^2}\right)\Theta_0.$$

Consequently, (30) reduces to

$$\Theta^0_{1,zz} + \frac{2\alpha}{R}\Theta^0_{1,z} - \frac{2}{R}\Theta^0_1 = \frac{\alpha}{2}\Theta_{0,z} + \frac{1}{2R}(2\alpha^2 - R)\Theta_0, \tag{32}$$

and we see that even for the pseudosteady cone, numerical methods will have to be used in general.

Further discussion is deferred to section 5. In the following we turn our discussion to thermal stress inside the crystal.

4. Thermal stress. The thermal stress experienced by the crystal during its growth leads to the generation of structural defects in the crystal [31]. If we want to eliminate these undesirable defects, then one must control the thermal stress. We begin with a brief introduction to the case of an isotropic body. Although InSb is

anisotropic with respect to its elasticity, this will be dealt with in a subsequent section. Fundamentals can be found elsewhere [21, 28].

From the elements of the stress tensor the characteristic amount of stress at a particular position can be described by the von Mises stress σ_{VM} with the relationship

$$2\sigma_{\mathrm{VM}}^2 = (\sigma_1 - \sigma_2)^2 + (\sigma_1 - \sigma_3)^2 + (\sigma_2 - \sigma_3)^2, \tag{33}$$

where σ_1, σ_2, σ_3 are the eigenvalues of the stress tensor. Being a function of the eigenvalues, the von Mises stress is invariant under coordinate transformations.

For a given temperature field, the resulting set of thermoelastic equations for the displacement vector are coupled, and a numerical method will be needed to solve the displacements before thermal stress can be computed. It is instructive to consider the special case where the displacement occurs in one of the three directions, due to the nature of temperature variation. In the following, we will address the thermal stress that arises from temperature variation in the radial direction, as this is the dominant contribution.

4.1. Thermal stress due to radial temperature variation. We assume the displacement vector is of the form $\vec{u} = \langle u(r), 0, 0\rangle$ and converting to nondimensional units u satisfies

$$\frac{\partial}{\partial r}\left[\frac{1}{r}\frac{\partial}{\partial r}(ru)\right] = \epsilon\frac{(1+\nu)}{(1-\nu)}\frac{\partial\Theta_1}{\partial r}, \qquad u(0) < \infty, \quad \sigma_{rr}(R) = 0,$$

where ν is the Poisson ratio. The condition $u(0) < \infty$ is due to the axisymmetry, and since the crystal surface is unstressed, $\sigma_{rr}(R) = 0$. The stress has been nondimensionalized by $\alpha_0\Delta TE/(1-\nu)$, E is the Young's modulus, and α_0 is the coefficient of thermal expansion. We have assumed here that $\Theta_z \simeq 0$ since we want to focus on the sole effect of any radial temperature variations. The solution satisfying the boundary conditions is

$$u(r) = \epsilon\frac{(1+\nu)}{(1-\nu)}\left[\frac{1}{r}\int_0^r \Theta_1(s)s\,ds + (1-2\nu)\frac{r}{R^2}\int_0^R \Theta_1(s)s\,ds\right],$$

and using (27), the corresponding nontrivial stresses are

$$\sigma_{rr} = \epsilon\left[\frac{1}{R^2}\int_0^R \Theta_1(s)s\,ds - \frac{1}{r^2}\int_0^r \Theta_1(s)s\,ds\right] = \frac{1}{4}\epsilon\Theta_1^1\left(R^2 - r^2\right), \tag{34a}$$

$$\sigma_{\theta\theta} = \epsilon\left[\frac{1}{R^2}\int_0^R \Theta_1(s)s\,ds + \frac{1}{r^2}\int_0^r \Theta_1(s)s\,ds - \Theta_1(r)\right] = \frac{1}{4}\epsilon\Theta_1^1\left(R^2 - 3r^2\right), \tag{34b}$$

$$\sigma_{zz} = \epsilon\left[\frac{2}{R^2}\int_0^R \Theta_1(s)s\,ds - \Theta_1(r)\right] = \frac{1}{2}\epsilon\Theta_1^1\left(R^2 - 2r^2\right) \tag{34c}$$

with σ_{zz} modified using St. Venant's principle.

Using (33) to compute the von Mises stress gives

$$\sigma_{\mathrm{VM}}(r,z,t) = \frac{1}{4}\epsilon\left|\Theta_1^1\right| R^2\left[1 - 4\left(\frac{r}{R}\right)^2 + 7\left(\frac{r}{R}\right)^4\right]^{1/2}. \tag{35}$$

The object in the square brackets is a shape factor, and it ranges from a value of

$\sqrt{3/7}$ at a radius of $r = \sqrt{2/7}R(z)$ to a maximum value of two at the outer edge of the crystal. For $\sqrt{4/7}R(z) < r \leq R(z)$ this factor is greater than one.

Remark 1. From (19a) and (28), $\Theta_1^1 = (\Theta_{0,t}/\,\mathrm{St} - \Theta_{0,zz})/4$, which reduces to $|\Theta_1^1| = |\Theta_{0,zz}|/4$ in the pseudosteady limit. This generalizes the classical result that the stress level is a characteristic of the concavity of the temperature in the axial direction [22, 34]. The stress is also linearly proportional to the Biot number ϵ indicating that an increase in the crystal radius will also increase the stress level, other conditions being equal. It also indicates that the increase of radius can be offset by reducing the heat transfer coefficient h_{gs}, suggesting that a possible way to reduce the stress is by changing the local heat flux from the crystal lateral surface.

Remark 2. Also from (28), it is clear that there are two components in the expression for Θ_1^1. Therefore both the temperature gradient and the crystal profile are important for stress reduction. Since circular cylindrical shape is normally adopted for the growth of more common crystals such as silicon, the shape effect has not been discussed much in the literature. However, for crystals such as InSb with relatively low resistance to thermal stress, finding the right shape is often part of an important strategy for growing defect-free crystals. We will address this issue again in later sections.

Remark 3. As other stresses, such as the total resolved stress, are often considered more relevant for causing defects, it is important to point out that the same characteristics remain for different representations of the thermal stress or for crystals being pulled in different directions. These issues will be the topic of the following two subsections.

4.2. Resolved stress. InSb crystallizes in a zincblende or $\bar{4}$3m structure. The structure description is two interpenetrating face-centered cubic (f.c.c.) sublattices of In and Sb separated by the displacement vector $a\langle 1,1,1\rangle/4$. Each In (Sb) atom is tetrahedrally coordinated with an Sb (In) atom. An alternative description of the structure is a f.c.c. sublattice of Sb atoms with one-half of the tetrahedral sites filled with In atoms. The nearest neighbor distance is $\sqrt{3}a/4$ and the lattice parameter is $a = 0.6476$ nm.

The preferred method of dislocation generation in InSb, as in all III-V semiconductors, is through the generation of slip defects, in particular the $\{111\}$, $\langle 1\bar{1}0\rangle$ slip system [19]. This system consists of four glide planes within which atoms can slip in one of three directions. For example, in the (111) plane the slip directions are $[10\bar{1}]$, $[\bar{1}10]$, and $[0\bar{1}1]$. Figure 4 looks down the z-axis of the tetrahedral structure of the crystal and shows each of the 12 permissible glide directions classified into five different categories.

The amount of stress in a particular slip direction $\vec{g}$ within a given glide plane with normal $\vec{n}$ is known as the resolved stress, σ_{RS}. If one assumes the crystallographic axes coincide with the coordinate axes, then σ_{RS} is computed by finding

$$\sigma_{\mathrm{RS}} = \vec{g}^T Q \sigma Q^T \vec{n}, \tag{36}$$

where Q is the coordinate transformation matrix that takes $(r,\theta,z) \to (x,y,z)$ and σ is the stress tensor in the (r,θ,z) coordinates. In summary, the five categories

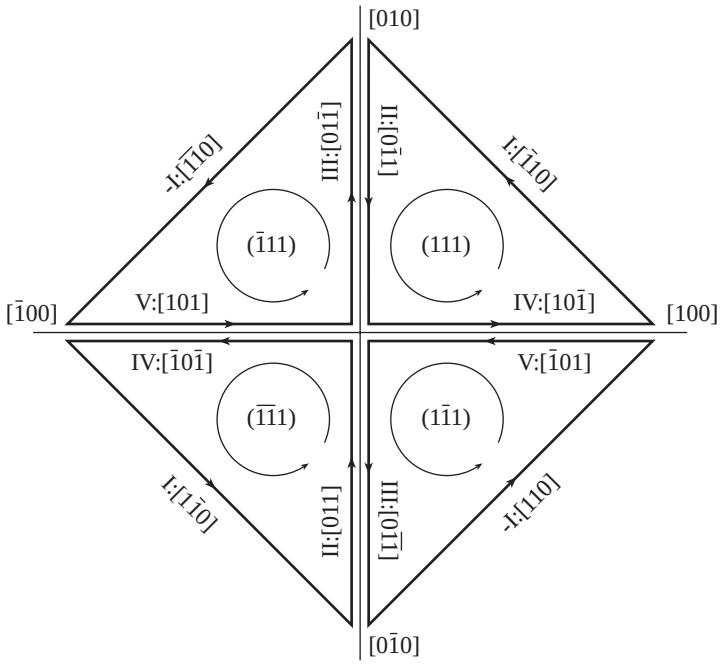


FIG. 4. *Illustrated are each of the* 12 *slip directions in the* $\{111\}$, $\langle 1\bar{1}0\rangle$ *slip system. The roman numerals refer to the functional form of the stress in the direction of the appropriate slip plane.*

illustrated in Figure 4 yield[1]

$$\sigma_{\mathrm{RS}}^{\mathrm{I}} = -\frac{1}{\sqrt{6}}(\sigma_{rr} - \sigma_{\theta\theta})\cos 2\theta, \tag{37a}$$

$$\sigma_{\mathrm{RS}}^{\mathrm{II}} = \frac{1}{\sqrt{6}}[(\sigma_{zz} - \sigma_{\theta\theta}) - (\sigma_{rr} - \sigma_{\theta\theta})(\sin^2\theta + \sin\theta\cos\theta)], \tag{37b}$$

$$\sigma_{\mathrm{RS}}^{\mathrm{III}} = -\frac{1}{\sqrt{6}}[(\sigma_{zz} - \sigma_{\theta\theta}) - (\sigma_{rr} - \sigma_{\theta\theta})(\sin^2\theta - \sin\theta\cos\theta)], \tag{37c}$$

$$\sigma_{\mathrm{RS}}^{\mathrm{IV}} = -\frac{1}{\sqrt{6}}[(\sigma_{zz} - \sigma_{\theta\theta}) - (\sigma_{rr} - \sigma_{\theta\theta})(\cos^2\theta + \sin\theta\cos\theta)], \tag{37d}$$

$$\sigma_{\mathrm{RS}}^{\mathrm{V}} = \frac{1}{\sqrt{6}}[(\sigma_{zz} - \sigma_{\theta\theta}) - (\sigma_{rr} - \sigma_{\theta\theta})(\cos^2\theta - \sin\theta\cos\theta)]. \tag{37e}$$

Plastic deformation of the crystal occurs if the stress in any of the 12 slip directions exceeds the critical resolved shear stress, σ_{crss}. To leading order, the actual density of dislocations suffered by the crystal is proportional to the total excess stress at any given point within the crystal. In this sense, an estimation of where dislocations are likely to occur is given by the distribution of the total absolute stress:

$$|\sigma_{\mathrm{tot}}| = 4\left|\sigma_{\mathrm{RS}}^{\mathrm{I}}\right| + 2\left(\left|\sigma_{\mathrm{RS}}^{\mathrm{II}}\right| + \left|\sigma_{\mathrm{RS}}^{\mathrm{III}}\right| + \left|\sigma_{\mathrm{RS}}^{\mathrm{IV}}\right| + \left|\sigma_{\mathrm{RS}}^{\mathrm{V}}\right|\right). \tag{38}$$

An additional complication is that, in general, the elastic constants depend on the solidification direction since the thermal and crystallographic axes are not aligned. However, for crystals that belong to the cubic classes this does not play a role [2].

[1]Note: $\theta = -\varphi$, where φ is the angular coordinate used by Jordan, Caruso, and von Neida [19].

4.3. Crystal extraction in an arbitrary direction. The previous subsection supposes the crystal is extracted from the melt in a direction coincident with the crystallographic axes. If there is a misalignment between the frame defined by the crystallographic axes of the unit cell and the frame with its z-axis coincident with the solidification direction, then a coordinate transformation is required to align $\vec{n}$ and $\vec{g}$. These directions must change because they are with respect to the crystallographic axes and not the the temperature field that determines the stress tensor. Let U_{v_p} denote a coordinate transformation, depending on the pulling direction, that takes vectors in the crystallographic frame to the solidification frame. The total resolved stress of the $\{111\}$, $\langle 1\bar{1}0\rangle$ slip system becomes

$$|\sigma_{\text{tot}}| = \sum_{i=1}^{12} \left| \vec{g}_i^T U_{v_p}^T Q \sigma Q^T U_{v_p} \vec{n}_i \right|, \tag{39}$$

where the symmetry in expression (38) has been broken.

5. Numerical results and discussion. We first discuss the temperature solutions for the decoupled growth; i.e., the heat flux from the melt to the crystal is assumed a known (constant) value. In particular, we compare the pseudosteady $\big((19), (29)$ with $\Theta_{0,t} = \Theta_{1,t}^0 = \Theta_{1,t}^1 = 0\big)$ and the unsteady solution (12). Even though the Stefan number is not much bigger than unity, the results show that the pseudosteady solutions are in good agreement with the unsteady calculation. This indicates that the thermal stress can be reasonably estimated using the pseudosteady solution, which greatly simplifies the calculation. The case for coupled growth is also investigated, and we show the transient influence of the melt is only important towards the end of the growth. During the growth, the heat flux from the melt to the crystal changes slowly, suggesting that the temperature solutions for the decoupled case are good approximations. The thermal stress is computed based on the pseudosteady solution using the decoupled growth condition for simplicity.

Table 2 displays the various quantities used in the simulation and not found in the previous table.[2]

5.1. Temperature solutions.

5.1.1. Decoupled growth. In this section we attempt to justify the pseudosteady approximation. To begin, we assume $\gamma = \delta = 0$, decoupling the crystal from the melt in the crucible. Figure 5 compares the time dependence of the position of the crystal-melt interface $S(r,t)$ using (12) (the unsteady case) with its zeroth order approximation $S_0(t)$ using (20), (19b)–(19d) (the pseudosteady case). To determine the influence of both the crystal profile and the amount of heat transfer, a cylindrical and conical profile were assumed and, for each profile, two values of ϵ ($\epsilon_1 = 0.0066$, $\epsilon_2 = 0.026$) were considered. Using only the zeroth order approximation, the interface position is uniformly overestimated with the pseudosteady approximation, and the amount of overestimate is proportional to ϵ. The largest relative difference is about 10% and occurs at the end of the growth for a cylindrical crystal with the largest value of ϵ.

The radial dependence of the interface can be estimated with the first order perturbation, $S_0(t) + \epsilon S_1(r,t)$. Figure 6 compares this approximation with the radial de-

[2] h_{sl} and h_{cl} are based on estimated boundary layer thicknesses of 2.5 mm and 1.8 mm, respectively. The former is due to an Ekman layer (rotations of crucible and the crystal at 5 rpm) and the latter is due to natural convection ($\Delta T = 1$ K, $\nu = 3.3 \times 10^{-7}$ m^2/s, Grashof number $\simeq 6.0 \times 10^6$).

TABLE 2
Remaining liquid and growth parameters used in the simulations.

Data	Symbol	Value
Growing properties		
Ambient temperature	T_a	600 K
Seed radius	R_0	0.005 m
Seed length	Z_0	0.03 m
Crucible depth	$\tilde{Z}_c$	0.0875 m
Crucible radius	$\tilde{R}_c$	0.0875 m
Thermal expansion	α_0	5.5×10^{-6} /K
Liquid properties at $T = T_m$		
Density	ρ_l	6.47×10^3 kg/m^3
Thermal conductivity	k_l	9.23 W/m K
Heat capacity	$\rho_l c_l$	1.7×10^6 J/m^3 K
Heat transfer coefficients		
Solid-liquid	h_{sl}	3700 W/m^2 K
Gas-liquid	h_{gl}	2 W/m^2 K
Crucible-liquid	h_{cl}	5230 W/m^2 K

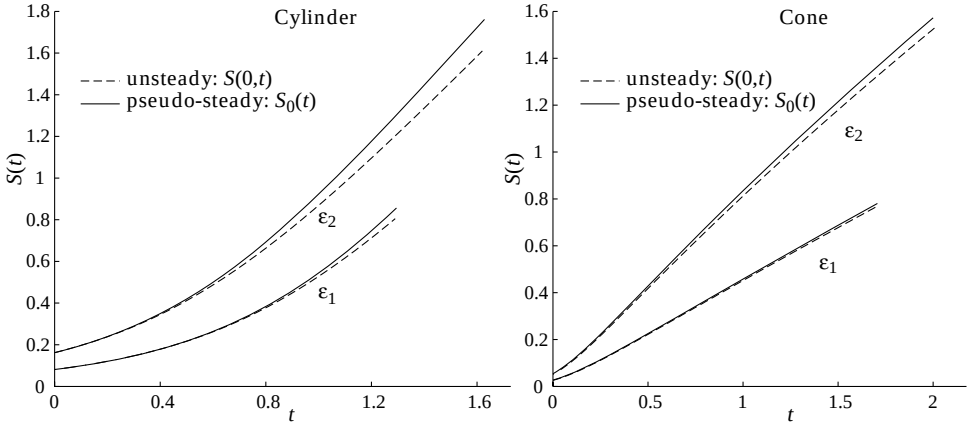


FIG. 5. *Time evolution of $S_0(t)$ for the pseudosteady approximation compared with $S(0,t)$. For both values of ϵ and both crystal profiles $S_0(t)$ closely approximates $S(0,t)$.*

pendence obtained by solving the unsteady equations at the end of the crystal growth, $t = t_f$. In both the case of a cylindrical and a conical crystal the growth interface is convex (viewed from inside the crystal). For the conical crystal the interface is flatter even though the curvature grows with time for both cases. The pseudosteady results closely track the unsteady solution with a maximum relative difference less than 5%.

As a final comparison, Figures 7 and 8 display the predicted thermal profile for the cylindrical and conical crystals, respectively. The crystals are displayed in the physically correct aspect ratio and with their respective solid-liquid interface. It can be seen that the pseudosteady and unsteady solutions are in close agreement.

From the previous results it is clear that without any coupling from the melt, the pseudosteady solution approximates the solution of the fully time dependent Stefan problem. When one considers coupling the melt in the crucible with the crystal, the pseudosteady approximation will remain valid if the predicted heat flux from the melt does not change appreciably over the growth of the crystal. In this case the temperature and flux of the melt are determined by (13).

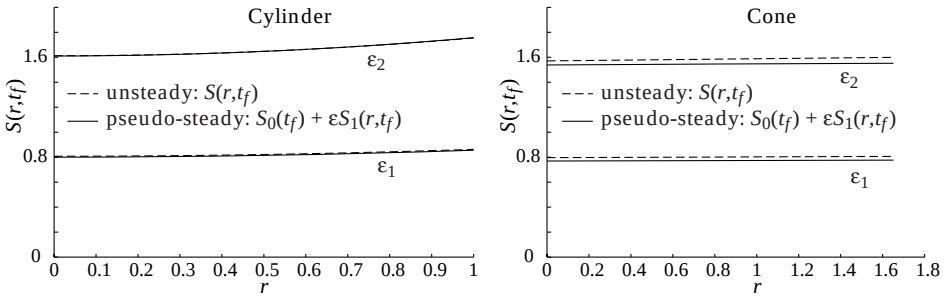


FIG. 6. *Comparison of the radial dependence of interface at the end of growth for the unsteady* $\left(S(r,t_f)\right)$ *and the first order approximation pseudosteady* $\left(S_0(r,t_f)+\epsilon S_1(r,t_f)\right)$ *cases.*

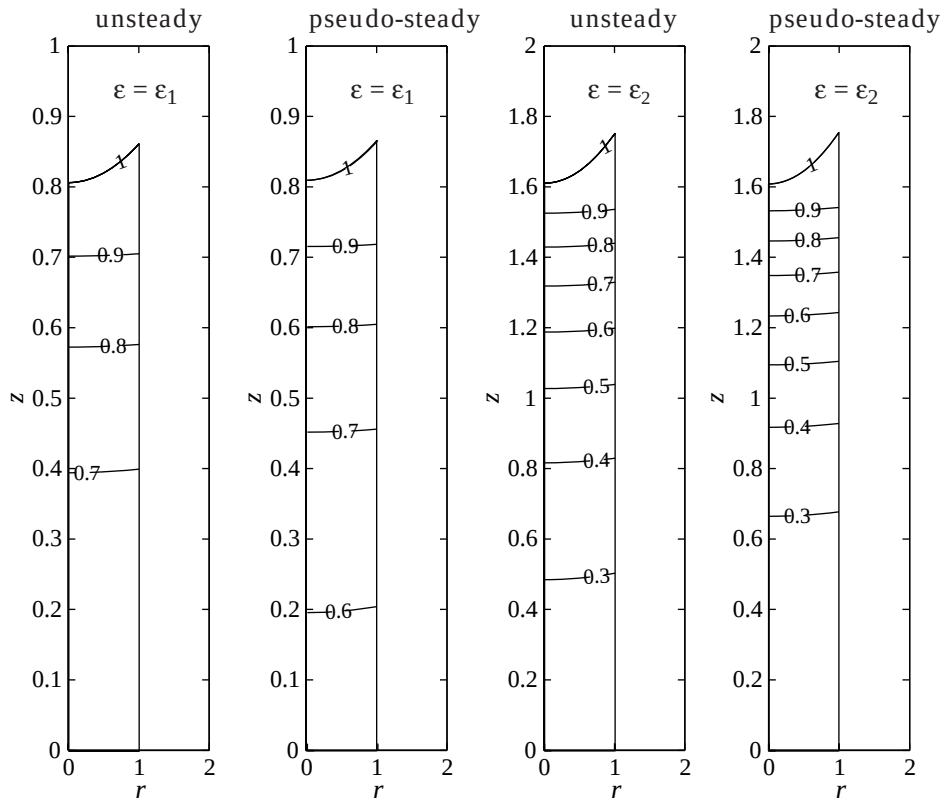


FIG. 7. *Nondimensional temperature contours for the cylinder at the end of the growth illustrated at the correct aspect ratio.*

5.1.2. Coupled growth with melt in the crucible. In this simulation we solve (19) with a cylindrical seed where $\gamma = \gamma(t)$ as determined by (13). In addition, $h_{gs} = 4$, $\delta = 0$, and the remaining heat transfer coefficients are listed in Table 2. Results for a cylindrical and conical crystal pulled from a parabolic crucible $\left(z = \tilde{Z}_c(r/\tilde{R}_c)^2\right)$ are displayed in Figure 9. In both cases the temperature of the crucible was initially $T_c(0) = T_m + 0.046$ K (dimensional) and reduced at a constant rate of

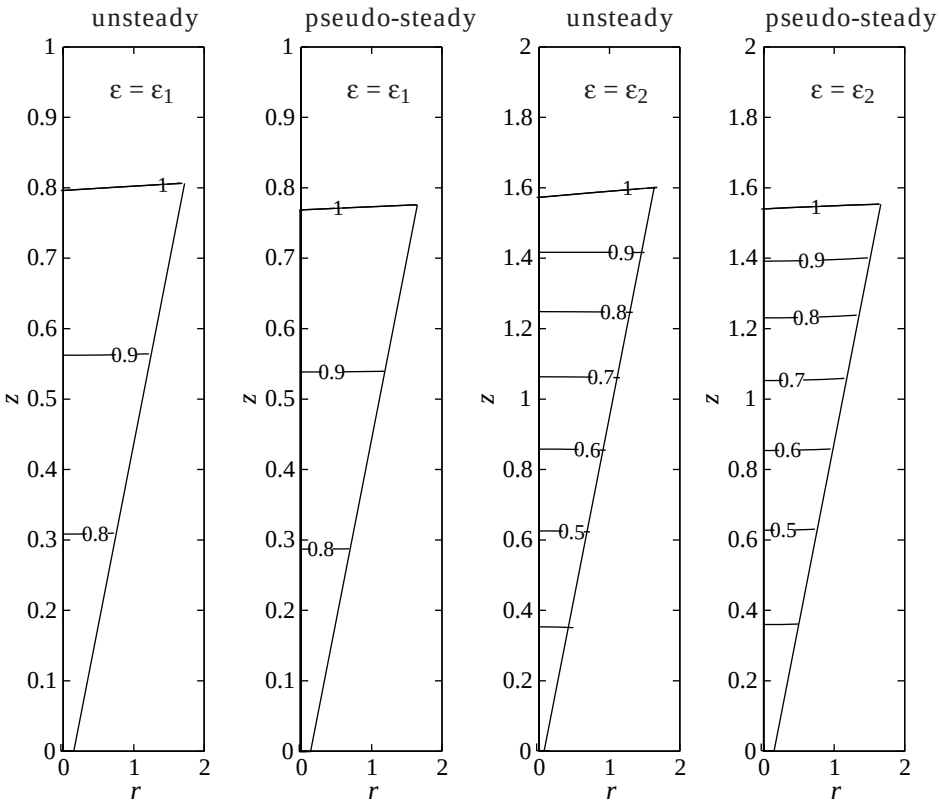


FIG. 8. *Nondimensional temperature contours for the cone at the end of the growth illustrated at the correct aspect ratio.*

0.2 K/hr.

On the far left of Figure 9 the grown crystal is illustrated at the physically correct aspect ratio with the initial and final melt levels. The initial level is chosen so that 25% of the melt remains in the crucible once the crystal achieves its final length. Growth times for each crystal are indicated on the far right.

The center two images display the time evolution of the crystal temperature with the pseudosteady position (St $\to \infty$) of the interface indicated with crosses. For the cylindrical case the initial rapid increase in the radius imprints an echo of the seed into the thermal field which will increase the stress in the shoulder of the crystal. Such an effect has recently been described elsewhere [22]. The dashed lines show the solution if $\gamma = 0$. By setting γ constant the growth rate of the crystal becomes essentially constant rather than accelerating as seen in the coupled case.

The far right shows the growth rate of the interface $S_0'(t)$, the corresponding pull rate $v_p(t)$, and the rate at which the melt drops $\lambda^2 v_m(t)$. For the cylindrical crystal, the growth rate rapidly decreases as the crystal shoulder is formed from the initial seed. This effect is decreased for the conical crystal, leading to a more uniform pull rate with this profile. As the crucible empties, the melt falls more rapidly, causing the pulling speed to reduce near the end of the growth. For the cone this is emphasized as the cross-sectional area increases with time.

Figure 10 illustrates both $\gamma(t)$ and the components of expression (13) during the

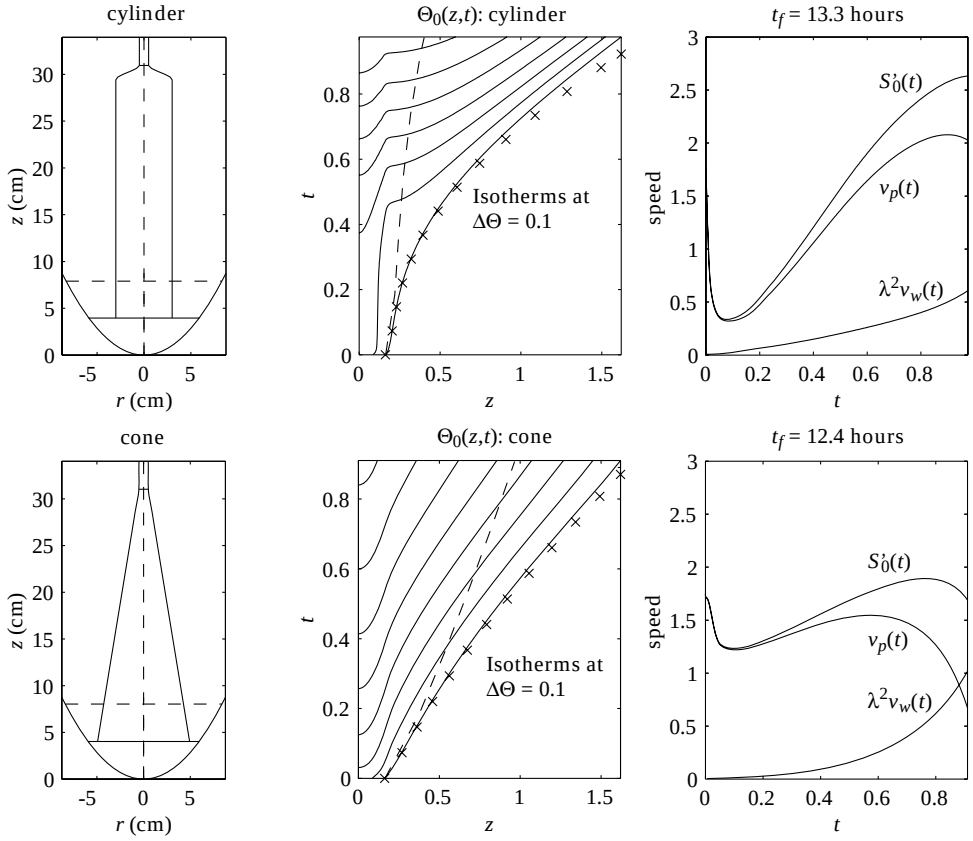


FIG. 9. *Final configuration of the crystal-melt system, thermal fields, and interface speeds for the coupled system* (13) *and* (19). *On the left is the final crystal position in the furnace. The temperature field is detailed in the center and the interfacial velocities are displayed to the right. Cylindrical and conical growth are on the top and bottom rows, respectively.*

growth cycle for the cylindrical and conical crystal profiles. The flux from the melt $\gamma(t)$ is shown to the left of the figure. Changes in the flux are driven primarily by the chosen rate of change of the crucible temperature and is relatively insensitive to the crystal profile. In both cases the magnitude of undercooling was $\simeq 2$ K. A more sophisticated model that considers the dynamics of the melt is clearly required to quantitatively estimate any undercooling effects. This research is currently underway.

Even with our very simple model for the melt, two separate growth regimes are clearly identified. By considering the source of the various heat fluxes acting on the melt one observes that in the initial states of growth, the heat loss to the ambient gas $Q_{gl} = h_{gl}(A_c - A)(T_l - T_g)$ and the heat gain from the crucible $Q_{cl} = h_{cl}A_l(T_c - T_l)$ are the dominant terms in (13). Once the crucible is cool enough, it becomes the dominant channel for heat loss, whereas the melt is heated by the solid-liquid interface at T_m and the decreased heat capacity through volume loss of the melt. The profile of the crystal changes $Q_{sl} = h_{sl}A(T_l - T_m)$ and Q_{gl}, while the shape of the crucible governs the behavior of Q_{cl}. $Q_{vl} = -\rho_l c_l T_l V_l'$ is dictated by the rate of growth. The nondimensional quantities in Figure 10 are found by dividing the Q_i by the dimensional factor $\tilde{R}_c{}^2 \tilde{Z}_c h_{gs} \Delta T / \overline{R}$.

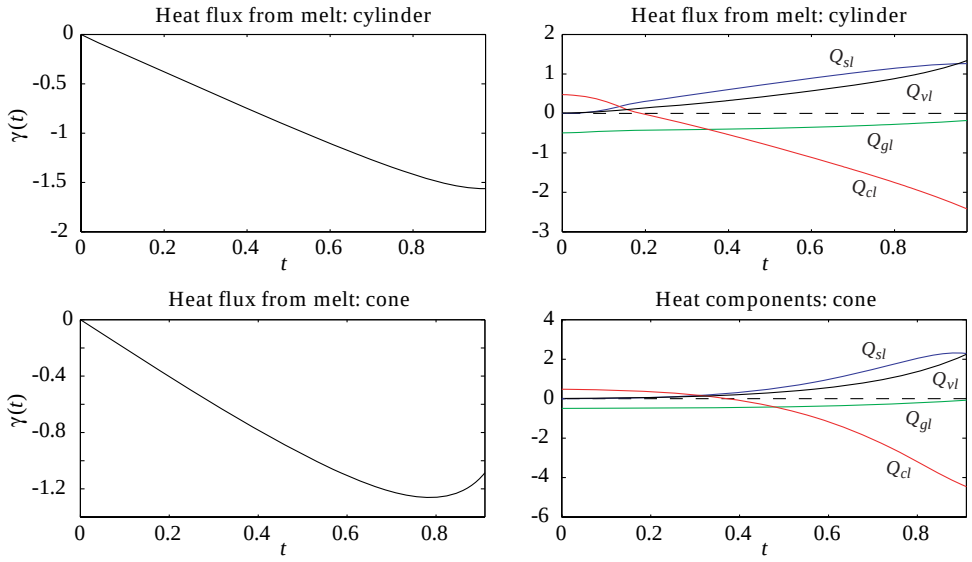


FIG. 10. *The nondimensional flux and the heat flux components of expression* (13) *through the growth cycle. The various components detailed to the right are discussed in section* 5.1.2.

5.2. Thermal stress. For an anisotropic crystal, Young's modulus E and Poisson's ratio ν depend on the specific orientation of the crystal. However, these values are invariant within the {111} planes [2]. For InSb one has

$$E_{\{111\}} = \frac{4(C_{11}+2C_{12})(C_{11}-C_{12})C_{44}}{(C_{11}+2C_{12})(C_{11}-C_{12})+2C_{11}C_{44}} = 6.18\times 10^4 \text{ MPa},$$
$$\nu_{\{111\}} = \frac{1}{3}\frac{(C_{11}+2C_{12})(C_{11}-C_{12})-2C_{44}(C_{11}-4C_{12})}{(C_{11}+2C_{12})(C_{11}-C_{12})+2C_{11}C_{44}} = 0.364,$$

where $C_{11} = 6.70\times 10^4$, $C_{12} = 3.65\times 10^4$, $C_{44} = 3.02\times 10^4$ are crystal stiffness constants[3] in MPa and, consequently, the dimensional constant for the stress calculations is $\alpha_0 \Delta T E/(1-\nu) \simeq 107$ MPa.

Figure 11 shows the stress contours of the von Mises stress for the cylinder and the cone at the end of the growth corresponding to the pseudosteady results in Figures 7 and 8. For a fixed value of ϵ the stress in the conical case is about one-half that of the cylindrical case. Also, increasing ϵ increases the stress level dramatically. By growing a conical crystal the stress can be reduced significantly. For a given temperature the amount of stress at which crystal deformation begins to occur is known as the critical resolved shear stress, σ_{crss}. In the case of InSb, σ_{crss} varies from 0.245 MPa [25] to 4.90 MPa [6] as the temperature varies from $T_m = 798.4$ K to 491 K, respectively, indicating that the conical crystal remains below this critical stress level.

An additional method of reducing the stress level in the crystal is to use the anisotropic nature of the crystal to our advantage by changing the direction in which the crystal is solidified. From expression (39) one can see that for a fixed vertical position in the crystal the total absolute resolved stress is a complicated function of

[3]The corresponding compliances are $S_{11} = 2.42\times 10^{-5}$ MPa^{-1}, $S_{12} = -8.55\times 10^{-6}$ MPa^{-1}, $S_{44} = 3.31\times 10^{-5}$ MPa^{-1}.

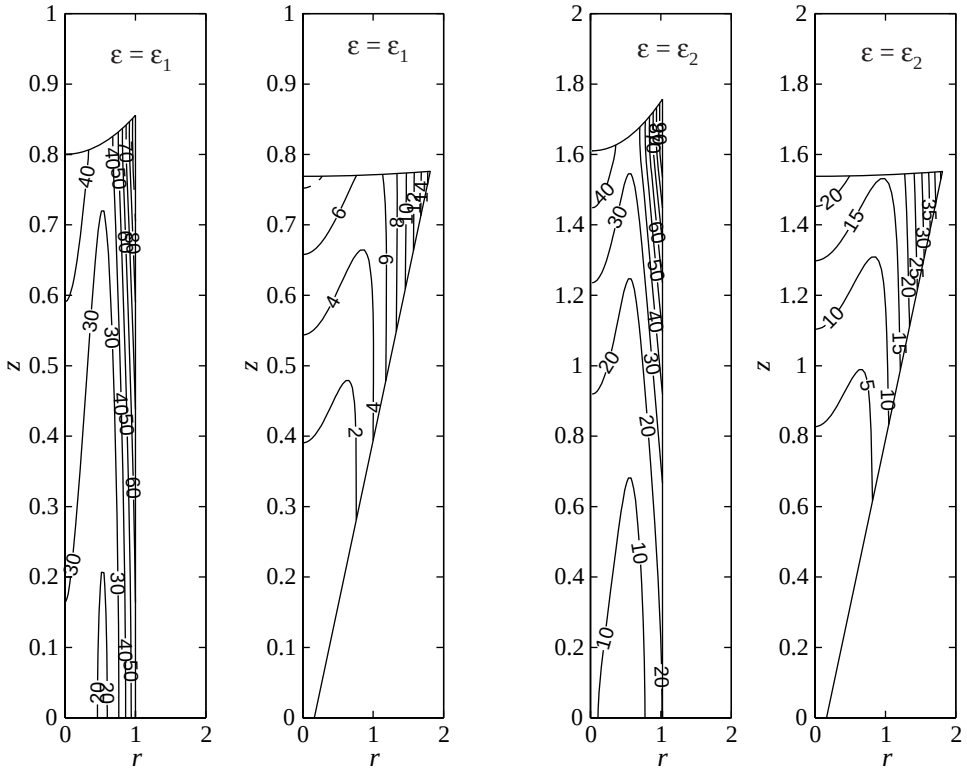


FIG. 11. *von Mises stress for the pseudosteady cylindrical and conical crystal cases. Isostress contours are labeled as a percent of the maximum stress (*0.176 *MPa for* ϵ_1 *and* 0.706 *MPa for* ϵ_2*).*

the angular coordinate. Figure 12 shows the stress pattern for a cylindrical crystal just inside the crystal-melt interface when the crystal is pulled in the directions [001], [111], $[10\bar{1}]$, and $[\bar{1}2\bar{1}]$ respectively. The temperature field corresponds to values of $h_{gs} = 4$, $\gamma = 0$, and $\delta = 0$ grown with the pseudosteady approximation. The $\langle 211 \rangle$ directions are preferred growth directions [23]. The other directions are for comparative purposes. Notice that the isostress contours are square for the [001] direction and hexagonal for the [111] direction, while the $[\bar{1}2\bar{1}]$ direction generates distorted rectangular isostress curves. If one assumes that the crystal will solidify in a manner consistent with minimizing the surface stress, then these curves should somewhat approximate the actual cross-sectional shape of the crystal as it is pulled from the melt. Clearly the crystal orientation can significantly reduce the stress. However, not all growth directions are amenable to crystal growth [23]. Because of these other issues, changes in the growth orientation are more effective at redistributing the stress within a particular cross-section than reducing the overall magnitude of stress. The issue of optimizing the growth conditions will be addressed in a subsequent paper.

6. Conclusion. In this study, we present a semianalytical approach for the temperature and thermal stress inside an InSb crystal. The purpose of the paper is twofold. By identifying the main physical features and using suitable mathematical models, we have gained useful insights into this complex manufacturing process. In particular, we have determined the dependence of the crystal stress on the evolving

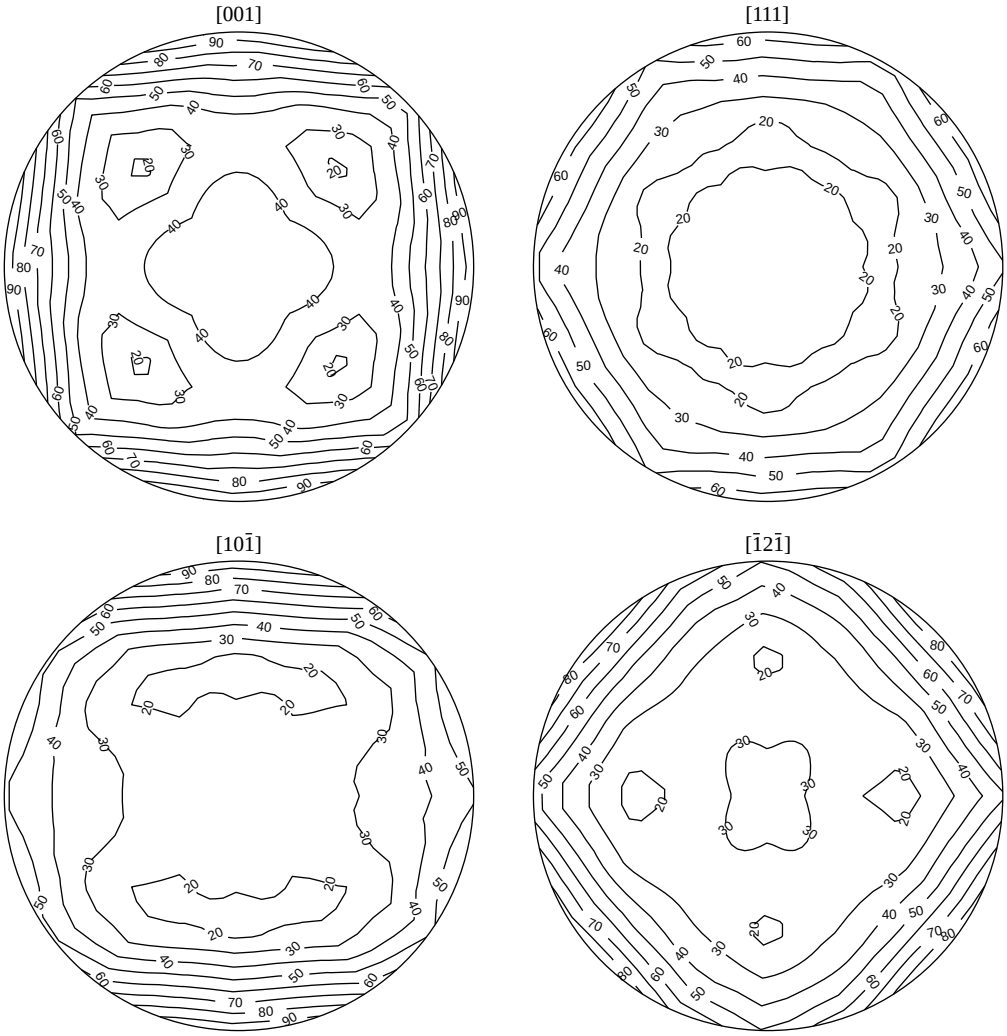


FIG. 12. *Total resolved stress distribution, computed using* (39)*, for a cylindrical crystal, at the indicated orientation, just inside the crystal-melt interface at the end of the growth. All reported stress values are expressed in percent with* 100% *occurring at the outer edge of a crystal grown in the* [001] *direction.*

crystal profile. By deriving semianalytical solutions, we have also provided a base that can be used to search for suitable growth conditions to improve the manufacturing procedure for all III-V compounds.

An important feature of our approach is that it allows us to derive explicit relationships between the thermal stress and relevant physical and geometrical parameters. This is achieved by using an asymptotic expansion of the solution in terms of the Biot number, characterizing the lateral heat flux. The asymptotic solution is obtained by solving essentially one-dimensional problems. The results show that the stress induced by radial temperature variation is related to the size (radius) and the profile (variation of the radius) of the crystal and heat flux through the side surface. On the other hand, the influence of the crystal radius on the stress induced by the

axial temperature variation is much weaker. The heat flux through the side surface is an important factor for reducing the overall thermal stress inside the crystal. The explicit nature of the thermal stress allows for a more efficient optimal control approach for finding better growth conditions, as shown in [15].

The other advantage of our semianalytical approach is that it can be extended to cases with more complicated models for the melt and gas flows. For example, the effect of the gas flow on the lateral heat flux between the crystal surface and the gas can be modeled by a nonconstant heat exchange coefficient h_{gs}. The motion of the melt can also be modeled by a similar approach, using a boundary layer argument [8, 9] or by solving the Navier–Stokes equations and temperature equation numerically. These will be the subject of a subsequent paper.

As pointed out earlier, we have assumed that the pulling rate can be adjusted to grow a crystal with a desirable shape. In practice, we may need to consider the dynamics and stability of the radial motion of the three-phase triple point. Models and computations have been carried out to capture the motion of the three-phase contact point [7, 32] and can be incorporated with the current model in a straightforward fashion. This work is currently underway.

Finally, we note that our study has its limitations and there is room for improvements and future investigations. For example, a nonuniform heat flux from the melt will introduce radial variation in the zeroth order temperature solution. As a result, the thermal stress will be determined not only by the axial gradient of the zeroth order solution but its radial variation as well. However, this radial variation is likely to be small, from our own observations and others, and an asymptotic solution can also be obtained, as indicated in [20]. Furthermore, we have not discussed the validity of the plane strain assumption. We believe that an asymptotic argument similar to that used for the temperature can be employed to derive the plane strain solution as part of the asymptotic series. We plan to address this issue also in a future study. Finally, the crystal grown in practice is not axisymmetric. It would be of practical interest to investigate the effect of anisotropy. Study is currently underway for a weakly anisotropic crystal.

Appendix: A simple model for the melt temperature. Starting with dimensional variables, we consider crystal growth in an axisymmetric setting where the rate of growth is small. We assume that the melt (liquid) in the crucible is well mixed and the temperature of the melt, $T_l(t)$, is uniform in space except in the thin layers near the crystal-melt and melt-ambient gas interfaces. We also assume that the ambient gas is well mixed and the temperature of the gas is a constant T_g. Furthermore, we will neglect the shape of the meniscus and assume that the crystal-melt and melt-gas interfaces are flat.[4] Therefore by adjusting the pulling speed v_p, the positions of the crystal-melt and melt-gas interfaces can be described by a single function $z = S(t)$.[5] Finally, we assume that the crystal radius $R(z)$ varies slowly in the z direction, $|R_z| \ll 1$. The coordinate system is fixed to the top of the growing crystal at $z = 0$, as described previously.

Figure 13 shows the three surfaces through which the melt can transfer heat

[4]For InSb crystals under consideration here, the typical length scale is $\overline{R} = 0.03$ m, the surface tension coefficient between the melt and gas is $\sigma_{gl} = 0.434$ N/m, and the melt density is $\rho_l = 6.47 \times 10^3$ kg/m^3. The Bond number is Bo $= \rho_l g \overline{R}^2/\sigma_{gl} \simeq 132 \gg 1$. Thus the meniscus is dominated by the gravity effect and the meniscus changes shape only near the three-phase contact point with a small capillary rise.

[5]The flat interface assumption allows one to drop the explicit r dependence of $S(t)$.

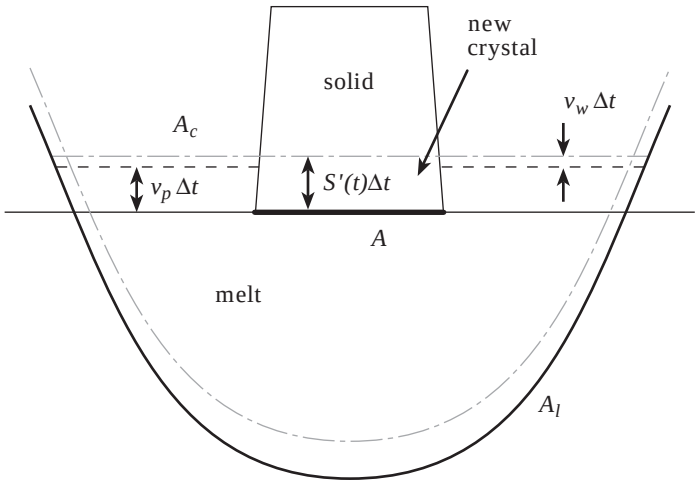


FIG. 13. *For a small time interval Δt the new crystal extends a distance $S'(t)\Delta t$ beyond the original interface location (dashed-dotted line). The corresponding drop in the melt is $v_m\Delta t$ (dashed line), and to realign the melt-gas interface and the crystal-melt the crystal must be extracted an additional distance of $v_p\Delta t$ (solid line).*

energy. These are denoted as $A = \pi R^2$, the area of the crystal-melt interface; $A_c = \pi R_c^2$, the cross-sectional area of the crucible ($A_c - A$ is the area of the melt-gas interface); and A_l, the surface area of the crucible in contact with the melt. Continuing to refer to the figure, conservation of mass implies that the rate at which the melt-gas interface drops due to the change in density upon solidification is

$$v_m(t) = \frac{\rho_s A}{\rho_l A_c} S'(t). \tag{40}$$

To ensure that the crystal-melt interface remains at the surface of the liquid the crucible is dropped at an optimal pull rate

$$v_p(t) = (S' - v_m)(t) = \frac{\rho_l A_c - \rho_s A}{\rho_l A_c} S'(t). \tag{41}$$

If the actual pull rate of the crystal exceeds v_p by a moderate amount, then the surface tension of the melt will cause the radius of the crystal to decrease. Similarly, pulling at a rate slower than v_p will cause the radius to increase.

Using these velocities, the position of the crystal-melt interface is $S = Z_0 + u_w + u_p$, where

$$u_w(t) = \int_0^t v_m(\tau)\, d\tau, \qquad u_p(t) = \int_0^t v_p(\tau)\, d\tau \tag{42}$$

are the displacements due to the loss of melt during the solidification and the growth of the crystal, respectively. The top of the crystal is fixed at $z = 0$.

During the growth period we assume that the heat flux from the melt to the crystal is given by

$$q_l = h_{sl}(T_l - T_m). \tag{43}$$

As a result the heat balance inside the melt yields

$$\frac{d}{dt}\left(\rho_l c_l T_l V_l\right) = -q_l A - h_{gl}(A_c - A)(T_l - T_g) + h_{cl} A_l (T_c - T_l), \tag{44}$$

where V_l is the time dependent volume of the melt, and T_c is the temperature of the crucible. The last term in (44) is the heat flux from the crucible to the liquid and is assumed to be a control parameter with T_c acting as the control.

For simulations in section 5.1.2 the crucible was assumed to be parabolic and filled to the extent that once the crystal reaches its final mass $\rho_s V_{\text{xtal}}$ and growth stops, there is a given proportion, p, of melt mass left in the crucible. As a result, $R_c(\xi) = \tilde{R}_c(\xi/\tilde{Z}_c)^{1/2}$, $0 \le \xi \le \tilde{Z}_c$, and the initial depth of the melt Z_{c0} is determined with the condition

$$\frac{\rho_s}{\rho_l} V_{\text{xtal}} = (1-p)\int_0^{Z_{c0}} \pi R_c^2(\xi)\, d\xi. \tag{45}$$

Since the shape of the crucible is known, A_l and V_l can be obtained as

$$A_l(t) = \int_0^{Z_{c0}-u_w(t)} 2\pi R_c(\xi)\sqrt{1+R_c'^2(\xi)}\, d\xi, \quad V_l(t) = \int_0^{Z_{c0}-u_w(t)} \pi R_c^2(\xi)\, d\xi \tag{46}$$

computed with respect to a local coordinate system fixed to the bottom of the crucible.

A nondimensionalized version of (44) is obtained by substituting the characteristic radius $\tilde{R}_c$ and depth $\tilde{Z}_c$ for the crucible so that

$$R_c = \tilde{R}_c \hat{R}_c, \qquad A = \overline{R}^2 \hat{A}, \qquad A_c = \tilde{R}_c^2 \hat{A}_c, \qquad A_l = \tilde{R}_c \tilde{Z}_c \hat{A}_l, \qquad V_l = \tilde{R}_c^2 \tilde{Z}_c \hat{V}_l.$$

Letting $T_l = T_g + \Delta T \Theta_l$ and $T_c = T_g + \Delta T \Theta_c$ and defining $\lambda = \overline{R}/\tilde{R}_c, \mu = \tilde{R}_c/\tilde{Z}_c, \phi = \rho_l c_l/\rho_s c_s$ one obtains

$$\frac{\phi}{\lambda\,\mathrm{St}}\frac{d}{d\hat{t}}\left[\left(\frac{T_g}{\Delta T}+\Theta_l\right)\hat{V}_l\right] = -\mu\lambda^2\hat{A}\frac{h_{sl}}{h_{gs}}(\Theta_l - 1) - \mu\frac{h_{gl}}{h_{gs}}(\hat{A}_c - \lambda^2\hat{A})\Theta_l + \frac{h_{cl}}{h_{gs}}\hat{A}_l(\Theta_c - \Theta_l) \tag{47a}$$

with $\Theta_l(0) = 1$ and a nondimensional heat flux given by

$$\gamma = \frac{q_{l,n}\overline{R}}{\epsilon^{1/2} k_s \Delta T} = \epsilon^{1/2}\frac{h_{sl}}{h_{gs}}(\Theta_l - 1). \tag{47b}$$

When commencing, the growth of the seed is slowly dropped until it contacts the melt surface and a meniscus is supported. Once the meniscus stabilizes and the seed reaches a thermal equilibrium with the melt and the crystal, the seed is extracted and the furnace temperature is slowly decreased [23]. Assuming a cylindrical seed of length Z_0 and radius R_0 and using expression (22) one finds an initial interface speed of

$$S_0'(0) = k\frac{k\sinh(kZ_0) + \delta\cosh(kZ_0) - \delta\Theta_{ch}}{k\cosh(kZ_0) + \delta\sinh(kZ_0)} \tag{47c}$$

with $k^2 = 2/R_0$. The initial crucible temperature is chosen so that $\Theta_l'(0) = 0$ and from (47a) one has

$$\Theta_c(0) = \left[1 + \mu\frac{h_{gl}}{h_{cl}}\left(\frac{\hat{A}_c - \lambda^2\hat{A}}{\hat{A}_l}\right) + \frac{\phi}{\lambda\,\mathrm{St}}\frac{h_{gs}}{h_{cl}}\frac{1}{\hat{A}_l}\frac{d\hat{V}_l}{d\hat{t}}\left(\frac{T_g}{\Delta T}+1\right)\right](0). \tag{47d}$$

To simulate the cooling of the furnace the crucible temperature was dropped at a constant rate of 0.2 K/hr for all simulations.

Nondimensional versions of the interface speeds (40)–(41), displacements (42), initial depth condition (45), and geometrical factors (46) are derived using the substitutions

$$Z_{c0} = \tilde{Z}_c \hat{Z}_{c0}, \qquad \epsilon^{1/2} u_p = \overline{R} \hat{u}_p, \qquad \epsilon^{1/2} u_w = \lambda^3 \tilde{R}_c \hat{u}_w.$$

Acknowledgments. The authors wish to thank Bill Micklethwaite and his engineers at Firebird Semiconductors Ltd. for sharing their insights and many stimulating discussions. We are also grateful to the anonymous referees for their comments and suggestions.

REFERENCES

[1] H. Alexander and P. Hassen (1968), *Dislocation and plastic flow in the diamond structure*, Solid State Physics, 22, pp. 27–158.
[2] W. A. Brantley (1973), *Calculated elastic constraints for stress problems associated with semiconductor devices*, J. Appl. Phys., 44, pp. 534–535.
[3] K. Brattkus and S. H. Davis (1988), *Directional solidification with heat losses*, J. Crystal Growth, 91, pp. 538–556.
[4] R. A. Brown (1988), *Theory of transport processes in single crystal growth from the melt*, AIChE J., 34, pp. 881–911.
[5] Y. T. Chan, H. J. Gilbeling, and H. L. Grubin (1988), *Numerical simulation of Czochralski growth*, J. Appl. Phys., 64, pp. 1425–1439.
[6] A. R. Chaudhuri, J. R. Patel, and L. G. Rubin (1962), *Velocities and densities of dislocations in germanium and other semiconductor crystals*, J. Appl. Phys., 33, pp. 2736–2746.
[7] A. B. Crowley (1983), *Mathematical modeling of heat flow in Czochralski crystal pulling*, IMA J. Appl. Math., 30, pp. 173–189.
[8] P. A. Davidson (1993), *Similarities in the structure of swirling and buoyancy-driven flows*, J. Fluid Mech., 252, pp. 357–382.
[9] P. A. Davidson and S. C. Flood (1994), *Natural convection in an aluminum ingot: A mathematical model*, Metallurgical and Materials Transactions B, 25, pp. 293–302.
[10] J. J. Derby and R. A. Brown (1988), *On the quasi-steady-state assumption in modeling Czochralski crystal growth*, J. Crystal Growth, 87, pp. 251–260.
[11] F. Dupret and N. van den Bogaert (1994), *Modelling Bridgeman and Czochralski growth*, in Handbook of Crystal Growth, Vol. 2B, D. T. J. Hurle, ed., North–Holland, Amsterdam, pp. 875–1010.
[12] A. Friedman (1964), *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ.
[13] A. N. Gulluoglu and C. T. Tsai (1999), *Effect of growth parameters on dislocation generation in InP single crystal grown by the vertical gradient freeze process*, Acta Materialia, 47, pp. 2313–2322.
[14] S. I. Hariharan and G. W. Young (2001), *Comparison of asymptotic solutions of a phase-field model to a sharp-interface model*, SIAM J. Appl. Math., 62, pp. 244–263.
[15] H. Huang and S. Liang, *An optimal control approach for thermal stress reduction inside a Cz grown crystal*, J. Engrg. Math., submitted.
[16] D. T. J. Hurle, ed. (1993), *Handbook of Crystal Growth, Vol.* 1: *Fundamentals*, North–Holland, Amsterdam.
[17] D. T. J. Hurle, ed. (1994), *Handbook of Crystal Growth, Vol.* 2: *Bulk Crystal Growth*, North–Holland, Amsterdam.
[18] D. T. J. Hurle (1993), *Crystal Pulling from the Melt*, Springer-Verlag, Berlin.
[19] A. S. Jordan, R. Caruso, and A. R. von Neida (1980), *A thermoelastic analysis of dislocation generation in pulled GaAs crystals*, Bell System Tech. J., 59, pp. 593–637.
[20] H. K. Kuiken and P. J. Roksnoer (1979), *Analysis of the temperature distribution in FZ silicon crystals*, J. Crystal Growth, 47, pp. 29–42.
[21] L. D. Landau and E. M. Lifshitz (1970), *Theory of Elasticity*, 2nd ed. Pergamon Press, Oxford, UK.

[22] O. A. Louchev, S. Kumaragurubaran, S. Takekawa, and K. Kitamura (2004), *Thermally induced effects during initial stage crystal growth from melts*, J. Crystal Growth, 273, pp. 320–328.
[23] W. F. Micklethwaite (2003), *private communications*, Firebird Semiconductors Ltd.
[24] W. F. Micklethwaite and A. J. Johnson (2000), *InSb Materials and devices*, in Infrared Detectors and Emitters: Materials and Devices, P. Capper and C. T. Elliot, eds., Kluwer Academic Publishers, Boston, pp. 177–204.
[25] M. G. Mil'vidskii and E. P. Bochkarev (1978), *Creation of defects during the growth of semiconductor single crystals and films*, J. Crystal Growth, 44, pp. 61–74.
[26] G. Müller (2002), *Experimental analysis and modeling of melt growth processes*, J. Crystal Growth, 237-239, pp. 1628–1637.
[27] G. Müller (1988), *Convection and Inhomogeneities in Crystal Growth from the Melt*, Crystals: Growth, Properties and Applications 12, H. C. Freyhardt, ed., Springer-Verlag, Berlin.
[28] J. L. Nowinski (1978), *Theory of Thermoelasticity with Applications*, Sijthoff & Noordhoff International Publishers, Alphen aan den Rijn, The Netherlands.
[29] V. Prasad, H. Zhang, and A. Anselmo (1997), *Transport phenomena in Czochralski crystal growth process*, Advances in Heat Transfer, 30, pp. 313–435.
[30] F. Quirós and J. L. Vázquez (2000), *Asymptotic convergence of the Stefan problem to Hele-Shaw*, Trans. Amer. Math. Soc., 353, pp. 609–634.
[31] T. Sinno, E. Dornberg, W. von Ammon, R. A. Brown, and F. Dupret (2002), *Defect engineering of Czochralski single-crystal silicon*, Material Science and Engineering Reports, 28, pp. 149–198.
[32] V. A. Tatarchenko (1993), *Shaped Crystal Growth*, Kluwer Academic Publishers, Boston.
[33] K. Tanahasji, M. Kikuchi, T. Higashino, N. Inoue, and Y. Mizokawa (2000), *Concentration of point defects changed by thermal stress in growing CZ silicon crystal: Effect of the growth rate*, J. Crystal Growth, 210, pp. 45–48.
[34] J. Völkl (1994), *Stress in the cooling crystal*, in Handbook of Crystal Growth, Vol. 2B, D. T. J. Hurle, ed., North–Holland, Amsterdam, pp. 821–874.
[35] S. Weinbaum, and L. M. Jiji (1977), *Singular perturbation theory for melting or freezing in finite domains initially not at the fusion temperature*, J. Appl. Mech., 44, pp. 25–30.
[36] G. W. Young and A. Chait (1990), *Surface tension driven heat, mass, and momentum transport in a two-dimensional float-zone*, J. Crystal Growth, 106, pp. 445–466.
[37] G. W. Young and J. A. Heminger (1997), *Modeling the time-dependent growth of single-crystal fibers*, J. Crystal Growth, 178, pp. 410–421.
[38] G. W. Young and J. A. Heminger (2000), *A mathematical model of the edge-defined film-fed growth proccess*, J. Engrg. Math., 38, pp. 371–390.
[39] M. M. Yan and P. N. S. Huang (1979), *Perturbation solutions to phase change problem subject to convection and radiation*, J. Heat Transfer, 101, pp. 96–100.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1563–1587

DISCRETE-TIME SIS EPIDEMIC MODEL IN A SEASONAL ENVIRONMENT*

JOHN E. FRANKE† AND ABDUL-AZIZ YAKUBU‡

Abstract. We study the combined effects of seasonal trends and diseases on the extinction and persistence of discretely reproducing populations. We introduce the epidemic threshold parameter, $\mathcal{R}_0$, for predicting disease dynamics in periodic environments. Typically, in periodic environments, $\mathcal{R}_0 > 1$ implies disease persistence on a cyclic attractor, while $\mathcal{R}_0 < 1$ implies disease extinction. We also explore the relationship between the demographic equation and the epidemic process. In particular, we show that in periodic environments, it is possible for the infective population to be on a chaotic attractor while the demographic dynamics is nonchaotic.

Key words. epidemics, infectives, periodic environments, susceptibles

AMS subject classifications. 37G15, 37G35, 39A11, 92B05

DOI. 10.1137/050638345

1. Introduction.

The complexities of a periodic environment can significantly affect the regulation of species [26]. In periodic environments, population sizes are often either enhanced via *resonance* or diminished via *attenuance* [5, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 23, 24, 25, 30, 31, 32, 33, 34, 35, 37, 38, 45, 48, 50]. However, most epidemic models in the literature (with a few exceptions) neglect seasonal factors [3, 4, 12]. For example, Allen and Burgin [1], Allen [2], and Castillo-Chavez and Yakubu [7, 8, 9] studied disease invasions in discretely reproducing populations that live on attractors in constant (nonperiodic) environments. Cushing and Henson [14], Elaydi and Sacker [17, 18, 19, 20], Franke and Yakubu [23, 24], Kocic [35], Kocic and Ladas [36], Kon [37, 38], and others have studied the effects of periodic environments on ecological models without explicit disease dynamics [46].

In this paper, we focus on the impact of seasonal factors on a discrete-time SIS (susceptible-infected-susceptible) epidemic model. The model reduces to the SIS epidemic model of Castillo-Chavez and Yakubu when the environment is constant (nonperiodic) [7, 8, 9]. To understand the impact of seasonality and disease on life-history outcomes, we study the long-term dynamics of our model under specific functional forms for the periodic recruitment function. The periodic Beverton–Holt [6], the periodic constant, and the periodic Malthus (geometric growth) models are the periodic recruitment functions for this study [7, 8, 9].

We assume that a disease invades and subdivides the target population into two classes: susceptibles (noninfectives) and infectives. Prior to the time of disease invasion, the population is assumed to be governed by a periodically forced demographic equation with a periodic recruitment function. Hence, the population is assumed to be either at a demographic "steady state" (an attracting cycle) or growing at a periodic

*Received by the editors August 17, 2005; accepted for publication (in revised form) January 30, 2006; published electronically June 9, 2006. This work was partially supported by funds from the NOAA-NEFSC in Woods Hole, MA.
http://www.siam.org/journals/siap/66-5/63834.html
†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (franke@math.ncsu.edu).
‡Department of Mathematics, Howard University, Washington, DC 20059 (ayakubu@howard.edu).

geometric rate. The transition from susceptible to infective is a function of the contact rate α (between individuals) and the proportion of infectives (prevalence) in the population. We derive the epidemic threshold parameter, $\mathcal{R}_0$, for predicting disease persistence or extinction in periodic environments. We also explore the relationship between the demographic equation and the epidemic process. Castillo-Chavez and Yakubu, in [7, 8, 9], show that in constant environments the demographic equation drives the disease dynamics. In stark contrast, we use numerical simulations to show that in periodic environments the demographic equation does not always drive the disease dynamics. We show that, in periodic environments, it is possible for the infective population to be on a chaotic attractor while the demographic dynamics is nonchaotic.

The paper is organized as follows. In section 2, we introduce the periodically forced demographic equation for the study. The equation, a nonautonomous nonlinear difference equation with periodic recruitment function, describes the dynamics of the (total) population before disease invasion. We review, in section 2, the results of Franke and Yakubu on periodically forced recruitment functions. The main model, a periodically forced discrete-time SIS epidemic model, is constructed in section 3. When the recruitment function is either a periodic constant or the periodic Beverton–Holt model, then the total population is persistent and lives on a globally attracting cycle. Autonomous discrete-time models do not support (nontrivial) globally stable cycles [21]. In section 4, the basic reproductive number $\mathcal{R}_0$ is introduced and used to predict the (uniform) persistence or extinction of the infective population, where the recruitment function is either a periodic constant or the periodic Beverton–Holt model. Section 5 covers the SIS epidemic model under asymptotically cyclic demographic dynamics, while sections 6 and 7 describe the epidemic model under geometric demographic dynamics. As in section 4, in section 6, $\mathcal{R}_0$ is used to predict the (uniform) persistence or extinction of the proportion of infectives in the total population. Conditions for disease persistence on cyclic attractors are introduced in section 7.

Periodically forced population models support multiple attractors, and we use numerical simulations to show, in section 8, that our periodic epidemic model supports multiple attractors. Section 9 is on period-doubling bifurcations in the epidemic model where the demographic dynamics is simple and nonchaotic. The implications of our results are discussed in section 10.

2. Demographic equations with seasonality. In constant environments, theoretical discrete-time epidemic models are usually formulated under the assumption that the dynamics of the total population size in generation t, denoted by $N(t)$, is governed by equations of the form

$$N(t+1) = f(N(t)) + \gamma N(t), \tag{1}$$

where $\gamma \in (0,1)$ is the constant "probability" of surviving per generation and $f : \mathbb{R}_+ \to \mathbb{R}_+$ models the birth or recruitment process [7, 9].

Seasonality can be introduced into (1) by writing the recruitment function as a p-periodically forced function. This is modeled with the *p-periodic demographic equation*

$$N(t+1) = f(t, N(t)) + \gamma N(t), \tag{2}$$

where $\exists\ p \in \mathbb{N}$ such that

$$f(t, N(t)) = f(t+p, N(t)) \quad \forall t \in \mathbb{Z}_+.$$

We assume throughout that $f(t,N) \in C^2(\mathbb{Z}_+ \times \mathbb{R}_+, \mathbb{R}_+)$ and $\gamma \in (0,1)$ [25].

Franke and Yakubu, in [25], studied model (2) with the periodic constant recruitment function

$$f(t,N(t)) = k_t(1-\gamma)$$

and with the periodic Beverton–Holt recruitment function

$$f(t,N(t)) = \frac{(1-\gamma)\mu k_t N(t)}{(1-\gamma)k_t + (\mu - 1 + \gamma)N(t)},$$

where the carrying capacity k_t is *p-periodic*, $k_{t+p} = k_t$ for all $t \in \mathbb{Z}_+$ [14, 25]. Franke and Yakubu proved that the periodically forced recruitment functions generate globally attracting cycles in model (2) [25]. We summarize their results in the following two theorems.

THEOREM 1. *Model* (2) *with* $f(t,N(t)) = k_t(1-\gamma)$ *has a globally attracting positive s-periodic cycle that starts at*

$$\overline{x}_0 = \frac{(1-\gamma)\left(k_{p-1} + k_{p-2}\gamma + \cdots + k_0\gamma^{p-1}\right)}{1-\gamma^p},$$

where s divides p.

THEOREM 2. *Model* (2) *with* $f(t,N(t)) = \frac{(1-\gamma)\mu k_t N(t)}{(1-\gamma)k_t + (\mu-1+\gamma)N(t)}$ *and* $\mu > 1$ *has a globally attracting positive s-cycle, where s divides p.*

Theorems 1 and 2 imply that the total population is asymptotically periodic (bounded) and lives on a cyclic attractor when the recruitment function is either a periodic constant or the Beverton–Holt model. Denote this cycle by $\{\overline{N}_0, \overline{N}_1, \ldots, \overline{N}_{s-1}\}$. When new recruits arrive at the periodic positive per-capita growth rate λ_t, then

$$f(t,N(t)) = \lambda_t N(t),$$

where $\lambda_{t+p} = \lambda_t$ for all $t \in \mathbb{Z}_+$. The solution to the demographic equation is

$$N(t) = \left(\prod_{J=0}^{t-1} (\lambda_J + \gamma)\right) N(0),$$

and the demographic basic reproductive number is

$$\mathcal{R}_d = \frac{\prod_{J=0}^{p-1} (\lambda_J + \gamma) - \gamma^p}{1-\gamma^p}. \tag{3}$$

$\mathcal{R}_d$ gives the average number of descendants produced by a typically small initial population over a *p-cycle*. If $\mathcal{R}_d < 1$, the total population goes extinct at a geometric rate, and if $\mathcal{R}_d > 1$, the total population explodes at a geometric rate. In constant environments, $p = 1$, $\lambda_J = \lambda$, and $\mathcal{R}_d$ reduces to

$$\mathcal{R}_d = \frac{\lambda}{1-\gamma}.$$

In [7, 8, 9], Castillo-Chavez and Yakubu used $\mathcal{R}_d = \frac{\lambda}{1-\gamma}$ to study the long-term behavior of geometrically growing populations in constant environments.

3. SIS epidemic model in periodic environments. In this section, we introduce the main model, an SIS epidemic model with periodic forcing. To do this, we assume that a nonfatal infectious disease has invaded a population living in a seasonal environment. The population is governed by (2). To model the disease, we build a simple SIS epidemic process on "top" of the periodic demographic equation. We let $S(t)$ denote the population of susceptibles; $I(t)$ denotes the population of the infected, assumed infectious; $N(t) \equiv S(t) + I(t)$ denotes the total population size at generation t, N_∞ denotes the demographic steady state or attracting population, and $\overline{N}_0$ the initial point on a globally attracting cycle, when they exist. We assume that individuals survive with constant probability γ each generation, and infected individuals recover with constant probability $(1-\sigma)$.

Let $\phi : [0,\infty) \to [0,1]$ be a monotone concave probability function with $\phi(0) = 1$, $\phi'(x) < 0$, and $\phi''(x) \geq 0$ for all $x \in [0,\infty)$. We assume that the susceptible individuals become infected with nonlinear probability $\left(1 - \phi\left(\alpha \frac{I}{N}\right)\right)$ per generation, where the transmission constant $\alpha > 0$. When infections are modeled as Poisson processes, then $\phi\left(\alpha \frac{I}{N}\right) = e^{-\alpha \frac{I}{N}}$ [7, 8, 9].

Our assumptions and notation lead to the following SIS epidemic model in period p environments:

$$\left.\begin{aligned} S(t+1) &= f(t, N(t)) + \gamma\phi\left(\alpha \tfrac{I(t)}{N(t)}\right) S(t) + \gamma(1-\sigma) I(t), \\ I(t+1) &= \gamma\left(1 - \phi\left(\alpha \tfrac{I(t)}{N(t)}\right)\right) S(t) + \gamma\sigma I(t), \end{aligned}\right\} \tag{4}$$

where $0 < \gamma, \sigma < 1$ and $N(t) > 0$. When the environment is constant, $f(t, N(t)) = f(N(t))$ and model (4) reduces to the model of Castillo-Chavez and Yakubu [7, 8, 9]. The total population in generation $t+1$, $S(t+1) + I(t+1)$, the sum of the two equations of model (4), is the demographic equation (2). Using the substitution $S(t) = N(t) - I(t)$, the I-equation in model (4) becomes

$$I(t+1) = \gamma\left(1 - \phi\left(\alpha \frac{I(t)}{N(t)}\right)\right)(N(t) - I(t)) + \gamma\sigma I(t).$$

Let

$$F_N(I) = \gamma\left(1 - \phi\left(\alpha \frac{I}{N}\right)\right)(N - I) + \gamma\sigma I.$$

When F_N has a unique positive fixed point and critical point, we denote them by I_N and C_N, respectively.

$$I(t+1) = F_{N(t)}(I(t)),$$

and the set of iterates of the nonautonomous map $F_{N(t)}$ is the set of density sequences generated by the infective equation. In the next section, we use the map F_N to study disease dynamics in the periodic SIS epidemic model, model (4).

4. Disease extinction versus disease persistence. The classical theory of disease epidemics usually involves computation of an epidemic threshold parameter, the basic reproductive number $\mathcal{R}_0$ [3]. Here, we introduce $\mathcal{R}_0$ and use it to predict the successful invasion or extinction of the disease modeled in (4). In constant environments $f(t, N(t)) = f(N(t))$, and

$$\mathcal{R}_0 = \frac{-\gamma\alpha\phi'(0)}{1 - \gamma\sigma}. \tag{5}$$

$\mathcal{R}_0$ is the average number of secondary infections generated by an initial population of infected (assumed infectious) individuals over their lifetimes [7, 8, 9].

In our periodic model, we use the same $\mathcal{R}_0$ to prove that $\mathcal{R}_0 < 1$ implies disease extinction and $\mathcal{R}_0 > 1$ implies disease persistence. To prove this result we need the following definition [49].

DEFINITION 3. *The total population in model* (2) *is persistent if*

$$\lim_{t\to\infty} \inf N(t) > 0$$

whenever $N(0) > 0$. *The total population is uniformly persistent if there exists a positive constant* η *such that*

$$\lim_{t\to\infty} \inf N(t) \geq \eta$$

whenever $N(0) > 0$.

By this definition, when the recruitment function is either a periodic constant or the Beverton–Holt model, the total population is uniformly persistent. Also, when new recruits arrive at the periodic positive per-capita growth rate λ_t and $\mathcal{R}_d > 1$, the total population is uniformly persistent. However, the population goes extinct when $\mathcal{R}_d < 1$.

The following auxiliary lemmas will be used to prove our results.

LEMMA 4. *If* $0 < I(t) \leq N(t)$ *in model* (4), *then* $I(t+1) < \min\{N(t), N(t+1)\}$.

Proof. In model (4),

$$I(t+1) = \gamma\left(1 - \phi\left(\alpha\frac{I(t)}{N(t)}\right)\right)S(t) + \gamma\sigma I(t)$$

and

$$N(t+1) = S(t+1) + I(t+1) = f(t, N(t)) + \gamma N(t).$$

Therefore,

$$\begin{aligned} I(t+1) &= \gamma\left(1 - \phi\left(\alpha\frac{I(t)}{N(t)}\right)\right)(N(t) - I(t)) + \gamma\sigma I(t) \\ &< \gamma(N(t) - I(t)) + \gamma I(t) = \gamma N(t) \\ &= N(t+1) - f(t, N(t)) \leq N(t+1). \end{aligned}$$

Hence,

$$I(t+1) < \min\{N(t), N(t+1)\}. \qquad \square$$

LEMMA 5. *If* $I(0) > 0$ *in model* (4), *then* $I(t) > 0$ *for all* $t \in \mathbb{Z}_+$.

Proof. $I(t+1) = \gamma\left(1 - \phi\left(\alpha\frac{I(t)}{N(t)}\right)\right)(N(t) - I(t)) + \gamma\sigma I(t)$. By Lemma 4, $N(t) - I(t) \geq 0$ for all $t \in \mathbb{Z}_+$. Therefore, $\gamma\left(1 - \phi\left(\alpha\frac{I(t)}{N(t)}\right)\right)(N(t) - I(t)) \geq 0$. $I(0) > 0$ implies $\gamma\sigma I(0) > 0$, and hence $I(1) > 0$. By induction, $I(t) > 0$ and $\gamma\sigma I(t) > 0$. Hence, $I(t+1) > 0$. $\square$

LEMMA 6.

$$F_N(I) = \gamma\left(1 - \phi\left(\alpha\frac{I}{N}\right)\right)(N - I) + \gamma\sigma I$$

satisfies the following conditions:

(a) $F_N'(0) = -\alpha\gamma\phi'(0) + \gamma\sigma$ *and* $F_N'(N) > -1$.
(b) $F_N(I)$ *is concave down on* $[0, N]$.
(c) $F_N(I) \le F_N'(0)I$ *on* $[0, N]$.
(d) *If* $F_N'(0) > 1$, *then* F_N *has a unique positive fixed point* I_N *in* $[0, N]$.
(e) *Let* $\Psi_N(I) = \frac{I}{N}$. *Then* $F_1(\Psi_N(I)) = \Psi_N(F_N(I))$. *That is,* Ψ_N *is a topological conjugacy between* F_1 *and* F_N.
(f) *If* $N_0 < N_1$ *and* $(-\alpha\gamma\phi'(0) + \gamma\sigma) > 1$, *then* $I_{N_0} < I_{N_1}$ *where* I_{N_i} *is the positive fixed point of* F_{N_i} *in* $[0, N_i]$.
(g) *If* C_1 *exists, then* $C_N = NC_1$.
(h) *If* $N_0 < N_1$, *then* $F_{N_0}(I) < F_{N_1}(I)$ *for all* $I \in (0, N_0]$.

Proof. (a)

$$F_N'(I) = -\frac{\alpha\gamma}{N}\phi'\left(\alpha\frac{I}{N}\right)(N - I) - \gamma\left(1 - \phi\left(\alpha\frac{I}{N}\right)\right) + \gamma\sigma,$$

$$\begin{aligned} F_N'(0) &= -\frac{\alpha\gamma}{N}\phi'(0)(N - 0) - \gamma(1 - \phi(0)) + \gamma\sigma \\ &= -\alpha\gamma\phi'(0) + \gamma\sigma, \end{aligned}$$

$$\begin{aligned} F_N'(N) &= -\frac{\alpha\gamma}{N}\phi'\left(\alpha\frac{N}{N}\right)(N - N) - \gamma\left(1 - \phi\left(\alpha\frac{N}{N}\right)\right) + \gamma\sigma \\ &= -\gamma(1 - \phi(\alpha)) + \gamma\sigma > -\gamma > -1. \end{aligned}$$

(b)

$$F_N''(I) = -\left(\frac{\alpha}{N}\right)^2\gamma\phi''\left(\alpha\frac{I}{N}\right)(N - I) + 2\frac{\alpha\gamma}{N}\phi'\left(\alpha\frac{I}{N}\right).$$

Since $\phi' < 0$ and $\phi'' \ge 0$ on $[0, \infty)$, we have

$$F_N''(I) < 0 \quad \text{on } [0, N].$$

(c) $F_N(0) = 0$ implies that $y = F_N'(0)I$ is the tangent line to the graph of $F_N(I)$ at 0. Since F_N is concave down on $[0, N]$, its graph is below the tangent line at the origin on $[0, N]$. Hence,

$$F_N(I) \le F_N'(0)I \quad \text{on } [0, N].$$

(d) $F_N(N) = \gamma\sigma N < N$. Since $F_N'(0) > 1$, the graph of $F_N(I)$ starts out higher than the diagonal and must cross it before $I = N$. The concavity property of $F_N(I)$ (see (b)) implies that there is a unique positive fixed point.

(e) $F_1(I) = \gamma(1 - \phi(\alpha I))(1 - I) + \gamma\sigma I$. Thus,

$$F_1(\Psi_N(I)) = \gamma\left(1 - \phi\left(\alpha\frac{I}{N}\right)\right)\left(1 - \frac{I}{N}\right) + \gamma\sigma\frac{I}{N} = \frac{1}{N}F_N(I) = \Psi_N(F_N(I)).$$

(f) Since $F_{N_0}'(0) = (-\alpha\gamma\phi'(0) + \gamma\sigma) > 1$, I_{N_0} exists with $F_{N_0}(I_{N_0}) = I_{N_0}$. Thus $\Psi_{N_0}(F_{N_0}(I_{N_0})) = \Psi_{N_0}(I_{N_0}) = F_1(\Psi_{N_0}(I_{N_0}))$. That is $\Psi_{N_0}(I_{N_0}) = I_1$, the unique positive fixed point of F_1, and $I_{N_0} = N_0 I_1$. Similarly, $I_{N_1} = N_1 I_1$. Hence, $N_0 < N_1$ implies $I_{N_0} < I_{N_1}$.

(g) Topological conjugacy preserves critical points. The result follows from (e).

(h) Let $N_0 < N_1$ and $I \in (0, N_0]$. The topological conjugacy in part (e) shows that $F_{N_0}(I) = N_0 F_1(\frac{I}{N_0})$ and $F_{N_1}(I) = N_1 F_1(\frac{I}{N_1})$. Note that $\frac{I}{N_1} < \frac{I}{N_0}$. Since the graph of F_1 goes through the origin with positive slope and is concave down, the ray through the origin and $\big(\frac{I}{N_1}, F_1(\frac{I}{N_1})\big)$ has a larger slope than the ray through the origin and $\big(\frac{I}{N_0}, F_1(\frac{I}{N_0})\big)$. The first ray contains the point $\big(I, N_1 F_1(\frac{I}{N_1})\big)$, while the second ray contains $\big(I, N_0 F_1(\frac{I}{N_0})\big)$. Hence, $F_{N_1}(I) = N_1 F_1(\frac{I}{N_1}) < N_0 F_1(\frac{I}{N_0}) = F_{N_0}(I)$. □

THEOREM 7. *Let the total population in model* (2) *be uniformly persistent.*

(a) *If* $\mathcal{R}_0 < 1$, *then in model* (4), $\lim_{t\to\infty} I(t) = 0$ *whenever* $I(0) \le N(0)$. *That is, the disease goes extinct.*

(b) *If* $\mathcal{R}_0 > 1$, *then in model* (4), $\exists\ \eta > 0$ *such that* $\lim_{t\to\infty} \inf I(t) \ge \eta$ *whenever* $N(0) \ge I(0) > 0$. *That is, the disease persists uniformly.*

Proof. Since $I(0) \le N(0)$, Lemma 4 implies that $I(t) \le N(t)$ for all $t \in \mathbb{Z}_+$.

(a) $\mathcal{R}_0 = \frac{-\gamma\alpha\phi'(0)}{1-\gamma\sigma} < 1$ is equivalent to $-\alpha\gamma\phi'(0) + \gamma\sigma < 1$. Lemma 6 gives $F'_N(0) = F'_{N(t)}(0) = -\alpha\gamma\phi'(0) + \gamma\sigma < 1$ and $I(t+1) = F_{N(t)}(I(t)) \le F'_{N(t)}(0) I(t)$. Thus, the sequence $\{I(t)\}$ is dominated by the geometrically decreasing sequence $\{(-\alpha\gamma\phi'(0) + \gamma\sigma)^t I(0)\}$, and hence

$$\lim_{t\to\infty} I(t) = 0.$$

(b) Lemma 5 implies that $I(t) > 0$ for all $t \in \mathbb{Z}_+$. Lemma 6 gives $F'_N(0) = F'_{N(t)}(0) = -\alpha\gamma\phi'(0) + \gamma\sigma > 1$. Since $I(t+1) = F_{N(t)}(I(t))$, $I(t+1) > I(t)$ on the open interval $\big(0, I_{N(t)}\big)$. If $I(t) \in (I_{N(t)}, N(t))$, $I(t+1) \ge \min\{I_{N(t)} = N(t)I_1, F_{N(t)}(N(t)) = \gamma\sigma N(t)\}$. Since the total population is uniformly persistent, $\exists\ \widehat{\eta} > 0$ satisfying $\lim_{t\to\infty} \inf N(t) \ge \widehat{\eta}$ whenever $N(0) > 0$. This implies that $\exists\ \eta > 0$ such that

$$\lim_{t\to\infty} \inf\left(\min\{N(t)I_1, \gamma\sigma N(t)\}\right) \ge \eta > 0.$$

Thus, the orbit $\{I(t)\}$ increases when it is small and eventually gets larger and remains larger than a fixed positive number. Hence, $\exists\ \eta > 0$ satisfying

$$\lim_{t\to\infty} \inf I(t) \ge \eta. \qquad \square$$

A slight modification of the proof of Theorem 7 reveals that uniform persistence can be replaced with persistence in the hypothesis and conclusion. That is, if the total population persists, then the disease persists whenever $\mathcal{R}_0 > 1$.

When the recruitment function is either a periodic constant or the periodic Beverton–Holt model, then the (total) population is uniformly persistent. If, in addition, $\mathcal{R}_0 < 1$, then in model (4), $\lim_{t\to\infty} I(t) = 0$, and the disease goes extinct. However, if $\mathcal{R}_0 > 1$, then in model (4), $\lim_{t\to\infty} \inf I(t) \ge \eta > 0$, and the disease persists uniformly (Theorem 7).

In constant environments, when the total population lives on a globally attracting positive fixed point, $\mathcal{R}_0 > 1$ implies uniform persistence of the infectives on a globally attracting positive fixed point [7, 8, 9]. With the advent of periodicity, when the total population lives on an attracting cycle, $\mathcal{R}_0 > 1$ implies uniform persistence of the infectives on a globally attracting cycle (section 5), multiple cyclic attractors (section 8), or a chaotic attractor (section 9). We summarize these results in the following corollary.

COROLLARY 8. *If the demographic equation, model* (2)*, has a globally attracting p-cycle* ($p > 1$) *and* $\mathcal{R}_0 > 1$*, then the uniform persistent infective population in model* (4) *is not on a fixed point attractor.*

Proof. By Theorem 7, the infective population in model (4) is uniformly persistent when $\mathcal{R}_0 > 1$. To establish this result, we use a contradiction proof to show that the infective population in model (4) has no positive fixed point when the demographic equation, model (2), has a globally attracting p-cycle ($p > 1$).

Assume that $(N(0), I(0))$ is an initial condition where $\{I(t)\}$ is constantly fixed at $I(0) > 0$. Now $I(t+1) = F_{N(t)}(I(t)) = I(0)$. Lemma 6 gives the fixed point of $F_{N(t)} = N(t)I_1 = I(0)$. Hence, $\{N(t)\}$ is constantly fixed at $N(0)$. Since all initial total populations are attracted to a nontrivial cycle, we have a contraction. Hence, the uniformly persistent infective population in model (4) is not on a fixed point attractor. □

5. Asymptotically cyclic epidemics. We now study the long-term disease dynamics for a population living in a seasonal environment, where the p-periodic demographic equation has a globally attracting positive cycle $\{\overline{N}_0, \overline{N}_1, \ldots, \overline{N}_{p-1}\}$. For example, when the recruitment function is either periodically constant or periodic Beverton–Holt, the demographic equation is asymptotically cyclic (Theorems 1 and 2). If in addition $\mathcal{R}_0 > 1$, we show that it is possible for the uniformly persistent epidemic to live on a globally attracting cycle. That is, the demographic dynamics drives the disease dynamics. To predict this long-term dynamics of the epidemic process, we use the very general "limiting systems" theory of Franke and Yakubu [23].

The general theory of Franke and Yakubu uses the following periodic hierarchical system:

$$\left.\begin{array}{rcll} x(t+1) & = & g(t,x(t)), & x(0) = x \in R_+^n, \\ y(t+1) & = & h(t,x(t),y(t)), & (x(0),y(0)) = (x,y) \in V \subseteq R_+^{n+m}, \end{array}\right\} \tag{6}$$

where $g : Z_+ \times R_+^n \to R_+^n$ and $h : Z_+ \times V \to R_+^m$ are smooth functions and where there exist smallest positive integers T_1 and T_2 satisfying $g(t+T_1, x(t)) = g(t, x(t))$ and $h(t+T_2, x(t), y(t)) = h(t, x(t), y(t))$, respectively.

Let

$$V = \{(N, I) : I \leq N\}.$$

Then V is a connected set, and for each $N \in \mathbb{R}_+$

$$\{I \in \mathbb{R}_+ : (N, I) \in V\}$$

is a connected set. Lemma 4 shows that the (N, I) system,

$$\left.\begin{array}{l} N(t+1) = f(t, N(t)) + \gamma N(t), \\ I(t+1) = \gamma \left(1 - \phi\left(\alpha \frac{I(t)}{N(t)}\right)\right)(N(t) - I(t)) + \gamma\sigma I(t), \end{array}\right\} \tag{7}$$

is an example of model (6).

System (6) is the sequence of maps $\{G_i\}$, where for each $i \in \mathbb{Z}_+$, $G_i : V \to V$ is defined by

$$G_i(x, y) = (g(i \bmod(T_1), x), h(i \bmod(T_2), x, y)) \equiv (g_i(x), h_i(x, y)).$$

G_i has period $T = \operatorname{lcm}(T_1, T_2)$.

To define a limiting system for system (6), we assume that $\{x_0, x_1, \dots, x_{k-1}\}$ is a *k-cycle* for the T_1*-periodic* dynamical system $\{g_0, g_1, \dots, g_{T_1-1}\}$. For each $i \in \mathbb{Z}_+$, define the sets $V_i = \{y \in \mathbb{R}_+^m : (x_{i \bmod(k)}, y) \in V\}$. Also, define the periodically forced (nonautonomous) maps

$$\widehat{G}_i : \mathbb{R}_+^n \times V_i \to \mathbb{R}_+^n \times V_{i+1} \quad \text{by} \quad \widehat{G}_i(x,y) = (g_i(x), h_i(x_{i \bmod(k)}, y)),$$

and

$$\widehat{H}_i : V_i \to V_{i+1} \quad \text{by} \quad \widehat{H}_i(y) = h_i(x_{i \bmod(k)}, y).$$

Note that

$$\widehat{H}_{kT_2-1} \circ \cdots \circ \widehat{H}_1 \circ \widehat{H}_o : V_0 \to V_0.$$

The periodic system $\{\widehat{G}_0, \widehat{G}_1, \dots, \widehat{G}_{q-1}, \dots\}$ is a limiting system of model (6) when the *k-cycle* is attracting.

Inserting cycle $\{\overline{N}_0, \overline{N}_1, \dots, \overline{N}_{p-1}\}$ into model (7) produces the limiting system

$$\left.\begin{aligned} N(t+1) &= f(t, N(t)) + \gamma N(t), \\ I(t+1) &= \gamma \left(1 - \phi\left(\alpha \tfrac{I(t)}{\overline{N}_t}\right)\right)(\overline{N}_t - I(t)) + \gamma\sigma I(t). \end{aligned}\right\} \tag{8}$$

The second equation of system (8) is $F_{\overline{N}_t}(I(t))$.

The following straightforward generalization of a theorem of Franke and Yakubu gives conditions under which the long-term qualitative dynamics of the nonautonomous system (6) is equivalent to that of the limiting system.

THEOREM 9 (see [23]). *Assume that all orbits of system* (6) *are bounded, V is a connected set, and for each* $x \in \mathbb{R}_+^n$

$$\{y \in \mathbb{R}_+^m : (x,y) \in V\}$$

is a connected set. Then system (6) *has*

$$\{(x_0, y_0), (x_1, y_1), \dots, (x_{l-1}, y_{l-1}), \dots\}$$

as a globally attracting cycle if and only if

$$\{x_0, x_1, \dots, x_{k-1}, \dots\}$$

is a globally attracting k-cycle of the T_1-periodic *dynamical system* $\{g_0, g_1, \dots, g_{T_1-1}\}$ *and* y_0 *is a globally attracting fixed point of the composition* $\widehat{H}_{kT_2-1} \circ \cdots \circ \widehat{H}_1 \circ \widehat{H}_o$.

To apply Theorem 9, we need the following result.

COROLLARY 10. *Assume that all orbits of system* (7) *are bounded. Then system* (7) *has*

$$\{(\overline{N}_0, \overline{I}_0), (\overline{N}_1, \overline{I}_1), \dots, (\overline{N}_{p-1}, \overline{I}_{p-1}), \dots\}$$

as a globally attracting cycle if and only if $\{\overline{N}_0, \overline{N}_1, \dots, \overline{N}_{p-1}, \dots\}$ *is a globally attracting cycle for the p-*periodic *dynamical system* $\{g_0, g_1, \dots, g_{p-1}\}$, *where*

$$g_i(N) = f(i, N) + \gamma N,$$

and $\overline{I}_0$ *is a globally attracting fixed point of the composition* $\widehat{H}_{p-1} \circ \cdots \circ \widehat{H}_1 \circ \widehat{H}_o$, *where*

$$\widehat{H}_i(I) = F_{\overline{N}_i}(I) = \gamma \left(1 - \phi\left(\alpha \frac{I}{\overline{N}_i}\right)\right)\left(\overline{N}_i - I\right) + \gamma\sigma I.$$

Now we derive conditions for disease persistence on a globally attracting cycle in periodic environments. In the following result, we prove that the disease lives on a globally attracting cycle when F_1 is a monotone map with no critical points.

THEOREM 11. *If* F_1 *has no critical points in* $[0,1]$ *and* $\mathcal{R}_0 > 1$, *then the composition map*

$$F_{\overline{N}_{p-1}} \circ \cdots \circ F_{\overline{N}_1} \circ F_{\overline{N}_0}$$

has a globally attracting positive fixed point $\overline{I}_0$, *and the uniformly persistent epidemic lives on a globally attracting cycle.*

Proof. By Lemma 6, each $F_{\overline{N}_i}$ is increasing, concave down, and has no critical point on $[0, \overline{N}_i]$. $\mathcal{R}_0 > 1$ is equivalent to $F'_{\overline{N}_i}(0) > 1$. $F_{\overline{N}_i}(0) = 0$ and $F_{\overline{N}_i}(\overline{N}_i) = \gamma\sigma\overline{N}_i < \overline{N}_i$. Thus, each positive initial condition converges under $F_{\overline{N}_i}$ iterations monotonically to the positive fixed point. That is, $F_{\overline{N}_i}$ has a globally attracting positive fixed point on $[0, \overline{N}_i]$.

By Lemma 4,

$$F_{\overline{N}_{p-1}} \circ \cdots \circ F_{\overline{N}_1} \circ F_{\overline{N}_0} : [0, \overline{N}_0] \to [0, \overline{N}_0).$$

Using the chain rule on the composition map $F_{\overline{N}_{p-1}} \circ \cdots \circ F_{\overline{N}_1} \circ F_{\overline{N}_0}$ shows that it is increasing, concave down, and has derivative at the origin larger than 1. So, as in the previous paragraph, $F_{\overline{N}_{p-1}} \circ \cdots \circ F_{\overline{N}_1} \circ F_{\overline{N}_0}$ has a unique globally attracting positive fixed point, $\overline{I}_0$. By Corollary 10, the uniformly persistent epidemic lives on a globally attracting cycle. □

To give a specific example of disease persistence on a globally attracting cycle as predicted by Theorem 11, we assume that infections are modeled as Poisson processes [7, 8, 9]. Then $\phi\left(\alpha \frac{I}{N}\right) = e^{-\alpha \frac{I}{N}}$ and

$$F_{\overline{N}_i}(I) = \gamma \left(1 - e^{-\alpha \frac{I}{\overline{N}_i}}\right)\left(\overline{N}_i - I\right) + \gamma\sigma I. \tag{9}$$

EXAMPLE 12. *In* (9), *set the following parameter values:*

$$\alpha = 2, \quad \gamma = 0.9, \quad \sigma = .9.$$

From the graph of F_1 (see Figure 1) it is clear that F_1 has no critical points in $[0,1]$ and $\mathcal{R}_0 > 1$.

Hence, with these parameters, the composition map

$$F_{\overline{N}_{p-1}} \circ \cdots \circ F_{\overline{N}_1} \circ F_{\overline{N}_0}$$

has a globally attracting positive fixed point, $\overline{I}_0$, and the uniformly persistent epidemic lives on a globally attracting cycle (Theorem 11). Numerical experiments show that this result is also true when $\alpha \in [1, 2.1]$, $\gamma = [0.88, 1)$, and $\sigma = [0.88, 1)$; as well as on other intervals.

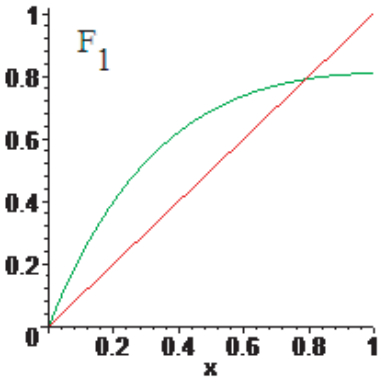


FIG. 1. *Graph of F_1 satisfies the hypotheses of Theorem* 11.

If in Example 12 the recruitment function is the periodic Beverton–Holt model

$$f(t,N(t)) = \frac{(1-\gamma)\mu k_t N(t)}{(1-\gamma)k_t + (\mu - 1 + \gamma)N(t)}$$

with

$$\mu = 2, \quad \gamma = 0.9, \quad \alpha = 2, \quad \sigma = 0.9, \quad p = 2, \quad k_0 = 2, \quad k_1 = 8,$$

then, as predicted by Theorem 11, the total population, susceptible population, and infective population live on the globally attracting 2-cycle

$$\{(8.087, 1.903, 6.184), (7.788, 1.437, 6.351)\}.$$

In the following result, we prove that the disease lives on a globally attracting cycle when F_1 has a critical point with an image (under F_1 iteration) smaller than the critical point.

THEOREM 13. *Let F_1 have a critical point, C_1, in $(0,1)$. If*

$$C_1 > F_1(C_1), \qquad F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}}) < C_{\min\{\overline{N}_i\}},$$

and $\mathcal{R}_0 > 1$, then the composition map

$$F_{\overline{N}_{p-1}} \circ \cdots \circ F_{\overline{N}_1} \circ F_{\overline{N}_0}$$

has a globally attracting positive fixed point $\overline{I}_0$, and the uniformly persistent epidemic lives on a globally attracting cycle.

Proof. By Lemma 6 and our hypothesis, each $F_{\overline{N}_i}$ is increasing, concave down, and has no critical point on $[0, F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}})]$. Since $C_1 > F_1(C_1)$, $I_1 < C_1$ and F_1 is increasing on $[I_1, C_1]$. Consequently, $I_1 < F_1(C_1)$, and by topological conjugacy and Lemma 6, $I_{\overline{N}_i} < F_{\overline{N}_i}(C_{\overline{N}_i}) \le F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}}) < C_{\min\{\overline{N}_i\}}$. Further, $\mathcal{R}_0 > 1$ is equivalent to $F'_{\overline{N}_i}(0) > 1$. Thus, each positive initial condition converges under $F_{\overline{N}_i}$ iterations monotonically to the positive fixed point. That is, $F_{\overline{N}_i}$ has a globally attracting positive fixed point on $[0, F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}})]$.

By the preceding arguments,

$$F_{\overline{N}_{p-1}} \circ \cdots \circ F_{\overline{N}_1} \circ F_{\overline{N}_0} : [0, F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}})] \to [0, F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}})).$$

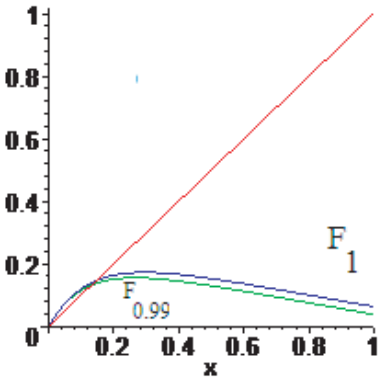


FIG. 2. *Graphs of F_1 and $F_{0.99}$ satisfy the hypotheses of Theorem* 13.

Using the chain rule on the composition map $F_{\overline{N}_{p-1}} \circ \cdots \circ F_{\overline{N}_1} \circ F_{\overline{N}_0}$ shows that it is increasing, concave down, and has derivative at the origin larger than 1. So, as in the previous paragraph, $F_{\overline{N}_{p-1}} \circ \cdots \circ F_{\overline{N}_1} \circ F_{\overline{N}_0}$ has a unique globally attracting positive fixed point on $[0, F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}})]$. Since $F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}})$ is the maximum value of all the $F_{\overline{N}_i}$, every point immediately gets into $[0, F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}})]$, and the composition map

$$F_{\overline{N}_{p-1}} \circ \cdots \circ F_{\overline{N}_1} \circ F_{\overline{N}_0}$$

has a globally attracting positive fixed point, $\overline{I}_0$. By Corollary 10, the uniformly persistent epidemic lives on a globally attracting cycle. □

Next we demonstrate, via a specific example, disease persistence on a globally attracting cycle, as predicted by Theorem 13.

EXAMPLE 14. *In* (9)*, set the following parameter values:*

$$\alpha = 7, \quad \gamma = 0.25, \quad \sigma = 0.25, \quad \max\{\overline{N}_i\} = 1, \quad \min\{\overline{N}_i\} = .99.$$

From the graphs of $F_{\max\{\overline{N}_i\}}$ and $F_{\min\{\overline{N}_i\}}$ (see Figure 2*) it is clear that $C_1 > F_1(C_1)$,*

$$F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}}) < C_{\min\{\overline{N}_i\}},$$

and $\mathcal{R}_0 > 1$.

Hence, with these parameters, the composition map

$$F_{\overline{N}_{p-1}} \circ \cdots \circ F_{\overline{N}_1} \circ F_{\overline{N}_0}$$

has a globally attracting positive fixed point $\overline{I}_0$, and the uniformly persistent epidemic lives on a globally attracting cycle. Numerical experiments show that this result is also true when $\alpha \in [7, 10]$, $\gamma = [0.15, 0.25]$, $\sigma = [0.15, 0.25]$, and $\overline{N}_J \in [0.9, 1]$; as well as on other intervals.

If in Example 14 the recruitment function is the periodic Beverton–Holt model

$$f(t, N(t)) = \frac{(1-\gamma)\mu k_t N(t)}{(1-\gamma)k_t + (\mu - 1 + \gamma)N(t)}$$

with

$$\mu = 2, \quad \gamma = 0.25, \quad \alpha = 7, \quad \sigma = 0.25, \quad p = 2, \quad k_0 = 0.665, \quad k_1 = 0.965,$$

then

$$\max\{\overline{N}_i\} = 0.995, \qquad \min\{\overline{N}_i\} = 0.905,$$

and, as predicted by Theorem 13, the total population, susceptible population, and infective population live on the globally attracting 2-cycle

$$\{(0.995, 0.860, 0.135), (0.905, 0.765, 0.140)\}.$$

Next, we prove that the disease lives on a globally attracting cycle when F_1 has a critical point with an image (under F_1 iteration) bigger than the critical point.

THEOREM 15. *Let F_1 have a critical point, C_1, in $(0,1)$. If*

$$C_1 < F_1(C_1), \qquad F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}}) < \min\{\overline{N}_i\},$$

$$C_{\max\{\overline{N}_i\}} < F_{\min\{\overline{N}_i\}} \circ F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}}),$$

and $\mathcal{R}_0 > 1$, then the composition map

$$F_{\overline{N}_{p-1}} \circ \cdots \circ F_{\overline{N}_1} \circ F_{\overline{N}_0}$$

has a globally attracting positive fixed point $\overline{I}_0$, and the uniformly persistent epidemic lives on a globally attracting cycle.

Proof. By Lemma 6 and our hypothesis, each $F_{\overline{N}_i}$ is decreasing on $[C_{\overline{N}_i}, \overline{N}_i] \supseteq [C_{\max\{\overline{N}_i\}}, F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}})]$ and $F_{\overline{N}_i}(C_{\max\{\overline{N}_i\}}) \le F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}})$. By our hypothesis, $C_{\max\{\overline{N}_i\}} < F_{\min\{\overline{N}_i\}} \circ F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}}) \le F_{\overline{N}_i} \circ F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}})$. Thus, each $F_{\overline{N}_i}$ is decreasing on $[C_{\max\{\overline{N}_i\}}, F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}})]$ and sends this interval into itself. Consequently, each $F_{\overline{N}_i}$ has a fixed point $I_{\overline{N}_i}$ in this interval. Since $F'_{\overline{N}_i}(I) \in (-1, 0]$ for all $I \in [C_{\max\{\overline{N}_i\}}, F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}})]$ and $\overline{N}_i \in \{\overline{N}_0, \overline{N}_1, \ldots, \overline{N}_{p-1}\}$ (Lemma 6), each $F_{\overline{N}_i}$ is a contraction on this interval. This implies that $F_{\overline{N}_{p-1}} \circ \cdots \circ F_{\overline{N}_1} \circ F_{\overline{N}_0}$ is a contraction with a unique fixed point, which is $\overline{I}_0$.

By Lemma 6 and our hypothesis, $F_{\overline{N}_i}(I) < F_{\overline{N}_i}(C_{\overline{N}_i}) \le F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}}) < \min\{\overline{N}_i\}$ for all $I \in [0, \overline{N}_i]$ and $\overline{N}_i \in \{\overline{N}_0, \overline{N}_1, \ldots, \overline{N}_{p-1}\}$. $F_{\overline{N}_i}(I) > I$ for all $I \in (0, I_{\overline{N}_i})$. Thus, all positive points below $C_{\max\{\overline{N}_i\}}$ increase until they are in $[C_{\max\{\overline{N}_i\}}, F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}})]$. Consequently, $F_{\overline{N}_{p-1}} \circ \cdots \circ F_{\overline{N}_1} \circ F_{\overline{N}_0}$ has a globally attracting positive fixed point, $\overline{I}_0$, and by Corollary 10, the uniformly persistent epidemic lives on a globally attracting cycle. ∎

Now we demonstrate, via a specific example, disease persistence on a globally attracting cycle, as predicted by Theorem 15.

EXAMPLE 16. *In (9) set the following parameter values:*

$$\alpha = 20, \quad \gamma = 0.5, \quad \sigma = .5, \quad \max\{\overline{N}_i\} = 1, \quad \min\{\overline{N}_i\} = .7.$$

From the graphs of $F_{\max\{\overline{N}_i\}}$ and $F_{\min\{\overline{N}_i\}}$ (see Figure 3) it is clear that

$$C_1 < F_1(C_1), \qquad F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}}) < \min\{\overline{N}_i\},$$

$$C_{\max\{\overline{N}_i\}} < F_{\min\{\overline{N}_i\}} \circ F_{\max\{\overline{N}_i\}}(C_{\max\{\overline{N}_i\}}),$$

and $\mathcal{R}_0 > 1$.

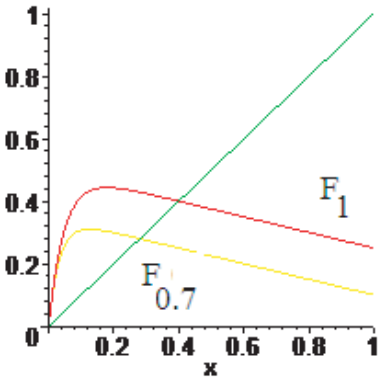


FIG. 3. *Graphs of F_1 and $F_{0.7}$ satisfy the hypotheses of Theorem* 15.

Hence, with these parameters, the composition map

$$F_{\overline{N}_{p-1}} \circ \cdots \circ F_{\overline{N}_1} \circ F_{\overline{N}_0}$$

has a globally attracting positive fixed point $\overline{I}_0$, and the uniformly persistent epidemic lives on a globally attracting cycle. Numerical experiments show that this result is also true when $\alpha \in [15, 25]$, $\gamma = [0.45, 0.6)$, $\sigma = [0.45, 0.6)$, and $\overline{N}_J \in [0.7, 1]$; as well as on other intervals.

If in Example 16 the recruitment function is the periodic Beverton–Holt model

$$f(t, N(t)) = \frac{(1-\gamma)\mu k_t N(t)}{(1-\gamma)k_t + (\mu - 1 + \gamma)N(t)}$$

with

$$\mu = 2, \quad \alpha = 20, \quad \gamma = 0.5, \quad \sigma = .5, \quad p = 2, \quad k_0 = 0.1, \quad k_1 = 0.9,$$

then

$$\max\{\overline{N}_i\} = 0.917, \qquad \min\{\overline{N}_i\} = 0.741,$$

and, as predicted by Theorem 15, the total population, susceptible population, and infective population live on the globally attracting 2-cycle

$$\{(0.917, 0.644, 0.273), (0.741, 0.352, 0.389)\}.$$

In all the above examples, we use the periodic Beverton–Holt model as the recruitment function to highlight uniform persistence via attracting cycles. Similar examples can be obtained using the periodic constant recruitment function.

6. Uniform persistence and geometric demographics. When new recruits arrive at the periodic positive per-capita growth rate λ_t, the demographic long-term dynamics is determined by the demographic basic reproductive number $\mathcal{R}_d$ (see (3)). In this case, we use proportions to study the epidemic process. We introduce the new variables

$$s(t) = \frac{S(t)}{N(t)}$$

and

$$i(t) = \frac{I(t)}{N(t)}.$$

In the new variables, when $f(t, N) = \lambda_t N$, then

$$N(t+1) = (\lambda_t + \gamma)\, N(t),$$

and model (4) becomes

$$\left.\begin{aligned} s(t+1) &= \tfrac{\lambda_t}{\lambda_t+\gamma} + \tfrac{\gamma\phi(\alpha i(t))}{\lambda_t+\gamma} s(t) + \tfrac{\gamma(1-\sigma)}{\lambda_t+\gamma} i(t), \\ i(t+1) &= \tfrac{\gamma(1-\phi(\alpha i(t)))}{\lambda_t+\gamma} s(t) + \tfrac{\gamma\sigma}{\lambda_t+\gamma} i(t). \end{aligned}\right\} \tag{10}$$

Since $i(t)+s(t) = 1$ for all t, the substitution $s(t) = 1-i(t)$ reduces the i-equation of the system to the one-dimensional nonautonomous equation

$$i(t+1) = \frac{\gamma\,(1-\phi\,(\alpha i(t)))}{\lambda_t+\gamma}\,(1-i(t)) + \frac{\gamma\sigma}{\lambda_t+\gamma} i(t).$$

Let

$$\widetilde{F}_\lambda(i) = \frac{\gamma\,(1-\phi\,(\alpha i))}{\lambda+\gamma}\,(1-i) + \frac{\gamma\sigma}{\lambda+\gamma} i.$$

Since $i \le 1$,

$$\widetilde{F}_\lambda(i) < 1 \quad \text{and} \quad \widetilde{F}_\lambda(i) = \frac{1}{\lambda+\gamma} F_1(i).$$

By Lemma 6,

$$\widetilde{F}'_\lambda(0) = \frac{-\alpha\gamma\phi'\,(0) + \gamma\sigma}{\lambda+\gamma}$$

and

$$\left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)'(0) = \frac{(-\alpha\gamma\phi'\,(0) + \gamma\sigma)^p}{\prod_{t=0}^{p-1}(\lambda_t+\gamma)}.$$

Let

$$\mathcal{R}_0 = \frac{-\alpha\gamma\phi'\,(0)}{(\mathcal{R}_d(1-\gamma^p)+\gamma^p)^{\frac{1}{p}} - \gamma\sigma}.$$

If $\mathcal{R}_d = 1$, the total population is bounded and uniformly persistent, and $\mathcal{R}_0$ reduces to $\frac{-\alpha\gamma\phi'(0)}{1-\gamma\sigma}$, which is (5). We will prove that $\mathcal{R}_0 > 1$ implies that $i(t)$ persists, and $\mathcal{R}_0 < 1$ implies $\lim_{t\to\infty} i(t) = 0$. First, we obtain local stability results when $\mathcal{R}_0 \neq 1$.

LEMMA 17. *$\mathcal{R}_0 > 1$ is equivalent to $\left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)'(0) > 1$, and $\mathcal{R}_0 < 1$ is equivalent to $\left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)'(0) < 1$.*

Proof. Assume $\mathcal{R}_0 > 1$; then

$$\frac{-\alpha\gamma\phi'\,(0)}{(\mathcal{R}_d(1-\gamma^p)+\gamma^p)^{\frac{1}{p}} - \gamma\sigma} > 1.$$

Since $-\alpha\gamma\phi'(0) > 0$, $(\mathcal{R}_d(1-\gamma^p)+\gamma^p)^{\frac{1}{p}} - \gamma\sigma > 0$ and $-\alpha\gamma\phi'(0) > (\mathcal{R}_d(1-\gamma^p)+\gamma^p)^{\frac{1}{p}} - \gamma\sigma$. This implies that

$$(-\alpha\gamma\phi'(0)+\gamma\sigma)^p > \mathcal{R}_d(1-\gamma^p)+\gamma^p = \frac{\prod_{J=0}^{p-1}(\lambda_J+\gamma)-\gamma^p}{1-\gamma^p}(1-\gamma^p)+\gamma^p$$
$$= \prod_{J=0}^{p-1}(\lambda_J+\gamma).$$

Hence,

$$\left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)'(0) = \frac{(-\alpha\gamma\phi'(0)+\gamma\sigma)^p}{\prod_{J=0}^{p-1}(\lambda_J+\gamma)} > 1.$$

Since all the steps are reversible, $\mathcal{R}_0 > 1$ is equivalent to $\left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)'(0) > 1$.

To prove the other inequality, note that $(\mathcal{R}_d(1-\gamma^p)+\gamma^p)^{\frac{1}{p}} - \gamma\sigma > (\gamma^p)^{\frac{1}{p}} - \gamma\sigma > 0$, and proceed as in the proof of the last inequality. □

THEOREM 18. *Let*

$$f(t,N) = \lambda_t N$$

in model (2)*, where* $\lambda_{t+p} = \lambda_t$.

(a) *If* $\mathcal{R}_0 < 1$*, then in model* (10)*,* $\lim_{t\to\infty} i(t) = 0$*. That is, the proportion of infectives in the total population goes extinct.*

(b) *If* $\mathcal{R}_0 > 1$*, then in model* (10)*,* $\exists\ \eta > 0$ *satisfying* $\lim_{t\to\infty} \inf i(t) \geq \eta$*. That is, the proportion of infectives in the total population uniformly persists.*

Proof. (a) Lemma 17 shows that $\mathcal{R}_0 < 1$ implies

$$\left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)'(0) = \frac{(-\alpha\gamma\phi'(0)+\gamma\sigma)^p}{\prod_{J=0}^{p-1}(\lambda_J+\gamma)} < 1.$$

Let $i > 0$; then $\widetilde{F}_\lambda(i) = \frac{1}{\lambda+\gamma}F_1(i) < \frac{1}{\lambda+\gamma}F_1'(0)i = \frac{-\alpha\gamma\phi'(0)+\gamma\sigma}{\lambda+\gamma}i$, by Lemma 6. Using this p times gives

$$i(p) = \left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)(i(0)) < \frac{(-\alpha\gamma\phi'(0)+\gamma\sigma)^p}{\prod_{J=0}^{p-1}(\lambda_J+\gamma)} i(0) < i(0).$$

Thus, the sequence $\{i(tp)\}$ is dominated by the geometrically decreasing sequence

$$\left\{\left(\frac{(-\alpha\gamma\phi'(0)+\gamma\sigma)^p}{\prod_{J=0}^{p-1}(\lambda_J+\gamma)}\right)^t i(0)\right\}$$

and hence, by continuity of the system,

$$\lim_{t\to\infty} i(t) = 0.$$

(b) Lemma 17 shows that $\mathcal{R}_0 > 1$ implies

$$\left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)'(0) > 1.$$

Since $\left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)(0) = 0$, the graph of $\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$ starts out above the diagonal. Since for each λ_J and each $i \in [0,1]$, $\widetilde{F}_{\lambda_J}(i) < 1$, $\left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)(1) < 1$. Thus $\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$ has at least one positive fixed point. Let i^* be the minimum of these positive fixed points; then $i((t+1)p) > i(tp)$ when $i(tp)$ is in the open interval $(0, i^*)$.

Let

$$m = \min\left\{\left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)(i) : i \in [i^*, 1]\right\}.$$

Note that $m > 0$. If $i(tp) \in (0, m)$, then $i((t+1)p) > i(tp)$, and the sequence $\{i(tp)\}$ continues to increase until the value is at least m. But then the sequence can never jump lower than m. Hence,

$$\lim_{t\to\infty} \inf i(tp) \geq m.$$

Now each of the maps $\widetilde{F}_{\lambda_0}, \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}, \ldots, \widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$ has positive minima on $[m, 1]$, and $\lim_{t\to\infty} \inf i(t)$ is at least the minimum of these minima. Hence, $\exists\ \eta > 0$ satisfying

$$\lim_{t\to\infty} \inf i(t) > \eta. \qquad \square$$

7. Cyclic attractors and geometric demographics. We assume that the total population is growing geometrically. That is, the recruitment function in the p-periodic demographic equation is $f(t, N(t)) = \lambda_t N(t)$. If, in addition,

$$\mathcal{R}_0 = \frac{-\alpha\gamma\phi'(0)}{(\mathcal{R}_d(1-\gamma^p) + \gamma^p)^{\frac{1}{p}} - \gamma\sigma} > 1,$$

we show that it is possible for the persistent i-population to live on a globally attracting cycle. This implies that both the *i-dynamics* under periodic geometric recruitment function and the *I-dynamics* under either periodic constant or periodic Beverton–Holt recruitment functions are capable of living on globally attracting cycles.

Next, we prove that the proportion of infectives live on a globally attracting cycle when $\widetilde{F}_{\lambda_t}$ is a monotone map with no critical points.

THEOREM 19. *If each $\widetilde{F}_{\lambda_t}$ has no critical points in $[0,1]$ and $\mathcal{R}_0 > 1$, then the composition map*

$$\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$$

has a globally attracting positive fixed point $\bar{i}_0$, and the uniformly persistent proportion of infectives in the total population lives on a globally attracting cycle.

Proof. By Lemma 6 each $\widetilde{F}_{\lambda_t}(i) = \frac{1}{\lambda_t+\gamma} F_1(i)$ is increasing, concave down, and has no critical point on $[0,1]$. Hence, $\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$ is also increasing, concave down, and has no critical point on $[0,1]$. $\mathcal{R}_0 > 1$ is equivalent to $\left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)'(0) > 1$. Since

$$\left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)(0) = 0,$$

$\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$ has a fixed point in $(0,1)$. Thus, each positive initial condition converges monotonically under iteration of $\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$ to the positive fixed

point. That is, $\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$ has a globally attracting positive fixed point on $[0,1]$, and the uniformly persistent proportion of infectives in the total population lives on a globally attracting cycle. □

To give a specific example of disease persistence on a globally attracting cycle as predicted by Theorem 19, following Example 12, we assume that infections are modeled as Poisson processes [7, 8, 9]. Then $\phi(\alpha i) = e^{-\alpha i}$, $\phi'(0) = -1$, and

$$\widetilde{F}_\lambda(i) = \frac{\gamma\left(1 - e^{-\alpha i}\right)}{\lambda + \gamma}(1 - i) + \frac{\gamma\sigma}{\lambda + \gamma} i. \tag{11}$$

EXAMPLE 20. *In* (11)*, set the following parameters:*

$$\alpha = 2, \quad \gamma = 0.9, \quad \sigma = 0.9, \quad \lambda_0, \lambda_1 \in [0.1, 0.75].$$

As in Example 12, $\widetilde{F}_\lambda$ has no critical point in $[0,1]$. For the special case $\alpha = 2$, $\gamma = 0.9$, $\sigma = 0.9$, $\lambda_0 = 0.5$, and $\lambda_1 = 0.6$ the proportion of infectives in the total population lives on the stable period 2 orbit $\{0.447, 0.469\}$ (see Theorem 19).

Now, we prove that the disease lives on a globally attracting cycle when F_1 has a critical point with an image (under $\widetilde{F}_{\max\{\lambda_t\}}$ iteration) smaller than the critical point.

THEOREM 21. *Let each F_λ have a critical point, C_1, in $(0,1)$. If*

$$\widetilde{F}_{\min\{\lambda_t\}}(C_1) < C_1$$

and $\mathcal{R}_0 > 1$, then the composition map

$$\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$$

has a globally attracting positive fixed point $\bar{i}_0$, and the uniformly persistent proportion of infectives in the total population lives on a globally attracting cycle.

Proof. By Lemma 6 and our hypothesis, each $\widetilde{F}_{\lambda_t}(i) = \frac{1}{\lambda_t + \gamma} F_1(i)$ has C_1 as its only critical point on $[0,1]$, is increasing on $[0, C_1]$, and concave down on $[0,1]$. We also have

$$\widetilde{F}_{\max\{\lambda_t\}}(i) \le \widetilde{F}_{\lambda_t}(i) \le \widetilde{F}_{\min\{\lambda_t\}}(i).$$

Since $\widetilde{F}_{\min\{\lambda_t\}}(C_1) < C_1$, the image of each $\widetilde{F}_{\lambda_t}$ is in $[0, C_1)$. Thus, $\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$ has C_1 as its only critical point, and its image is in $[0, C_1)$. Hence, $\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$ is increasing and concave down on $[0, C_1]$.

$\mathcal{R}_0 > 1$ is equivalent to $\left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)'(0) > 1$. Since $\left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)(0) = 0$, $\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$ has a fixed point in $(0, C_1)$. Thus, each positive initial condition gets into $[0, C_1]$ and converges monotonically under iteration of $\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$ to the positive fixed point. That is, $\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$ has a globally attracting positive fixed point on $(0,1]$, and the uniformly persistent proportion of infectives in the total population lives on a globally attracting cycle. □

To give a specific example of disease persistence on a globally attracting cycle as predicted by Theorem 21, following Example 14, we assume that infections are modeled as Poisson processes [7, 8, 9].

EXAMPLE 22. *In* (11)*, set the following parameter values:*

$$\alpha \in [7, 10], \quad \gamma \in [0.15, 0.25], \quad \sigma \in [0.15, 0.25], \quad \lambda_0, \lambda_1 \in [0.85, 1.0].$$

As in Example 14, all the conditions of Theorem 21 are satisfied. For the special case $\alpha = 7$, $\gamma = 0.25$, $\sigma = 0.25$, $\lambda_0 = 0.85$, and $\lambda_1 = 0.95$ the proportion of infectives in the total population lives on the stable period 2 orbit $\{0.107, 0.113\}$ (see Theorem 21).

Next, we prove that the disease lives on a globally attracting cycle when F_1 has a critical point with an image (under $\widetilde{F}_{\max\{\lambda_t\}}$ iteration) bigger than the critical point.

THEOREM 23. *Let F_1 have a critical point, C_1, in $(0,1)$. If*

$$C_1 < \widetilde{F}_{\max\{\lambda_t\}} \circ \widetilde{F}_{\min\{\lambda_t\}}(C_1),$$

then $\mathcal{R}_0 > 1$, and the composition map

$$\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$$

has a globally attracting positive fixed point $\bar{i}_0$, and the uniformly persistent proportion of infectives in the total population lives on a globally attracting cycle.

Proof. By Lemma 6 and our hypothesis, each $\widetilde{F}_{\lambda_t}(i) = \frac{1}{\lambda_t+\gamma}F_1(i)$ has C_1 as its only critical point on $[0,1]$, is increasing on $[0, C_1]$, and concave down on $[0,1]$. We also have

$$\widetilde{F}_{\max\{\lambda_t\}}(i) \le \widetilde{F}_{\lambda_t}(i) \le \widetilde{F}_{\min\{\lambda_t\}}(i).$$

Since $C_1 < \widetilde{F}_{\max\{\lambda_t\}} \circ \widetilde{F}_{\min\{\lambda_t\}}(C_1) < \widetilde{F}_{\max\{\lambda_t\}}(C_1) \le \widetilde{F}_{\lambda_t}(C_1)$ and $\widetilde{F}_{\lambda_t}(0) = 0$, each $\widetilde{F}'_{\lambda_t}(0) > 1$. Hence, $\left(\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}\right)'(0) > 1$ and, by Lemma 17, $\mathcal{R}_0 > 1$.

By our hypothesis, $C_1 < \widetilde{F}_{\max\{\lambda_t\}} \circ \widetilde{F}_{\min\{\lambda_t\}}(C_1) \le \widetilde{F}_{\lambda_t} \circ \widetilde{F}_{\min\{\lambda_t\}}(C_1)$ and $\widetilde{F}_{\lambda_t}(C_1) \le \widetilde{F}_{\min\{\lambda_t\}}(C_1)$ for each λ_t. Thus, each $\widetilde{F}_{\lambda_t}$ is decreasing on $[C_1, \widetilde{F}_{\min\{\lambda_t\}}(C_1)]$ and sends this interval into itself. Consequently, each $\widetilde{F}_{\lambda_t}$ has a fixed point $\bar{i}_{\lambda_t}$ in this interval:

$$\begin{aligned}\widetilde{F}'_{\lambda_i}(i) &= \frac{1}{\lambda_i+\gamma}F'_1(i) = \frac{1}{\lambda_i+\gamma}\left(-\alpha\gamma\phi'(\alpha i)(1-i) - \gamma(1-\phi(\alpha i)) + \gamma\sigma\right)\\ &> \frac{-\gamma}{\lambda_i+\gamma} > -1.\end{aligned}$$

Hence, $\widetilde{F}'_{\lambda_i}(i) \in (-1, 0]$ for all $i \in [C_1, \widetilde{F}_{\min\{\lambda_t\}}(C_1)]$ and all λ_i. Each $\widetilde{F}_{\lambda_i}$ is a contraction on this interval. This implies that $\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$ is a contraction with a unique fixed point in $[C_1, \widetilde{F}_{\min\{\lambda_t\}}(C_1)]$. Thus, all positive points below C_1 increase until they are in $[C_1, \widetilde{F}_{\min\{\lambda_t\}}(C_1)]$. Consequently, $\widetilde{F}_{\lambda_{p-1}} \circ \cdots \circ \widetilde{F}_{\lambda_1} \circ \widetilde{F}_{\lambda_0}$ has a globally attracting positive fixed point on $(0,1]$, and the uniformly persistent proportion of infectives in the total population lives on a globally attracting cycle. □

To give a specific example of disease persistence on a globally attracting cycle as predicted by Theorem 23, following Example 16, we assume that infections are modeled as Poisson processes [7, 8, 9].

EXAMPLE 24. *In (11), set the following parameter values:*

$$\alpha \in [15, 25], \quad \gamma \in [0.45, 0.6], \quad \sigma \in [0.45, 0.6], \quad \lambda_0, \lambda_1 \in [0.7, 1.0].$$

As in Example 16, all the conditions of Theorem 23 are satisfied. For the special case $\alpha = 20$, $\gamma = 0.5$, $\sigma = 0.5$, $\lambda_0 = 0.8$, and $\lambda_1 = 0.9$ the proportion of infectives in the total population lives on the stable period 2 orbit $\{0.289, 0.327\}$ (see Theorem 23).

FIG. 4. *Two attracting 4-cycles (multiple attractors). Red and black attractors.*

8. Multiple attractors. In constant environments, single patch discrete-time epidemic models typically support only one attractor [7, 8, 9]. Henson and coworkers [30, 31, 32, 33], Franke and Selgrade [22], and Franke and Yakubu [24] found multiple attractors in periodically forced models, where the corresponding models in constant environments have no multiple attractors. These multiple attractors are a result of periodic perturbations of the corresponding models in constant environments. In this section, we illustrate that our periodically forced discrete-time single patch epidemic model can generate multiple (coexisting) attractors. In this situation, the long-term disease dynamics depends on initial conditions.

Periodicity is not the only mechanism for generating multiple attractors. Migration and age-structure are known to induce multiple attractors in population models [5, 7, 9, 28, 29, 48, 50]. Also, epidemic models with "backward" bifurcations support multiple attractors [27, 47].

EXAMPLE 25. *Consider model* (7) *with 4-periodic constant recruitment function*

$$f(t,N) = k_t(1-\gamma)$$

and

$$\phi\left(\frac{\alpha I}{N}\right) = e^{-\frac{\alpha I}{N}},$$

where

$$\alpha = 250, \quad \gamma = 0.4, \quad \sigma = 0.02, \quad k_0 = 1, \quad k_1 = 200, \quad k_2 = 1, \quad k_3 = 210.$$

Example 25 has two coexisting 4-cycle attractors, a "red" attractor at

$$\{(60.32, 52.44), (144.13, 3.57), (58.25, 56.14), (149.30, 1.29)\}$$

and a "black" attractor at

$$\{(60.32, 58.19), (144.13, 1.32), (58.25, 51.32), (149.30, 3.18)\}.$$

In this example, the total population is on a globally attracting 4-cycle, while the infective population is on multiple 4-cycle attractors. That is, the disease dynamics has multiple outcomes, while the total population has a single long-term dynamics. Figure 4 displays the two attracting 4-cycles.

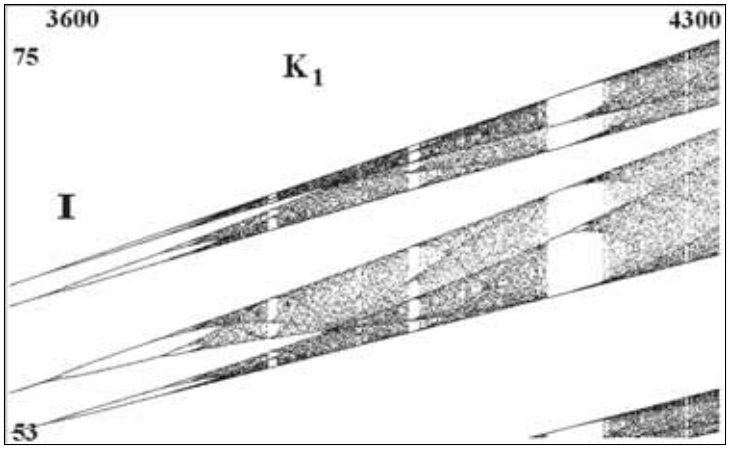


FIG. 5. *As k_1 varies between* 3600 *and* 4300*, the infective population in Example* 26 *undergoes period-doubling bifurcations route to chaos.*

9. Nonchaotic demographic dynamics generates chaotic disease dynamics. In constant environments, the demographic dynamics is capable of driving the disease dynamics [1, 2, 7, 8, 9]. That is, when the total population (in the absence of the disease) is on a cycle of period k, the population of infectives (in the presence of the disease) is also on a cycle of the same period k, albeit the amplitude of the total population is much larger than that of the infected population.

Our demographic equation with periodic constant or periodic Beverton–Holt or periodic geometric recruitment function can have either asymptotically bounded growth via globally attracting cycles (constant or Beverton–Holt models) or geometric growth (geometric model). In this section, we use numerical simulations to illustrate that the periodic epidemic model (4) can generate chaotic attractors where the periodic recruitment function is periodic constant or the periodic Beverton–Holt or periodic geometric function [39, 40, 41, 42, 43, 44, 45]. That is, in periodic environments, the demographic dynamics does not always drive the disease dynamics. We illustrate these cases in the following three examples.

EXAMPLE 26. *Consider model* (7) *with* 2*-periodic constant recruitment function*

$$f(t,N) = k_t(1-\gamma),$$
$$\phi\left(\frac{\alpha I}{N}\right) = e^{-\frac{\alpha I}{N}},$$

and

$$\alpha = 250, \quad \gamma = 0.44, \quad \sigma = 0.002, \quad k_0 = 1, \quad 3600 \le k_1 \le 4300.$$

Figure 5 shows parameter regimes of chaotic dynamics in the infective population of Example 26, where the total population is on a cyclic (nonchaotic) attractor. In this example, the recruitment function is a 2-periodic constant function. Next, we use numerical simulations to illustrate chaotic dynamics in the infective population where the recruitment function is the periodic Beverton–Holt model.

EXAMPLE 27. *Consider model* (7) *with* 2*-periodic geometric growth model*

$$f(t,N) = \frac{(1-\gamma)\mu k_t N(t)}{(1-\gamma)k_t + (\mu - 1 + \gamma)N(t)},$$

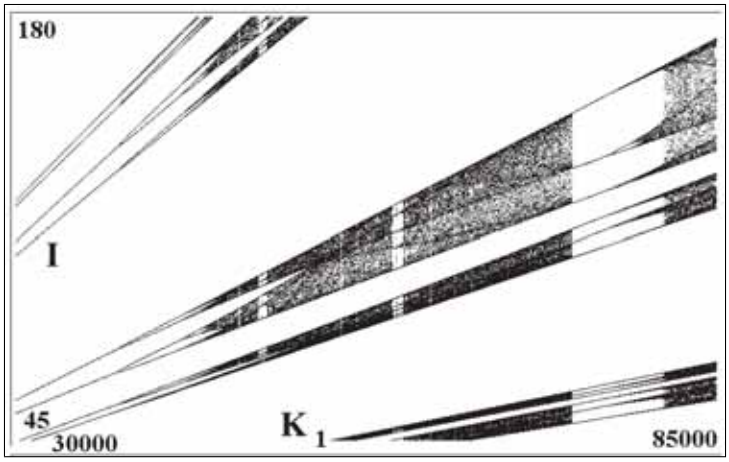


FIG. 6. *As k_1 varies between* 30000 *and* 85000*, the infective population in Example* 27 *undergoes period-doubling bifurcations route to chaos.*

$$\phi\left(\frac{\alpha I}{N}\right) = e^{-\frac{\alpha I}{N}},$$

and

$$\alpha = 250, \quad \gamma = 0.44, \quad \sigma = 0.002, \quad \mu = 2, \quad k_0 = 1, \quad 30000 \leq k_1 \leq 85000.$$

As we vary k_1 between 30000 and 85000, Figure 6 shows the infective population undergoing period-doubling bifurcations route to chaos. As in Figure 5, Figure 6 shows parameter regimes of chaotic dynamics in the infective population of Example 27, where the total population is governed by the 2-periodic Beverton–Holt model (nonchaotic dynamics). Next, we use numerical simulations to illustrate chaotic dynamics in the infective population where the recruitment function is the periodic geometric growth model.

EXAMPLE 28. *Consider model* (7) *with* 2*-periodic geometric growth model*

$$f(t, N) = \lambda_t N$$

and

$$\phi\left(\frac{\alpha I}{N}\right) = e^{-\frac{\alpha I}{N}},$$

where

$$\alpha = 250, \quad \gamma = 0.44, \quad \sigma = 0.002, \quad \lambda_0 = 0.0004, \quad 0.3 \leq \lambda_1 \leq 1.5.$$

As λ_1 varies between 0.3 and 1.5, the infective population in Example 28 undergoes period-doubling bifurcations route to chaos. As in Figures 5 and 6, Figure 7 shows parameter regimes of chaotic dynamics in the infective population of Example 28, where the total population is under geometric (nonchaotic) growth.

In periodic environments, Examples 26, 27, and 28 show that demographics dynamics does not always drive disease dynamics. In particular, they illustrate chaotic disease dynamics in the absence of chaotic dynamics in the demographic equation. These examples have only highlighted some of the complex interactions between disease and demographics dynamics in periodic environments.

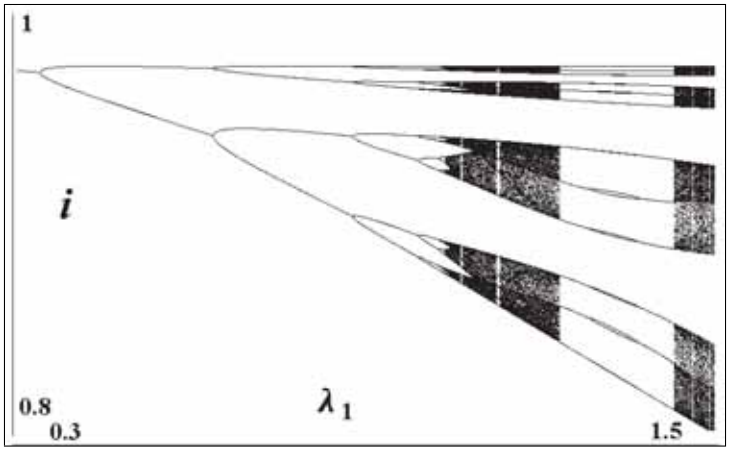


FIG. 7. *As* λ_1 *varies between* 0.3 *and* 1.5*, the infective population in Example* 28 *undergoes period-doubling bifurcations route to chaos.*

10. Conclusion. The study of the combined effects of seasonal trends and diseases on the extinction and persistence of discretely reproducing populations has received little attention. The focus has been on the impact of diseases on populations in constant (nonseasonal) environments [1, 2, 3, 7, 8, 9]. Most species live in seasonal environments, and the neglect of seasonal factors is apt to lead to a misunderstanding of how the population is interacting with its environment [26]. In this paper, we focus on the joint impact of periodic environments and disease epidemics on life-history outcomes of discretely reproducing populations. We formulated and analyzed a periodically forced discrete-time SIS epidemic model via the epidemic threshold parameter $\mathcal{R}_0$. We also investigated the relationship between the predisease invasion population dynamics and the diseases dynamics.

Fixed point (nonoscillatory) dynamics are rare in periodic environments. We use the periodic Beverton–Holt, the periodic constant, and the periodic Malthus (geometric growth) models as recruitment functions to highlight disease (uniform) persistence on globally attracting cycles whenever $\mathcal{R}_0 > 1$. The disease persists on fixed point attractors in the corresponding autonomous epidemic models [7, 8, 9].

In constant environments, Castillo-Chavez and Yakubu, in an earlier work, showed that the SIS discrete-time epidemic model supports only one attractor [7, 8, 9]. That is, the long-term epidemic dynamics is independent of initial population sizes. It is known that periodically forced (nonautonomous) population models without explicit disease dynamics are capable of generating multiple attractors via cusp bifurcations, where the corresponding autonomous models do not have multiple attractors [24]. In periodic environments, we use numerical simulations to show that the SIS model supports multiple attractors. That is, in periodic environments, the ultimate disease dynamics depends on initial population sizes. Seasonality is not the only mechanism for generating multiple attractors. Dispersal and age-structure are other factors that lead to the creation of multiple attractors in constant environments.

Castillo-Chavez and Yakubu, in [7, 8, 9], used the autonomous SIS discrete-time epidemic model to answer the following questions. Will the infective population survive? And if it does, will it settle on a particular attractor? What is the relationship between the population and epidemic attractors? Castillo-Chavez and Yakubu showed that in constant environments, infectives can survive on cyclic attractors. The period

of the (predisease invasion) population attractor is the same as the period of the infective population. In this paper, we show that it is possible for the disease dynamics to be chaotic, where the (predisease invasion) population is cyclic and nonchaotic. That is, with the advent of seasonality the demographic dynamics does not always drive the disease dynamics.

Acknowledgment. We thank the referees for useful comments and suggestions.

REFERENCES

[1] L. J. S. Allen and A. M. Burgin, *Comparison of deterministic and stochastic SIS and SIR models in discrete-time*, Math. Biosci., 163 (2000), pp. 1–33.
[2] L. J. S. Allen, *Some discrete-time SI, SIR and SIS epidemic models*, Math. Biosci., 124 (1994), pp. 83–105.
[3] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, Oxford, UK, 1992.
[4] N. T. J. Bailey, *The Mathematical Theory of Infectious Diseases and Its Applications*, Griffin, London, 1975.
[5] M. Begon, J. L. Harper, and C. R. Townsend, *Ecology: Individuals, Populations and Communities*, Blackwell Science, Williston, VT, 1996.
[6] R. J. H. Beverton and S. J. Holt, *On the Dynamics of Exploited Fish Populations*, Fish. Invest. Ser. II, H. M. Stationery Office, London, 1957.
[7] C. Castillo-Chavez and A. Yakubu, *Dispersal, disease and life-history evolution*, Math. Biosci., 173 (2001), pp. 35–53.
[8] C. Castillo-Chavez and A. Yakubu, *Discrete-time S-I-S models with complex dynamics*, Nonlinear Anal., 47 (2001), pp. 4753–4762.
[9] C. Castillo-Chavez and A. A. Yakubu, *Intraspecific competition, dispersal and disease dynamics in discrete-time patchy environments*, in Mathematical Approaches for Emerging and Reemerging Infectious Diseases: An Introduction to Models, Methods and Theory, C. Castillo-Chavez with S. Blower, P. van den Driessche, D. Kirschner, and A.-A. Yakubu, eds., Springer-Verlag, New York, 2002, pp. 165–181.
[10] B. D. Coleman, *On the growth of populations with narrow spread in reproductive age.* I. *General theory and examples*, J. Math. Biol., 6 (1978), pp. 1–19.
[11] C. S. Coleman and J. C. Frauenthal, *Satiable egg eating predators*, Math. Biosci., 63 (1983), pp. 99–119.
[12] K. L. Cooke and J. A. Yorke, *Some equations modelling growth processes of gonorrhea epidemics*, Math. Biosci., 58 (1973), pp. 93–109.
[13] R. F. Costantino, J. M. Cushing, B. Dennis, R. A. Desharnais, and S. M. Henson, *Resonant population cycles in temporarily fluctuating habitats*, Bull. Math. Biol., 60 (1998), pp. 247–273.
[14] J. M. Cushing and S. M. Henson, *Global dynamics of some periodically forced, monotone difference equations*, J. Differential Equations Appl., 7 (2001), pp. 859–872.
[15] S. N. Elaydi, *Discrete Chaos*, Chapman & Hall/CRC, Boca Raton, FL, 2000.
[16] S. N. Elaydi, *Periodicity and stability of linear Volterra difference equations*, J. Math. Anal. Appl., 181 (1994), pp. 483–492.
[17] S. N. Elaydi and R. J. Sacker, *Global stability of periodic orbits of nonautonomous difference equations and population biology*, J. Differential Equations, 208 (2005), pp. 258–273.
[18] S. N. Elaydi and R. J. Sacker, *Global stability of periodic orbits of nonautonomous difference equations in population biology and Cushing-Henson conjectures*, in Proceedings of the 8th International Conference on Difference Equations and Applications, Chapman & Hall/CRC, Boca Raton, FL, 2005, pp. 113–126.
[19] S. N. Elaydi and R. J. Sacker, *Nonautonomous Beverton-Holt equations and the Cushing-Henson conjectures*, J. Differential Equations Appl., 11 (2005), pp. 336–346.
[20] S. N. Elaydi and R. J. Sacker, *Periodic Difference Equations, Populations Biology and the Cushing-Henson Conjectures*, preprint, Trinity University.
[21] S. N. Elaydi and A.-A. Yakubu, *Global stability of cycles: Lotka-Volterra competition model with stocking*, J. Differential Equations Appl., 8 (2002), pp. 537–549.
[22] J. E. Franke and J. F. Selgrade, *Attractor for periodic dynamical systems*, J. Math. Anal. Appl., 286 (2003), pp. 64–79.
[23] J. E. Franke and A.-A. Yakubu, *Periodic dynamical systems in unidirectional metapopulation models*, J. Differential Equations Appl., 11 (2005), pp. 687–700.

[24] J. E. Franke and A.-A. Yakubu, *Multiple attractors via cusp bifurcation in periodically varying environments*, J. Differential Equations Appl., 11 (2005), pp. 365–377.
[25] J. E. Franke and A.-A. Yakubu, *Population models with periodic recruitment functions and survival rates*, J. Differential Equations Appl., 11 (2005), pp. 1169–1184.
[26] S. D. Fretwell, *Populations in a Seasonal Environment*, Princeton University Press, Princeton, NJ, 1972.
[27] K. P. Hadeler and P. van den Driessche, *Backward bifurcation in epidemic control*, Math. Biosci., 146 (1997), pp. 15–35.
[28] M. P. Hassell, *The Dynamics of Competition and Predation*, Studies in Biol. 72, The Camelot Press, Southampton, UK, 1976.
[29] M. P. Hassell, J. H. Lawton, and R. M. May, *Patterns of dynamical behavior in single species populations*, J. Animal Ecol., 45 (1976), pp. 471–486.
[30] S. M. Henson, *Multiple attractors and resonance in periodically forced population models*, Phys. D, 140 (2000), pp. 33–49.
[31] S. M. Henson, *The effect of periodicity in maps*, J. Differential Equations Appl., 5 (1999), pp. 31–56.
[32] S. M. Henson, R. F. Costantino, J. M. Cushing, B. Dennis, and R. A. Desharnais, *Multiple attractors, saddles, and population dynamics in periodic habitats*, Bull. Math. Biol., 61 (1999), pp. 1121–1149.
[33] S. M. Henson and J. M. Cushing, *The effect of periodic habitat fluctuations on a nonlinear insect population model*, J. Math. Biol., 36 (1997), pp. 201–226.
[34] D. Jillson, *Insect populations respond to fluctuating environments*, Nature, 288 (1980), pp. 699–700.
[35] V. L. Kocic, *A note on nonautonomous Beverton-Holt model*, J. Differential Equations Appl., 11 (2005), pp. 415–422.
[36] V. L. Kocic and G. Ladas, *Global Behavior of Nonlinear Difference Equations of Higher Order with Applications*, Math. Appl. 256, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.
[37] R. Kon, *A note on attenuant cycles of population models with periodic carrying capacity*, J. Differential Equations Appl., 10 (2004), pp. 791–793.
[38] R. Kon, *Attenuant cycles of population models with periodic carrying capacity*, J. Differential Equations Appl., 11 (2005), pp. 423–430.
[39] J. Li, *Periodic solutions of population models in a periodically fluctuating environment*, Math. Biosci., 110 (1992), pp. 17–25.
[40] R. M. May and G. F. Oster, *Bifurcations and dynamic complexity in simple ecological models*, Amer. Naturalist, 110 (1976), pp. 573–579.
[41] R. M. May, *Simple mathematical models with very complicated dynamics*, Nature, 261 (1977), pp. 459–469.
[42] R. M. May, *Stability and Complexity in Model Ecosystems*, Princeton University Press, Princeton, NJ, 1974.
[43] A. J. Nicholson, *Compensatory reactions of populations to stresses, and their evolutionary significance*, Aust. J. Zool., 2 (1954), pp. 1–65.
[44] R. M. Nisbet and W. S. C. Gurney, *Modelling Fluctuating Populations*, Wiley & Sons, New York, 1982.
[45] S. Rosenblat, *Population models in a periodically fluctuating environment*, J. Math. Biol., 9 (1980), pp. 23–36.
[46] J. F. Selgrade and H. D. Roberds, *On the structure of attractors for discrete, periodically forced systems with applications to population models*, Phys. D, 158 (2001), pp. 69–82.
[47] P. van den Driessche and J. Watmough, *A simple SIS epidemic model with a backward bifurcation*, J. Math. Biol., 40 (2000), pp. 525–540.
[48] A.-A. Yakubu, *Periodically forced nonlinear difference equations with delay*, in Difference Equations and Discrete Dynamical Systems, Proceedings of the 9th International Conference, University of Southern California, L. Allen, B. Aulbach, S. Elaydi, and R. Sacker, eds., World Scientific, River Edge, NJ, 2005, pp. 217–231.
[49] A.-A. Yakubu and M. Fogarty, *Spatially discrete metapopulation models with directional dispersal*, Math. Biosci., to appear.
[50] P. Yodzis, *Introduction to Theoretical Ecology*, Harper and Row, New York, 1989.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1588–1609

THE MOTION OF A THIN LIQUID FILM DRIVEN BY SURFACTANT AND GRAVITY*

R. LEVY† AND M. SHEARER‡

Abstract. We investigate wave solutions of a lubrication model for surfactant-driven flow of a thin liquid film down an inclined plane. We model the flow in one space dimension with a system of nonlinear PDEs of mixed hyperbolic-parabolic type in which the effects of capillarity and surface diffusion are neglected. Numerical solutions reveal distinct patterns of waves that are described analytically by combinations of traveling waves, some with jumps in height and surfactant concentration gradient. The various waves and combinations are strikingly different from what is observed in the case of flow on a horizontal plane. Jump conditions admit new shock waves sustained by a linear surfactant wave traveling upstream. The stability of these waves is investigated analytically and numerically. For initial value problems, a critical ratio of upstream to downstream height separates two distinct long-time wave patterns. Below the critical ratio, there is also an exact solution in which the height is piecewise constant and the surfactant concentration is piecewise linear and has compact support.

Key words. PDE, surfactants, hyperbolic-parabolic system

AMS subject classifications. 35G30, 35M10, 76Z05, 76D08

DOI. 10.1137/050637030

1. Introduction.

Spatial variations in surface tension on the free surface of a fluid induce a surface force, known as a Marangoni force [8]. Such variations can occur through temperature changes [2, 4] or by introducing surfactants. In thin film flow, surfactants have been studied in the context of industrial coating processes [9, 16] and as a component in the treatment of premature babies, whose lungs are in danger of collapse due to insufficient natural surfactant [3, 11]. Both applications have motivated extensive recent research on thin films driven by surfactant [12, 13, 17, 18].

In much of the research into surfactant spreading, the effect of gravity in driving the flow has been assumed to be negligible [10]. In this paper, we consider flow on an inclined plane, building on the work of Edmonstone, Craster, and Matar [6, 7] that explores the effect of adding gravity to the driving force. Our results demonstrate that gravity has a profound effect and probably should not be neglected in simulations of surfactant spreading.

For constant Marangoni force, as in studies of thermally driven thin films [2, 4, 8], the equations of motion are reasonably represented by a scalar fourth order PDE, known as the thin film equation. In the presence of surfactant, however, the Marangoni force is not constant; the density and motion of the surfactant molecules are modeled by an additional equation. The full model, derived using lubrication theory in [3], consists of a system of two nonlinear coupled PDEs for the film height and the surfactant concentration.

The PDE system exhibits a complicated combination of wave-like structures in the solution of initial value and boundary value problems [6, 7]. The complexity comes

*Received by the editors July 28, 2005; accepted for publication (in revised form) March 28, 2006; published electronically June 9, 2006. This work was supported by NSF grant DMS-0244491.
http://www.siam.org/journals/siap/66-5/63703.html
†Department of Mathematics, Duke University, Durham, NC 27708 (rachel.levy@gmail.com).
‡Department of Mathematics and Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695, and Department of Mathematics, Duke University, Durham, NC 27708 (shearer@math.ncsu.edu).

about in part because the underlying equation for the transport of surfactant is a degenerate *parabolic* equation resembling the porous medium equation, the degeneracy occurring at zero surfactant concentration. In contrast with the scalar thin film equation [15], we cannot appeal to the theory of hyperbolic equations to predict the long-time behavior of solutions, and indeed we find that it is different from the combination of shocks and rarefaction waves of the hyperbolic theory. Nonetheless, we explore the point of view that long-time behavior of the underlying equations should be predictable from basic information in the initial and boundary conditions, and it should be given by combinations of similarity solutions of the equations.

The underlying equations we study are derived from the full lubrication system (in section 2) by omitting terms related to surface diffusion and capillarity, i.e., taking the Péclet number to be infinite and the capillary number to be zero. We also neglect a second order diffusive term that is small for steep inclines. This enables us to focus on the underlying structure of solutions by analogy with the connection between hyperbolic systems and their parabolic regularization.

Numerical simulations of the reduced system, described in section 3, contain discontinuities in film height and surfactant concentration gradient. These solutions have a recognizable structure, and the goal of much of the rest of the paper is to explain the structure analytically. There is extensive numerical evidence [6] that the omitted regularizing terms primarily smooth the discontinuities without changing their structure or speeds, provided the coefficients are small.

The analysis of traveling waves in section 4 and jump conditions in section 5 reveals a surprisingly rich variety of individual waves. In section 6, we show how these waves can be combined to explain the wave patterns in the numerical simulations. However, the combination is possible only below a critical value of the ratio h_R/h_L of downstream height h_R to upstream height h_L. This analysis is used in section 7 to generate special exact solutions that are piecewise constant in h and piecewise linear in Γ, with Γ having compact support.

Above the critical value a different wave pattern emerges, described through numerical experiments in section 7. A notable feature of these new patterns is a hyperbolic precursor wave propagating ahead of the surfactant front. Also in this section we probe the critical value in PDE simulations and investigate the stability of individual waves predicted by jump conditions. In section 8 we summarize the catalogue of new individual waves and discuss the results in the context of ongoing research into the role of gravity in surfactant spreading.

2. The model. Consider a flat solid substrate, inclined as shown in Figure 1, in which $z = \tilde{h}(x, y, t)$ is the height of a thin film flowing down the slope. On the surface of the film is a layer of surfactant with concentration $\tilde{\Gamma}(x, y, t)$, which measures the density of surfactant molecules on the free surface. The surfactant is assumed to be immiscible and does not add to the height of the film.[1]

The full multidimensional model consists of a system of two nonlinear PDEs derived from the Navier–Stokes equations and the well-known lubrication approximation [1, 6, 13, 19]:

$$h_t + \nabla \cdot \left[\mathbf{C} \frac{h^3}{3} \nabla \nabla^2 h - \mathbf{G} \cos\theta \frac{h^3}{3} \nabla h - \frac{h^2}{2} \nabla \Gamma \right] + \left[\mathbf{G} \sin\theta \frac{h^3}{3} \right]_x = 0, \tag{2.1}$$

[1] The amount of surfactant is also assumed to be below the critical micelle concentration [5].

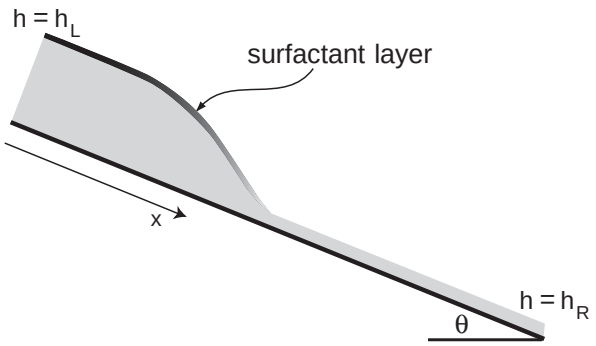


FIG. 1. *A thin film on an inclined substrate with partial coating by surfactant of varying concentration and negligible height.*

(2.2)
$$\Gamma_t + \nabla \cdot \left[\mathbf{C} \frac{h^2}{2} \Gamma \nabla \nabla^2 h - \mathbf{G} \cos\theta \frac{h^2}{2} \Gamma \nabla h - h \Gamma \nabla \Gamma \right] + \left[\mathbf{G} \sin\theta \frac{h^2}{2} \Gamma \right]_x - \mathbf{D} \nabla^2 \Gamma = 0.$$

Here h, Γ are dimensionless variables: $h = \tilde{h}/H$, $\Gamma = \tilde{\Gamma}/\Gamma_m$, where H is a characteristic length scale for the film thickness, and Γ_m is a typical surfactant concentration. The system contains nondimensional parameters associated with gravity, $\mathbf{G} = \frac{\rho g H L}{\Pi}$, surface diffusion $\mathbf{D} = \frac{1}{\mathbf{Pe}} = \frac{\mu D_s}{\Pi H}$ (where $\mathbf{Pe}$ is the Péclet number), and capillarity $\mathbf{C} = \frac{\epsilon^2 \sigma_m}{\Pi}$. These parameters depend on density ρ, gravity g, viscosity μ, and surface diffusivity D_s of the surfactant, all taken to be constant. They also depend on a characteristic length L of the film and on the small parameter $\epsilon = \frac{H}{L}$. The spreading pressure Π is given by $\Pi = \sigma_0 - \sigma_m$, where σ_0 is the surface tension in the absence of surfactant, and σ_m is a typical reduced surface tension in the presence of a typical concentration of surfactant. The spreading pressure Π is related to the Marangoni number, which after nondimensionalization is effectively set to 1. Note that in (2.1), (2.2) we have used a linear relation $\sigma = 1 - \Gamma$ (in nondimensional form) between surface tension and surfactant concentration. We refer the reader to [6] for typical values of the parameters.

In this paper, we consider a reduced model in which the variables are considered to be independent of the transverse variable y, and the regularizing effects of capillarity and surface diffusion are neglected by taking $\mathbf{C} = 0$ and $\mathbf{D} = 0$. Letting $\alpha = \mathbf{G} \sin\theta$, the reduced equations are

$$h_t - \frac{1}{2} \left(h^2 \Gamma_x \right)_x + \frac{\alpha}{3} \left(h^3 \right)_x = 0, \tag{2.3}$$

$$\Gamma_t - (h \Gamma \Gamma_x)_x + \frac{\alpha}{2} \left(h^2 \Gamma \right)_x = 0. \tag{2.4}$$

In this system we have also neglected the gravity terms with coefficient $\mathbf{G} \cos\theta$. This coefficient is small for θ near $\frac{\pi}{2}$, where it has a minor smoothing effect on solutions [6]. It is well known that fourth order diffusion gives rise to a capillary ridge [20], which is not captured in the reduced system. Neither capillarity nor surface diffusion affects wave speeds significantly, at least for small values of $\mathbf{C}$ and $\mathbf{D}$, as verified numerically in [6].

It is perhaps helpful to compare this system to familiar PDEs. For a given function $\Gamma(x,t)$, (2.3) would be a scalar conservation law

$$h_t + f(h)_x = 0,$$

with $f(h) = -\frac{1}{2}h^2\Gamma_x + \frac{\alpha}{3}h^3$. Similarly, for a given $h(x,t)$, (2.4) resembles a porous medium equation. Specifically, if h is constant, and $\alpha = 0$, then (2.4) is

$$\Gamma_t = \frac{1}{2}h\left(\Gamma^2\right)_{xx}.$$

Consequently, we may expect discontinuities in h and Γ_x, while Γ itself remains continuous; this would be typical in a quasi-linear conservation law for h and a porous medium equation for Γ.

As this discussion suggests, the system of equations is of parabolic-hyperbolic type. In a series of papers [17, 18], Renardy examined analytical issues such as local existence, propagation speed, and formation of shocks in the case $\alpha = 0$. While some of these results may generalize to the case $\alpha > 0$, we do not pursue this line of analysis in this paper.

3. Numerical experiments I. The numerical algorithm used to compute solutions of the system of PDEs (2.3), (2.4) is a first order composite finite difference scheme that couples a fully implicit time step and central spatial differences for second order derivatives with an upwind scheme for the more hyperbolic first order terms (with coefficient α in (2.3), (2.4)).

We define a spatial finite difference operator acting on $u_j^n = u(x_j, t_n)$ as

$$(\delta_x u)^n_{j+\frac{1}{2}} \equiv \frac{u^n_{j+1} - u^n_j}{\Delta x}. \tag{3.1}$$

We also use standard notation for spatial averages:

$$(\bar{u})^n_{j+\frac{1}{2}} \equiv \frac{u^n_{j+1} + u^n_j}{2}, \tag{3.2}$$

and let $\lambda \equiv \frac{\Delta t}{\Delta x}$. The finite difference scheme is then defined as

$$h_j^{n+1} = h_j^n + \lambda\left(\frac{1}{2}\left(\bar{h}^2\delta_x\Gamma\right)^{n+1}_{j+\frac{1}{2}} - \frac{1}{2}\left(\bar{h}^2\delta_x\Gamma\right)^{n+1}_{j-\frac{1}{2}} - \frac{\alpha}{3}\left((h^3)^n_j - (h^3)^n_{j-1}\right)\right), \tag{3.3}$$

$$\Gamma_j^{n+1} = \Gamma_j^n + \lambda\left(\left(\bar{h}\bar{\Gamma}\delta_x\Gamma\right)^{n+1}_{j+\frac{1}{2}} - \left(\bar{h}\bar{\Gamma}\delta_x\Gamma\right)^{n+1}_{j-\frac{1}{2}} - \frac{\alpha}{2}\left((h^2\Gamma)^n_j - (h^2\Gamma)^n_{j-1}\right)\right). \tag{3.4}$$

The resulting nonlinear system is solved with Newton's method; the composite scheme is formally first order in both time and space.

The goal of our analysis is to understand the structure of long-time solutions as a function only of boundary data. In order to simulate initial value problems with initial data in which h and Γ are constant outside an interval, and $\Gamma = 0$ downstream, we impose simple boundary conditions at the end points of the computational domain $[0, x_{max}]$:

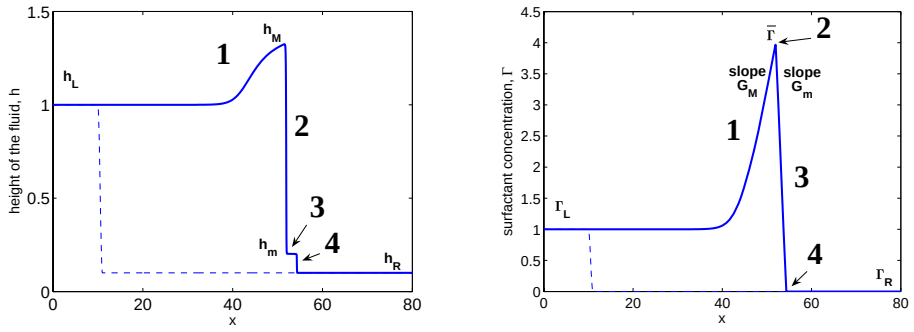


FIG. 2. *Typical solution profiles* $h(x,60)$ *(left) and* $\Gamma(x,60)$ *(right) with* $h_R = 0.1$. *Initial data* $h(x,0)$ *and* $\Gamma(x,0)$ *are represented by dashed lines. Numbered labels:* (1) *nonlinear traveling wave;* (2) *jump in* h, Γ_x*;* (3) *linear traveling wave with* Γ *linear and* h *constant;* (4) *jump in* h, Γ_x.

$$h(0,t) = h_L, \quad \Gamma(0,t) = \Gamma_L; \quad h(x_{max},t) = h_R, \quad \Gamma(x_{max},t) = 0.$$

In the PDE simulations, we exploit the following scalings in the equations. Let $h = H(x,t)$, $\Gamma = g(x,t)$ satisfy (2.3), (2.4) with $H(0,t) = 1$, $g(0,t) = 1$, and $\alpha = 1$. Then h, Γ given by

$$h(x,t) = h_L\, H\left(\frac{\alpha h_L}{\Gamma_L}x, \frac{\alpha^2 h_L^3}{\Gamma_L}t\right) \quad \text{and} \quad \Gamma = \Gamma_L\, g\left(\frac{\alpha h_L}{\Gamma_L}x, \frac{\alpha^2 h_L^3}{\Gamma_L}t\right) \tag{3.5}$$

satisfy (2.3), (2.4) with $h(0,t) = h_L$, $\Gamma(0,t) = \Gamma_L$, provided $\Gamma_L > 0$. Thus, to explore variation in wave structures, we can fix $\alpha = h_L = \Gamma_L = 1$ and vary h_R (note that $\Gamma_R = 0$). We employ this simplification in numerical simulations. In the analysis, we retain all the parameters so that their effect can be seen explicitly.

Figure 2 contains graphs of numerical simulations of (2.3), (2.4) at time $t = 60$. The initial data (indicated by dashed lines) contain a single (smoothed) jump in h from $h_L = 1.0$ to $h_R = 0.1$ and in Γ from $\Gamma_L = 1.0$ to $\Gamma_R = 0$. The same structure emerges from the same upstream and downstream heights for more general smooth height data (including oscillatory data) after transients have died away. The figure is annotated to show the broad wave structure of the solution. In the graph of h, note that at (1) the height increases monotonically from the fixed boundary value h_L to a maximum height h_M; at (2) there is a jump from h_M to the height h_m of the horizontal "step" at (3). Finally, the step contains a second jump at (4) from h_m to the precursor layer of fixed height h_R.

The graph of the surfactant concentration Γ has jumps at the same locations as those of the height, but the jumps are in the slope Γ_x (henceforth denoted by G, as in the figure), while Γ is continuous. The surfactant concentration increases monotonically (1) to a maximum concentration Γ_0 at the corner (2), where there is a jump in Γ_x from $G_M > 0$ to $G_m < 0$. In (3), Γ appears to be linear, extending down to $\Gamma = 0$, where there is a second jump (4) in Γ_x from $G_m < 0$ to zero.

As the spatial grid is refined, the discontinuities are better resolved, and no additional data points appear in the discontinuities. We demonstrate this for the jump in height in Figure 3, in which the number of grid points is doubled from 2500 ($\Delta x = 0.012$) to 5000 ($\Delta x = 0.006$).

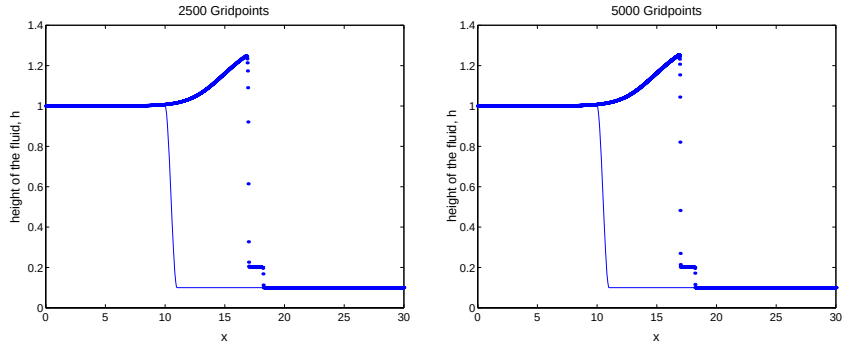


FIG. 3. *Grid refinement to distinguish the discontinuity in height from a steep gradient. The coarse grid on the left has* 2500 *points; the fine grid on the right has* 5000 *points. The number of points in the shock is approximately the same in both cases.*

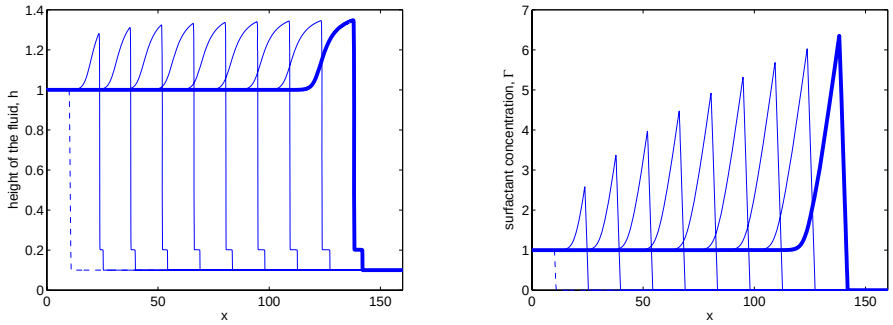


FIG. 4. *Numerical solutions showing the evolution of* h *and* Γ *for* $h_R = 0.1$. *The profiles are separated by* 40 *time steps, and the final profiles at* $t = 180$ *are in bold. The computational time step is* $\Delta t = 0.001$, *and the grid is fairly coarse with* $\Delta x = 0.04$.

Figure 4 illustrates the evolution of the solution, with a dashed line indicating the initial profile. After a rapid initial transient between the first two plots, the maximum height h slowly asymptotes to a constant value, whereas the maximum surfactant concentration Γ increases without bound. (Note that surfactant is supplied from the left boundary.) The step height h_m develops at a very early time and remains constant, although the width of the step increases slowly.

To visualize the data from Figure 4 in a different way, in Figure 5 we plot the curves $(h(x,t), \Gamma(x,t))$, $0 < x < x_{max}$, for various values of $t > 0$. Starting from $(h_L, \Gamma_L) = (1,1)$, there is a sequence of curves on the right that appears to be converging. A jump in h occurs at the maximum value of Γ, which is increasing in time. On the left of the figure, Γ decreases to zero at a height h which is constant in space and time. The final step occurs on the h-axis.

4. Traveling waves. In this section we explore analytical solutions of the PDE (2.3), (2.4) that capture some of the features observed in the numerical simulations. The traveling waves we consider are smooth solutions $h = \hat{h}(x - ct)$, $\Gamma = \hat{\Gamma}(x - ct)$ that move with constant speed c. Substituting into system (2.3), (2.4) and integrating

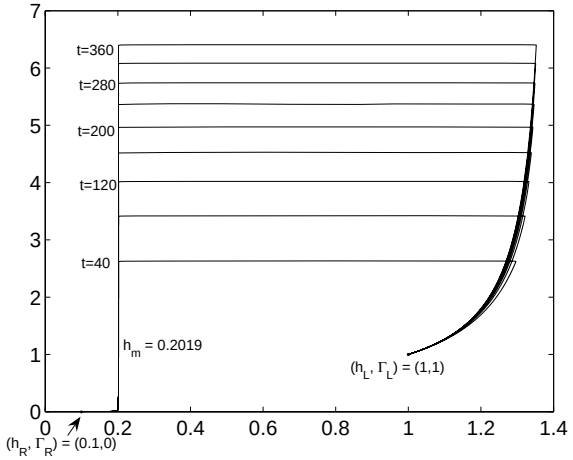


FIG. 5. *Phase portrait of numerical PDE solutions* (h, Γ) *with* $h_R = 0.1$.

once, we obtain the ODE system (dropping the hats)

$$-ch - \frac{1}{2}h^2\Gamma' + \frac{\alpha}{3}h^3 = K_1, \tag{4.1}$$

$$-c\Gamma - h\Gamma\Gamma' + \frac{\alpha}{2}h^2\Gamma = K_2 \tag{4.2}$$

in which $\Gamma' = \frac{d\Gamma}{d\xi}$, $\xi = x - ct$. First, we observe that there are traveling waves in which h is constant and Γ is linear. We refer to these traveling waves as *simple traveling waves*. Note that simple traveling waves should be considered to be defined only for values of $x - ct$ where Γ is nonnegative.

THEOREM 4.1. *Let* $h > 0$ *be constant, and let* $\Gamma = \Gamma_0 + G\xi$, *with* $G = \Gamma' =$ *constant. Then* (4.1) *is satisfied identically with* $K_1 = -ch - \frac{1}{2}h^2G + \frac{\alpha}{3}h^3$, *and* (4.2) *is satisfied if and only if* $K_2 = 0$ *and*

$$c = -hG + \frac{\alpha}{2}h^2. \tag{4.3}$$

Proof. The only complication is in the Γ equation, where the left-hand side has a constant term and a term linear in ξ. The restrictions on K_2 and c come from equating coefficients. ∎

Interestingly, the speed in (4.3) is the transport velocity in (2.4); it is neither the transport velocity $-\frac{1}{2}hG + \frac{\alpha}{3}h^2$ nor the characteristic speed $-hG + \alpha h^2$ for the conservation law (2.3) with $\Gamma_x = G$ constant. This is not a contradiction, since h is constant.

There are also *nonlinear traveling waves* corresponding to solutions of system (4.1), (4.2) in which h is not constant and Γ is not linear. Consider a solution $h(\xi)$, $\Gamma(\xi)$ with boundary conditions at $\xi = -\infty$:

$$h(-\infty) = h_L, \quad \Gamma(-\infty) = \Gamma_L, \quad \Gamma'(-\infty) = 0. \tag{4.4}$$

These conditions serve to determine the constants K_1, K_2, so that (4.1), (4.2) become

$$-ch - \frac{1}{2}h^2\Gamma' + \frac{\alpha}{3}h^3 = -ch_L + \alpha\frac{h_L^3}{3}, \tag{4.5}$$

$$-c\Gamma - h\Gamma\Gamma' + \frac{\alpha}{2}h^2\Gamma = -c\Gamma_L + \frac{\alpha}{2}h_L^2\Gamma_L. \tag{4.6}$$

Eliminating Γ', we obtain an invariant curve through (h_L, Γ_L) in the (h, Γ)-plane:

$$\Gamma(h) = \frac{h\Gamma_L\left(c - \frac{\alpha}{2}h_L^2\right)}{c(2h_L - h) + \alpha\left(\frac{1}{6}h^3 - \frac{2}{3}h_L^3\right)}. \tag{4.7}$$

On the curve $\Gamma(h)$, the flow is given by solving (4.1) for Γ' and using the chain rule $\frac{d\Gamma}{d\xi} = \frac{d\Gamma}{dh}\frac{dh}{d\xi}$ to obtain an expression for h'. Differentiating (4.7), we obtain

$$\frac{d\Gamma}{dh} = \frac{6\Gamma_L(2c - \alpha h_L^2)(6ch_L - \alpha h^3 - 2\alpha h_L^3)}{(12ch_L - 6ch + \alpha h^3 - 4\alpha h_L^3)^2}. \tag{4.8}$$

Therefore, after some processing, we find

$$\frac{d\Gamma}{d\xi} = \frac{\frac{2}{3}(h_L - h)(3c - \alpha(h^2 + hh_L + h_L^2))}{h^2}, \tag{4.9}$$

$$\frac{dh}{d\xi} = \frac{-\frac{1}{9}(h - h_L)(3c - \alpha(h^2 + hh_L + h_L^2))(12ch_L - 6ch + \alpha h^3 - 4\alpha h_L^3)^2}{h^2\Gamma_L(2c - \alpha h_L^2)(6ch_L - \alpha h^3 - 2\alpha h_L^3)}. \tag{4.10}$$

In particular, the point (h_L, Γ_L) is an equilibrium; whether it is stable or unstable depends on the signs of the various factors in (4.9), (4.10). It is unstable for the traveling waves we seek.

As suggested by the numerical results of the previous section, the traveling waves of most interest to us in this paper have h and Γ increasing, with h approaching finite limits at $\pm\infty$, and Γ unbounded at $+\infty$. The following result establishes the existence of such waves.

THEOREM 4.2. *For each $h_M \in (h_L, 2^{\frac{2}{3}}h_L)$, there is a traveling wave solution $h(\xi)$, $\Gamma(\xi)$, $\xi = x - ct$ satisfying*

(i) *$h'(\xi) > 0$; $\Gamma'(\xi) > 0$;*

(ii) *$h(-\infty) = h_L$, $\Gamma(-\infty) = \Gamma_L$, $h(\infty) = h_M$, $\Gamma(\infty) = \infty$, $\Gamma'(\infty) = G_M$*

with speed

$$c = -h_M G_M + \frac{\alpha}{2}h_M^2, \tag{4.11}$$

where

$$G_M = \frac{\alpha}{3}\frac{(h_L - h_M)}{h_M(2h_L - h_M)}(h_M^2 - 2h_Lh_M - 2h_L^2). \tag{4.12}$$

Proof. Let

$$d(h; c, h_L) = c(2h_L - h) + \frac{\alpha}{6}(h^3 - 4h_L^3),$$

the denominator in the expression (4.7) for $\Gamma = \Gamma(h)$. Solving $d(h; c, h_L) = 0$, we find

$$c = c(h) = \frac{\alpha}{6}\left(\frac{h^3 - 4h_L^3}{h - 2h_L}\right). \tag{4.13}$$

By considering the graph of this rational function of h, we find c has a positive local maximum at $h = h_L$, a zero at $h = 2^{\frac{2}{3}}h_L < 2h_L$, and $c(0) = \frac{\alpha}{3}h_L^2$. Consequently, for h_M in the range of the theorem, there is a single zero of d, parameterized by either $h = h_M$ or by $c \in (0, \frac{\alpha}{2}h_L^2)$. For $\frac{\alpha}{3}h_L^2 < c < \frac{\alpha}{2}h_L^2$, there is a second positive root $h_0 < h_L$, which crosses $h = 0$ at $c = \frac{\alpha}{3}h_L^2$.

In summary, the zero of d that we seek is given by (4.13) with $h = h_M$:

$$c = c(h_M) = \frac{\alpha}{6}\left(\frac{h_M^3 - 4h_L^3}{h_M - 2h_L}\right), \quad h_L < h_M < 2^{\frac{2}{3}}h_L. \tag{4.14}$$

Then $\Gamma(h)$ has a vertical asymptote at $h = h_M$ and, moreover, Γ stays positive: $\Gamma(h) \to \infty$ as $h \nearrow h_M$. Examining the flow (4.9), (4.10), we find that (h_L, Γ_L) is an unstable equilibrium. Let $(h(\xi), \Gamma(\xi))$ be the corresponding trajectory satisfying (4.5), (4.6) with $(h(-\infty), \Gamma(-\infty)) = (h_L, \Gamma_L)$, $h' > 0$, $\Gamma' > 0$. Then $h(\infty) = h_M$ and $\Gamma(\infty) = \infty$.

The final step is to set $h = h_M$ in (4.9), from which we find (after substitution for c from (4.14)) $\Gamma'(\xi) \to G_M$, as given in (4.12). Formula (4.11) is then easily checked. □

Remarks. 1. Formula (4.11) suggests that the traveling wave approaches a simple traveling wave as $\xi \to \infty$. Indeed, $h(\xi) \to h_M$, a constant, and $\Gamma'(\xi) \to G_M$, as $\xi \to \infty$.

2. Equations (4.11), (4.12) link the asymptotic behavior as $\xi \to \infty$ to the equilibrium (h_L, Γ_L) without needing to integrate the ODEs. The formulae are independent of Γ_L, reflecting the scale invariance of the equations. Moreover, these formulae persist as $\Gamma_L \to 0$. In this limit, the traveling wave approaches a simple nonsmooth wave (see Corollary 5.5 below).

3. From (4.14), we observe that $c(h_M) \in \left(\frac{\alpha}{3}h_L^2, \frac{\alpha}{2}h_L^2\right)$, an interval that collapses onto zero as $\alpha \to 0$. Consequently, the nonlinear traveling waves of the theorem do not appear in the case $\alpha = 0$ of the horizontal substrate.

5. Jump conditions. As discussed above, and as observed in numerical simulations, solutions of system (2.3), (2.4) typically contain discontinuities in h and Γ_x. In this section, we begin a systematic study of the Rankine–Hugoniot jump conditions for the system.

Consider the Cauchy problem for system (2.3), (2.4) with initial data

$$h(x,0) = h_0(x), \quad \Gamma(x,0) = \Gamma_0(x), \quad -\infty < x < \infty. \tag{5.1}$$

By a *weak solution* of the Cauchy problem, we mean a function $(h, \Gamma) : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+^2$ with $h, \Gamma, \Gamma_x \in L^\infty \cap L^1_{loc}$ such that for every test function $\phi \in C_c^\infty(\mathbb{R} \times \mathbb{R}_+)$,

$$\begin{aligned} \int_0^\infty \int_{-\infty}^\infty \left\{h\phi_t - (\tfrac{1}{2}h^2\Gamma_x - \tfrac{\alpha}{3}h^3)\phi_x\right\} dx\, dt &= \int_{-\infty}^\infty h_0\phi(x,0)dx, \\ \int_0^\infty \int_{-\infty}^\infty \left\{\Gamma\phi_t - (h\Gamma\Gamma_x - \tfrac{\alpha}{2}h^2\Gamma)\phi_x\right\} dx\, dt &= \int_{-\infty}^\infty \Gamma_0\phi(x,0)dx. \end{aligned} \tag{5.2}$$

Note that in this definition of weak solutions, $\Gamma(x,t)$ is continuous with respect to x for each time t.

In analyzing jump conditions, we will always assume that h is C^1 and Γ is C^2, apart from a finite number of curves $x = \gamma(t)$ across which h and Γ_x can have jump discontinuities. We refer to such solutions as *piecewise-smooth*. We shall be

particularly interested in solutions in which surfactant spreads into a region with initially uniform film height. Let (h, Γ) be a piecewise-smooth weak solution, and suppose the leading edge of the surfactant is located on a curve $x = \gamma_\ell(t)$:

$$\Gamma(x,t) = 0, \quad x \geq \gamma_\ell(t). \tag{5.3}$$

The following theorem determines the speed at the leading edge of the surfactant.

THEOREM 5.1. *Let h, Γ be a piecewise-smooth weak solution of* (2.3), (2.4) *satisfying* (5.3). *If $G_- \equiv \Gamma_x(\gamma_\ell(t)-, t) < 0$, then*

$$\gamma_\ell'(t) = -h_- G_- + \frac{\alpha}{2} h_-^2, \tag{5.4}$$

where $h_- = h(\gamma_\ell(t)-, t)$.

Proof. Differentiating $\Gamma(\gamma_\ell(t)-, t) = 0$, we have

$$\Gamma_x \gamma_\ell'(t) + \Gamma_t = 0.$$

But from (2.4), $\Gamma_t = (h\Gamma\Gamma_x)_x - \frac{\alpha}{2}\left(h^2\Gamma\right)_x = h_- G_-^2 - \frac{\alpha}{2} h_-^2 G_-$, since $\Gamma = 0$ at the leading edge $x = \gamma_\ell(t)$. The formula (5.4) now follows, since $\Gamma_x = G_- \neq 0$. ∎

Now we derive a set of equations from the Rankine–Hugoniot conditions for the discontinuities in h and Γ_x and an equation enforcing the continuity of Γ at the discontinuities. Consider a solution (h, Γ) that is smooth away from a differentiable curve $C = \{x = \gamma(t)\}$ such that Γ is continuous across C, and h, Γ_x have well-defined one-sided limits at C. Let $[u]$ denote the jump in a function u across C, and let $c = \gamma'(t)$. It is also convenient to use the notation

$$G = \Gamma_x, \quad \overline{\Gamma} = \Gamma(\gamma(t), t).$$

Then the jump conditions are as follows. From (2.3) we obtain

$$-c[h] + \left[-\frac{1}{2} h^2 G + \frac{\alpha}{3} h^3\right] = 0, \tag{5.5}$$

while (since Γ is continuous) (2.4) yields

$$\overline{\Gamma}\left[-hG + \frac{\alpha}{2} h^2\right] = 0. \tag{5.6}$$

The continuity of Γ provides an additional equation matching the left and right limits of Γ at $x = \gamma(t)$:

$$\Gamma(\gamma(t)-, t) = \Gamma(\gamma(t)+, t). \tag{5.7}$$

We refer to weak solutions that consist of a simple traveling wave (see section 4) on either side of a jump in h, Γ_x as a *simple jump*. In a simple jump, both h and G are constant, so that the jump condition (5.5) implies the speed is constant also. By translation invariance, we may without loss of generality take the simple jump to lie along the line $x = ct$:

(5.8)

$$h(x,t) = \begin{cases} h_- & \text{if } x < ct, \\ h_+ & \text{if } x > ct, \end{cases} \qquad \Gamma(x,t) = \begin{cases} \Gamma_0 + G_-(x - c_- t) & \text{if } x < ct, \\ \Gamma_0 + G_+(x - c_+ t) & \text{if } x > ct. \end{cases}$$

In this solution, we are leaving open the possibility that the simple traveling waves on either side of the jump travel with different speeds. In this case, continuity of Γ (5.7) requires

$$G_-(c-c_-) = G_+(c-c_+). \tag{5.9}$$

The jump conditions (5.5), (5.6) for a simple jump (5.8) become

$$-c(h_+ - h_-) - \frac{1}{2}(h_+^2 G_+ - h_-^2 G_-) + \frac{\alpha}{3}(h_+^3 - h_-^3) = 0, \tag{5.10}$$

$$(\Gamma_0 + G_-(c-c_-)t)\left(-h_+G_+ + h_-G_- + \frac{\alpha}{2}(h_+^2 - h_-^2)\right) = 0. \tag{5.11}$$

Equating coefficients in the second equation, we have a pair of conditions:

$$\begin{aligned} \Gamma_0\left(-h_+G_+ + h_-G_- + \tfrac{\alpha}{2}(h_+^2 - h_-^2)\right) &= 0, \\ G_-(c-c_-)\left(-h_+G_+ + h_-G_- + \tfrac{\alpha}{2}(h_+^2 - h_-^2)\right) &= 0. \end{aligned} \tag{5.12}$$

The analysis of the jump conditions is organized as follows. First, we show that for $\alpha = 0$, and $\overline{\Gamma} = 0$, there is a simple jump in h; the height doubles across the jump, and the speed is determined by the height and surfactant concentration gradient on one side of the jump. However, when $\overline{\Gamma} > 0$, there are no simple jumps when $\alpha = 0$. For $\alpha > 0$, there are two simple jumps: one at the leading edge, but with the jump in height coupled to the surfactant concentration gradient, and another in the bulk, where the jump conditions can be solved explicitly.

The horizontal substrate: $\alpha = 0$. First, we investigate the structure of the leading edge of the surfactant when $\alpha = 0$. The speed $c = \gamma_\ell'(t)$ of the leading edge $x = \gamma_\ell(t)$ was characterized in Theorem 5.1.

THEOREM 5.2.[2] *Let h, Γ be a piecewise-smooth weak solution of* (2.3), (2.4) *with $\alpha = 0$ satisfying* (5.3). *If $G_- = \Gamma_x(\gamma_\ell(t)-, t) < 0$, then*

$$h_- = 2h_+, \tag{5.13}$$

where $h_\pm = h(\gamma_\ell(t)\pm, t)$.

Proof. The result follows immediately by substituting the expression for the speed at the jump (5.4) into the height jump condition (5.5). Specifically, (5.5) with $\alpha = 0$ becomes

$$-c(h_+ - h_-) + \frac{1}{2}h_-^2 G_- = 0, \tag{5.14}$$

since $G_+ = \Gamma_x(\gamma_\ell(t)+, t) = 0$. But from (5.4), $c = -h_-G_-$. Substitution into (5.14) leads to the result, since $G_- \neq 0$. ∎

Next, we show there are no simple jumps with $\overline{\Gamma} > 0$.

THEOREM 5.3. *Let $\alpha = 0$, $\overline{\Gamma} > 0$. There are no simple jumps with $c \neq 0$. For $c = 0$, there is a nonphysical solution with a stationary jump in h and no jump in Γ_x.*

Proof. Set $\alpha = 0$ in the jump conditions (5.10), (5.12) to get

$$-c(h_+ - h_-) - \frac{1}{2}h_+^2 G_+ + \frac{1}{2}h_-^2 G_- = 0 \tag{5.15}$$

[2] These jumps were first characterized by Borgas and Grotberg [3].

and

$$\begin{aligned} \Gamma_0\,(-h_+G_+ + h_-G_-) &= 0, \\ G_-(c-c_-)\,(-h_+G_+ + h_-G_-) &= 0. \end{aligned} \tag{5.16}$$

Since $\Gamma_0 > 0$, we see that

$$h_-G_- = h_+G_+. \tag{5.17}$$

Consequently, the speeds c_-, c_+ of the simple traveling waves on each side of the simple jump are equal (see Theorem 4.1 with $\alpha = 0$):

$$c_- = -h_-G_- = c_+ = -h_+G_+. \tag{5.18}$$

Substituting (5.17) into (5.15), we conclude that either $h_+ = h_-$, in which case $G_+ = G_-$ and there is no jump, or $h_+ \neq h_-$ and a jump in h and G would have speed

$$c = -\frac{1}{2}h_-G_-. \tag{5.19}$$

But then (5.9), (5.18) imply that either $c = c_- = c_+ = 0$, so that $G_- = G_+ = 0$, or $G_- = G_+ \neq 0$. In the former case, we regard the solution with an arbitrary stationary jump in h as unphysical, since the fluid discontinuity would collapse in the presence of additional smoothing terms such as capillarity or second order diffusion. In the latter case, (5.17) implies $h_- = h_+$, and we are back to the simple traveling wave of the previous section with no jump in h or Γ_x. ∎

The inclined substrate: $\alpha > 0$. When we consider a film flowing down an inclined substrate in which $\alpha > 0$, we find solutions of the jump conditions that are strikingly different from the $\alpha = 0$ case. We explore the wave structures analytically here and numerically in section 7. As with $\alpha = 0$, we treat the leading edge of surfactant separately.

THEOREM 5.4. *Let h, Γ be a piecewise-smooth weak solution of* (2.3), (2.4) *with $\alpha > 0$ satisfying* (5.3). *Let $h_\pm = h(\gamma_\ell(t)\pm, t)$, and suppose $G_- = \Gamma_x(\gamma_\ell(t)-, t) < 0$. Then*

$$\text{(a)}\quad h_- < h_+ \quad \textit{or} \quad \text{(b)}\quad 2h_+ < h_- < (1+\sqrt{3})h_+; \tag{5.20}$$

in both cases

$$G_- = \frac{\alpha}{3}\frac{(h_- - h_+)(h_-^2 - 2h_-h_+ - 2h_+^2)}{h_-(h_- - 2h_+)}, \qquad \gamma_\ell'(t) = -h_-G_- + \frac{\alpha h_-^2}{2}. \tag{5.21}$$

Proof. The speed in (5.21) is given by Theorem 5.1 (see (5.4)). Substituting $c = \gamma_\ell'(t)$ and $G_+ = 0$ into the jump condition (5.10) gives the formula for G_- in (5.21). The inequalities (5.20) are equivalent to $G_- < 0$. (The inequality $G_- \leq 0$ is needed to ensure $\Gamma > 0$ behind the leading edge of the surfactant.) ∎

Remarks. 1. For $\alpha > 0$, the jump in h and the surfactant concentration gradient are linked, in contrast to the $\alpha = 0$ case, in which the jump in h is determined, and G_- is a free parameter that influences only the speed.

2. Conditions (5.21) correspond to the limit $\Gamma_L \to 0$ of the nonlinear traveling wave in Theorem 4.2.

3. The case $h_- < h_+$ appears to be unphysical; both gravity and the surfactant concentration gradient act downwards and cannot sustain such a jump in the height.

We conjecture that this jump is unstable, as suggested by the numerical evidence in Figure 8. Further explanation for the instability of the wave is suggested by the observation that the speed of the discontinuity in (5.21) is smaller than the characteristic speed αh_+^2 in the surfactant-free region ahead of the discontinuity.

The following corollary parallels the result of Theorem 5.4, except that it concerns the trailing edge of the surfactant, so the roles of h_+, h_- are exchanged. However, the cases are not symmetric, since gravity and the Marangoni force are now opposed. In particular, $h_+ > h_-$.

COROLLARY 5.5. *Let h, Γ be a piecewise-smooth weak solution of* (2.3), (2.4) *with $\alpha > 0$ satisfying $\Gamma(x,t) = 0$, $x \le \gamma_T(t)$. Let $h_\pm = h(\gamma_T(t)\pm, t)$, and suppose $G_+ = \Gamma_x(\gamma_T(t)+, t) > 0$. Then*

$$\text{(a)} \quad h_+ > (1+\sqrt{3})h_- \quad \textit{or} \quad \text{(b)} \quad h_- < h_+ < 2h_-; \tag{5.22}$$

in both cases

$$G_+ = \frac{\alpha}{3}\frac{(h_+ - h_-)(h_+^2 - 2h_-h_+ - 2h_-^2)}{h_+(h_+ - 2h_-)}, \qquad \gamma_T'(t) = -h_+G_+ + \frac{\alpha h_+^2}{2}. \tag{5.23}$$

Proof. The proof parallels that of Theorem 5.4 with h_-, h_+ switched; the bounds on $h_\pm$ ensure $G_+ > 0$. ∎

THEOREM 5.6. *For $\alpha > 0$, let h, Γ be a piecewise-smooth weak solution of* (2.3), (2.4) *with a simple jump across the line $x = ct$, where $\Gamma = \Gamma_0 > 0$. Then either*
(a) $\Gamma = \Gamma_0 + G(x - st)$ *is linear, in which case*

$$G = \frac{\alpha}{2}(h_+ + h_-); \quad s = -\frac{\alpha}{2}h_+h_-; \quad c = \frac{\alpha}{12}(h_+ - h_-)^2, \tag{5.24}$$

or (b) Γ_x *has a jump with*

$$\begin{gathered} G_- = -\frac{\alpha}{6h_-}(h_+ - h_-)(h_+ + 2h_-); \quad G_+ = \frac{\alpha}{6h_+}(h_+ - h_-)(2h_+ + h_-); \\ c = \frac{\alpha}{6}(h_+^2 + h_+h_- + h_-^2). \end{gathered} \tag{5.25}$$

Proof. In case (a), we substitute $\Gamma = \Gamma_0 + G(x - st)$, $h = h_\pm$ into (2.4) to find

$$s = -h_+G + \frac{\alpha}{2}h_+^2 = -h_-G + \frac{\alpha}{2}h_-^2.$$

(That is, $(h_\pm, \Gamma)$ is a simple traveling wave with the same speed s and surfactant gradient G on each side of $x = ct$ but with a jump in h.) The second equality gives the formula for G (consistent with (5.6)) from which the expression for s follows. The speed of the jump in h comes from the jump condition (5.10) in which $G_\pm = G$.

In case (b), since $\Gamma > 0$, the jump condition (5.11) together with Theorem 4.1 for simple traveling waves implies

$$c_+ = -h_+G_+ + \frac{\alpha}{2}h_+^2 = -h_-G_- + \frac{\alpha}{2}h_-^2 = c_-.$$

But now continuity of Γ (see (5.9)) and $G_+ \ne G_-$ imply $c = c_+ = c_-$. Substituting into the jump condition (5.10) leads to the result (noting that $h_+ = h_-$ implies $G_+ = G_-$). ∎

Remark. The motion of the film in this case is somewhat surprising, since the linear wave with slope $G = \frac{\alpha}{2}(h_+ + h_-)$ moves to the left with speed $s = -\frac{\alpha}{2}h_+h_-$, while simultaneously the jump in height moves to the right with speed $c = \frac{\alpha}{12}(h_+ - h_-)^2$.

6. Combining waves when $\alpha > 0$. The numerical experiments of section 3 suggest that as time increases, the solution approaches a combination of traveling waves and simple jumps. We can now interpret this structure in terms of the traveling waves of section 4 and the jumps of section 5. Each of the four numbered features in the figure is associated with formulae relating the parameters labeled in the figure to each other and to the wave speeds. In processing the equations, we observe that the jump condition (5.6) effectively equates wave speeds of simple traveling waves on either side of a jump. Consequently, the four wave speeds are all equal to a single speed c. We combine the following:

(1) A traveling wave with speed c connecting h_L to h_M (see Theorem 4.2):

$$c = \frac{\alpha\left(\frac{1}{6}h_M^3 - \frac{2}{3}h_L^3\right)}{h_M - 2h_L}, \tag{6.1}$$

$$G_M = \frac{\alpha}{3}\frac{(h_L - h_M)(h_M^2 - 2h_L h_M - 2h_L^2)}{h_M(2h_L - h_M)}. \tag{6.2}$$

(2) A simple jump in h, Γ_x (see Theorem 5.6):

$$G_M = -\frac{\alpha}{6h_M}(h_m - h_M)(h_m + 2h_M), \tag{6.3}$$

$$G_m = \frac{\alpha}{6h_m}(h_m - h_M)(2h_m + h_M). \tag{6.4}$$

(3) A simple traveling wave with Γ descending to zero and (4) a simple jump in h (see Theorems 4.1 and 5.4):

$$G_m = \frac{\alpha}{3}\frac{(h_m - h_R)(h_m^2 - 2h_m h_R - 2h_R^2)}{h_m(h_m - 2h_R)}. \tag{6.5}$$

Noting the symmetry in the expressions for G_m and G_M, we solve the equations by treating h_L and h_R as parameters. We can then simplify the equations by equating G_M in (6.2), (6.3) and equating G_m in (6.4), (6.5), giving two simultaneous equations for h_M, h_m:

$$2h_L(h_m^2 + h_M^2 - 2h_L^2) - h_M h_m(h_M + h_m - 2h_L) = 0, \tag{6.6}$$

$$2h_R(h_m^2 + h_M^2 - 2h_R^2) - h_M h_m(h_M + h_m - 2h_R) = 0. \tag{6.7}$$

THEOREM 6.1. *The polynomial equations* (6.6), (6.7) *have two positive solutions for* $0 < h_R/h_L < r^*$, *where* $r^* = \frac{1}{2}(\sqrt{3} - 1)$, *and no solution for* $r^* < h_R/h_L < 1$. *The only relevant solution has* $h_m < h_M$.

Proof. It is perhaps easier to see the structure if we rewrite the variables in (6.6), (6.7):

$$x = h_M, \quad y = h_m, \quad u = h_L, \quad v = h_R.$$

Then the system becomes

$$\begin{aligned} 2u(x^2 + y^2 - 2u^2) - xy(x + y - 2u) &= 0, \qquad \text{(a)} \\ 2v(x^2 + y^2 - 2v^2) - xy(x + y - 2v) &= 0. \qquad \text{(b)} \end{aligned} \tag{6.8}$$

Since $u \neq v$, taking $v(\mathrm{a}) - u(\mathrm{b})$ leads to

$$xy(x+y) = 4uv(u+v). \tag{6.9}$$

Substituting back into (6.8(a)), we get

$$x^2 + y^2 + xy = 2(u^2 + uv + v^2). \tag{6.10}$$

We can eliminate x, since $xy(x+y) = y(x^2+xy)$. Thus, $x^2 + xy = 4uv(u+v)/y$. Substituting into (6.10),

$$y^2 + 4uv(u+v)/y = 2(u^2 + uv + v^2),$$

leaving a cubic in y:

$$y^3 - ay + b = 0, \tag{6.11}$$

where $a = 2(u^2 + uv + v^2)$, $b = 4uv(u+v)$. Consequently, the equation has one negative root and zero, one, or two positive roots, since $u > 0$, $v > 0$. The number of solutions of (6.11) changes from one to three precisely on the curve

$$27b^2 = 4a^3.$$

In terms of u, v this equation is

$$27u^2v^2(u+v)^2 = 2(u^2 + uv + v^2)^3. \tag{6.12}$$

The polynomial $27u^2v^2(u+v)^2 - 2(u^2+uv+v^2)^3$ has three quadratic factors:

$$\begin{aligned} &27u^2v^2(u+v)^2 - 2(u^2+uv+v^2)^3 \\ &\quad = \left(v^2 + 4\,uv + u^2\right)\left(v^2 - 2\,uv - 2\,u^2\right)\left(u^2 - 2\,uv - 2\,v^2\right). \end{aligned}$$

The first factor is positive in the first quadrant, and the other factors have conjugate roots stemming from

$$\frac{v}{u} = \frac{1}{2}\left(\sqrt{3} - 1\right) \equiv r^*,$$

the threshold beyond which we can no longer obtain solutions consistent with (6.1) through (6.5). □

To understand better how the various parameters relate to each other, first we note that $x = h_m$ and $y = h_M$ are interchangeable in (6.8). Without loss of generality, we take $u = h_L = 1$ and rewrite (6.11) as a quadratic equation for v in terms of $y = h_m$ (or h_M):

$$v^2 + v + \frac{y(y^2-2)}{2(2-y)} = 0. \tag{6.13}$$

For $0 < y < \sqrt{2}$, the product of the two roots is negative, so they are real and of opposite sign. In Figure 6, we plot $v = h_R$ on the horizontal axis to clarify that for each choice of h_R, $0 < h_R < r^*$, there are two values of y corresponding to h_m, h_M. From this graph or from (6.13), we see that, as the precursor height $h_R = v \to 0$, we have $h_m \to 0$ and $h_M \to \sqrt{2}$. From (6.3), we find that $G_m \to -\infty$, $G_M \to \frac{\sqrt{2}}{3}$.

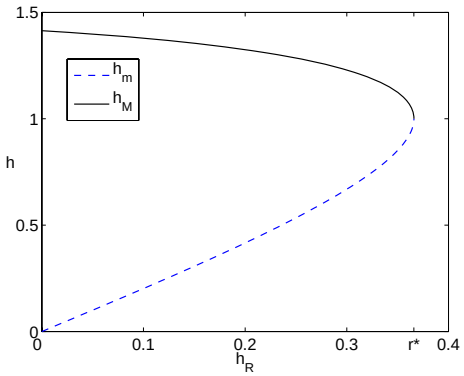


FIG. 6. *Plot of* h_R *versus* h_m, h_M *from* (6.13).

Correspondingly, the speed $c \to \frac{1}{3}$, a value that appears in the proof of Theorem 4.2. That is, if h_M is restricted to lie in the interval $(h_L, \sqrt{2}h_L)$ in Theorem 4.2, then the traveling wave speed is in the interval $\left(\frac{\alpha}{3}h_L^2, \frac{\alpha}{2}h_L^2\right)$. In Figure 6 we observe that h_m, h_M both approach $h = 1$ as $v = h_R$ approaches the threshold $r^* = \frac{1}{2}(\sqrt{3} - 1)$ of Theorem 6.1.

For a film propagating onto a prewetted surface (i.e., with a small precursor height h_R), it might seem reasonable from Figure 6, relating the values of h_M and h_m to h_R, to approximate the height of the step to be twice that of the precursor height. However, for larger precursors this is not a good approximation. The step is always greater than twice the precursor height; in fact, for a given h_R, as h_m approaches $2h_R$, G_M approaches ∞ and the speed of the wave goes to ∞. Moreover, $h_m \to 1$ as $h_R \to r^*$.

In the next section, numerical simulations illustrate the wave structures introduced in the analysis of sections 5 and 6.

7. Numerical simulations II. PDE simulations of (2.3), (2.4) for the inclined substrate illustrate a number of issues presented in the above analysis. First, we explore a combination of linear waves and simple jumps using the analysis of section 6 to choose appropriate initial film profiles. Then we test the stability of linear waves and simple jumps, presented in Theorems 5.4 and 5.6. Next, we explore simulations near the critical threshold r^*. Finally, we vary h_R to observe changes in the wave structures that occur above the threshold.

PDE simulations for linear waves and simple jumps. Equations (6.1)–(6.5) were derived to explain the numerical simulations shown in Figure 4. However, the same equations apply to a combination of linear waves and simple jumps with $\Gamma_L = 0 = \Gamma_R$, all traveling with speed c given by (6.1). The connection between the two structures stems from the limit $\Gamma_L \to 0$ in the nonlinear traveling wave, which does not affect the formulae in (6.1), (6.2).

In the PDE simulations of Figure 7, we choose initial data in which h is piecewise constant and Γ is piecewise linear, consistent with (6.1)–(6.5). The intermediate parameters h_m, h_M, G_m, and G_M annotated in Figure 2 are calculated from (6.6), (6.7), (6.3), and (6.4) for the most common boundary conditions in the simulations,

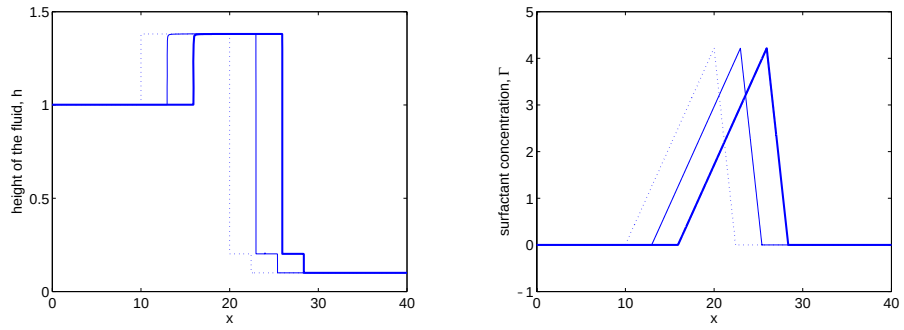


FIG. 7. *PDE simulations of simple traveling waves. Initial profiles are dashed, final profiles are in bold, and plots are separated by* 8 *time units.*

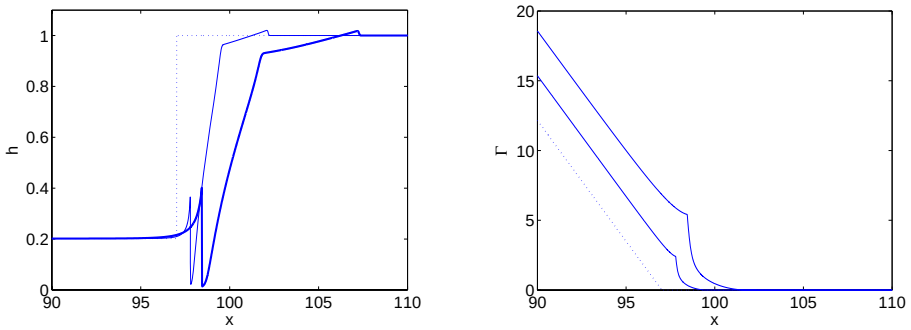


FIG. 8. *Numerical evidence that the linear wave and jump combination of* (5.20) *with* $h_- = 0.2$, $h_+ = 1.0$ *is unstable as in Theorem* 5.4. *Initial data are dotted, and final profiles in bold are at time* $t = 10$.

$h_L = 1.0$, $h_R = 0.1$:

$$h_M = 1.3787; \quad h_m = 0.2019; \quad G_M = 0.4210; \quad G_m = -1.7316. \tag{7.1}$$

The initial data have jumps at $x = 10$, $x = 20$; the initial location of the leading edge of the surfactant is then determined by the data.

In Figure 7, the entire structure moves to the right with speed 0.37 predicted by Theorem 5.4. Note that the maximum value of Γ remains constant, as expected.

Figure 8 provides numerical evidence that the waves conjectured to be unstable in Theorem 5.4 are indeed so. In these plots notice that for the wave emerging from h_L, it appears that $h_x \to \infty$, followed by a jump in the height corresponding to a corner in Γ_x. To the right of the discontinuity, there is a smooth wave preceded by a hydrodynamic wave where there is no surfactant present.

The numerical experiments shown in Figures 9 and 10 explore the wave structure introduced in Theorem 5.6(a). In this case, the positive surfactant gradient creates a linear profile for Γ that travels up the incline and maintains the jump in film height, which can have either sign. The depth-averaged velocity $-\frac{1}{2}hG + \frac{\alpha}{3}h^2$ changes sign across the jump, but the surface velocity $s = -\frac{\alpha}{2}h_+h_-$ is negative, while the jump speed $c = \frac{\alpha}{12}(h_+ - h_-)^2$ is positive. Since $G = \Gamma_x$ is constant, the jump in h is a shock wave solution of (2.3). In general, a jump down satisfies the Lax entropy condition

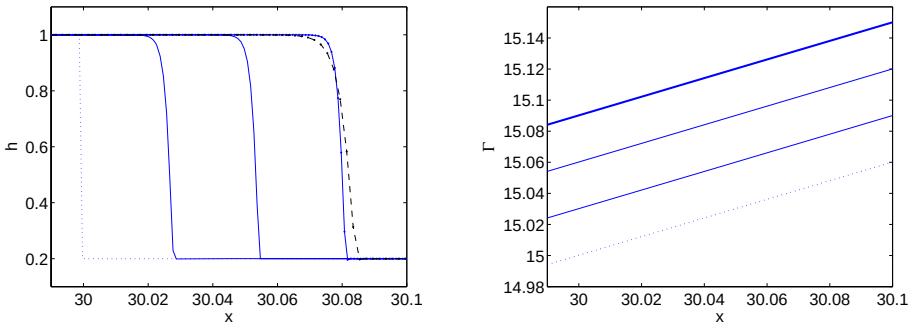


FIG. 9. *PDE simulation of linear wave in* Γ *moving to the left as jump in* h *moves to the right. The initial condition is dotted, and the final profile at time* $t = 1.5$ *is in bold* ($\Delta x = .00025$). *For the height plot, the dashed profile at* $t = 1.5$ *has a more coarse grid* ($\Delta x = .001$); *the consistent location of grid points in the final profile at both mesh sizes and steepening with mesh refinement indicate the presence of a shock.*

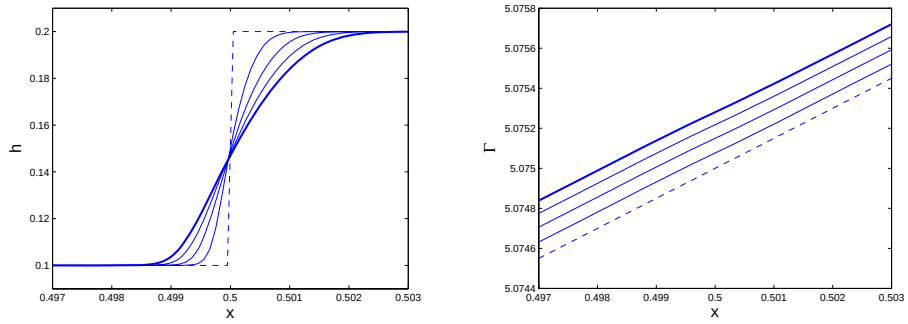


FIG. 10. *PDE simulation of initially linear* Γ *with a jump up in* h *demonstrating the instability of jump-up shocks.*

and is stable, as in Figure 9, while a jump up is unstable, as shown in Figure 10. In the latter case, it is tempting to think of the spreading solution as a rarefaction, with Γ_x constant, but, in fact, the rarefaction solution fails to satisfy the surfactant equation (2.4).

PDE simulations varying h_R. Figure 11 contains the results of numerical simulation of the PDE with $h_L = 1$ and with h_R just above and just below the threshold of the theorem. For $h_R = 0.365$, below the threshold, the step in h is clearly emerging, but for $h_R = 0.367$, above the threshold, the step fails to emerge. To quantify this observation, in Figure 12 we plot the difference in height Δh between the local maximum near the leading shock and the local minimum immediately behind. Plotted as a function of time in the figure, we observe that Δh approaches zero for $h_R = 0.365$, whereas Δh appears to be converging to a positive value for $h_R = 0.367$.

Numerical PDE simulations with $h_L = 1$, $\Gamma_L = 1$, and $\alpha = 1$ and varying values of h_R below the critical threshold $h_R = \frac{\sqrt{3}-1}{2}$ of Theorem 6.1 reproduce the same pattern of waves shown in Figure 4. In Figure 13 we show the results of PDE simulations for selected values of h_R *above* the critical threshold. These wave structures have not been observed previously, since h_R is generally taken to be the thickness of a precursor

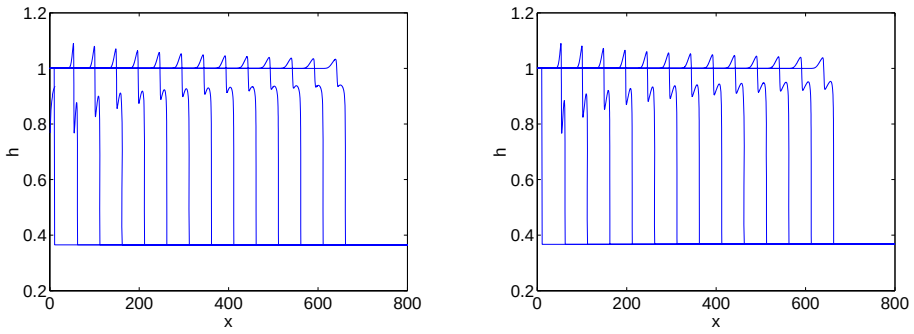


FIG. 11. *Solutions near the transition between solution types.* $h_L = 1$, $h_R = 0.365$ *(left);* $h_L = 1$, $h_R = 0.367$ *(right). Plots are separated by* 100 *time units.*

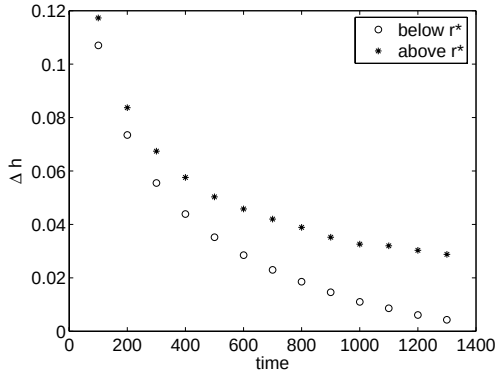


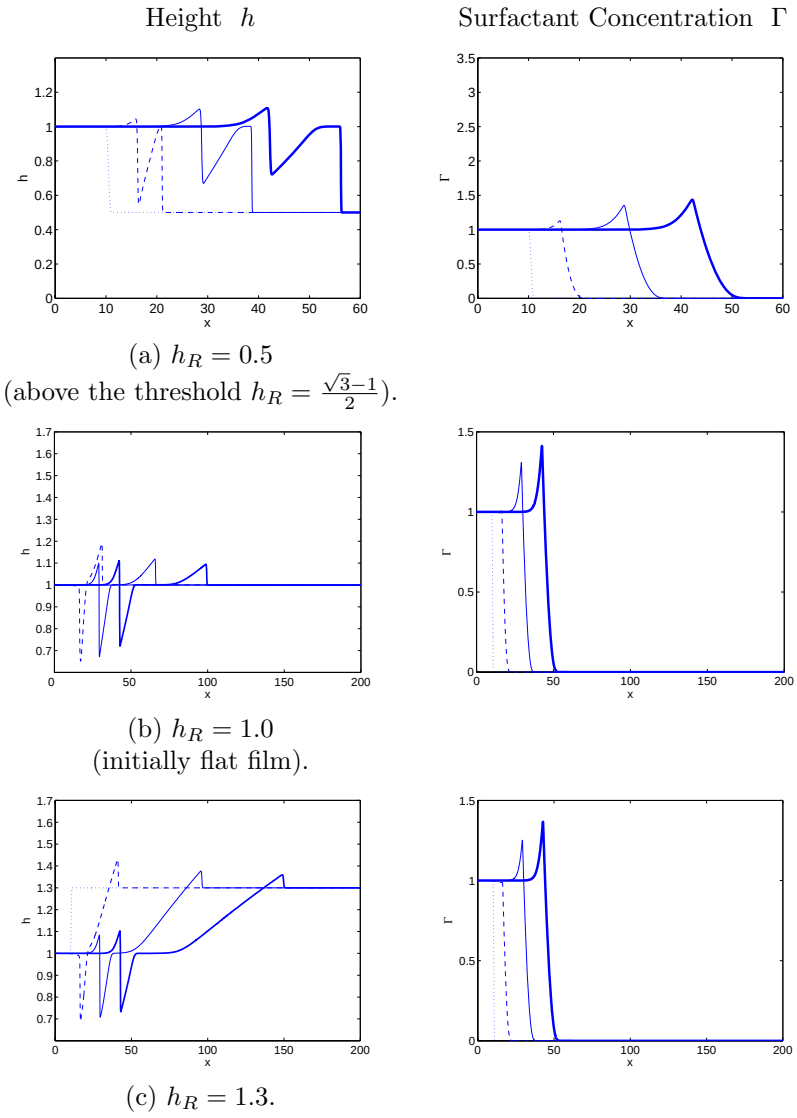
FIG. 12. *Comparison of step heights for the plots in Figure* 11. *As the step forms for the solution below the critical threshold* ($h_R < r^*$)*, the difference* Δh *in maximum and minimum heights within the step goes to zero. Above the threshold* ($h_R > r^*$) *the height difference decreases much more slowly, evidence that no step emerges.*

layer, and therefore much smaller than h_L. In all of the height plots, a distinctive jog in the height develops, which on the incline resembles a Z-shaped wave. We call this new wave a *Z-wave*, in which a shock is between two smooth waves. By comparison, an N-wave of hyperbolic conservation laws consists of a rarefaction bounded by two shocks. The middle height to the right of the Z is the same as h_L, instead of the step height h_m observed below the threshold.

In Figure 13(a), we observe a leading shock (i.e., a jump in h) ahead of the leading edge of surfactant. The shock is a solution of the conservation law

$$h_t + \frac{1}{3}(h^3)_x = 0, \tag{7.2}$$

i.e., (2.3) with $\alpha = 1$ and $\Gamma = 0$. Similarly, in Figure 13(b), the fluid surface is initially flat. The surface tension gradient and gravity force a volume of fluid out ahead of the surfactant, forming a rarefaction wave interacting with a shock, both solutions of (7.2) and decaying in time. For larger values of h_R, as in Figure 13(c), there is also a leading shock being eroded by the rarefaction wave behind it, once again ahead of the surfactant.



(a) $h_R = 0.5$
(above the threshold $h_R = \frac{\sqrt{3}-1}{2}$).

(b) $h_R = 1.0$
(initially flat film).

(c) $h_R = 1.3$.

FIG. 13. *Variation in wave structure for fixed upstream height* $h_L = 1$ *and various downstream heights* h_R. *Each plot has graphs for* $t = 0$ *(dotted line),* $t = 15$ *(dashed line),* $t = 45$ *(thin line), and* $t = 75$ *(bold line).*

The structure of the surfactant profiles is remarkably unchanged over the entire range of h_R. However, Γ appears to reach $\Gamma = 0$ smoothly without the corner observed in simulations with h_R below the critical value.

8. Discussion. The mixed hyperbolic-parabolic system of this paper is derived from the lubrication approximation for the influence of surfactant on flow of a thin liquid film on an inclined plane, neglecting smoothing terms of capillarity and surface diffusion. The analysis of traveling waves and jump conditions leads to the identification of a variety of individual waves.

There are traveling waves in which h and Γ are smooth and nonlinear (Theo-

rem 4.2), and there are three cases in which h is piecewise constant and Γ is piecewise linear:

Neither h nor Γ_x jumps (Theorem 4.1). h is constant and Γ is linear.

Only h jumps (Theorem 5.6(a)). In numerical simulations, we find that jumps down are stable and jumps up are unstable. The stable waves are counterpropagating in that the thin film and surfactant flow up the inclined plane but the jump propagates downwards.

Both h and Γ_x jump (Theorems 5.4 and 5.6). Jumps at $\Gamma = 0$, the leading edge of the surfactant, are related to the step in film height discovered by Borgas and Grotberg [3] for horizontal substrates.

We have constructed an exact solution of the PDE with three jumps that is piecewise constant in h and piecewise linear in Γ and propagates with constant speed. However, this construction is possible only for the ratio of downstream to upstream height below a critical value r^*. This ratio also limits the construction of wave combinations that mimic numerical simulations of initial value problems in which surfactant is supplied from upstream. The supply of surfactant from the boundary has the effect of allowing the maximum surfactant concentration to grow without bound, as shown in Figure 4. This effect is not explained by the analysis, but Taylor series expansions can be used to capture the increase locally in space and time [14].

Above the critical ratio, solutions approach a different structure that we plan to analyze in a future paper. These solutions have a distinctive Z-wave pattern, together with precursor waves propagating into the undisturbed film with no surfactant. As the Z-wave pattern forms, the film height has large dips, which might lead to dewetting. This is a concern in surfactant replacement therapy [13], in which a coating of surfactant is required for healthy lung function. This risk of dewetting does not occur for boundary height ratios below the critical ratio.

While we have discussed stability of waves numerically and to some extent analytically, much remains to be done. It would be interesting to analyze stability of the individual waves for $C > 0$ and $D > 0$. In a future paper, we plan to analyze the constant volume case in which a thin liquid drop on an inclined substrate is spread by the influence of surfactant and gravity.

Acknowledgments. We would like to thank Barry Edmonstone, Richard Craster, and Omar Matar, who introduced us to the problem presented in this paper and who generously shared preliminary versions of their papers. We also thank Andrea Bertozzi and Tom Witelski for helpful suggestions regarding numerics.

REFERENCES

[1] R. Aris, *Vectors, Tensors, and the Basic Equations of Fluid Dynamics*, Prentice–Hall, Englewood Cliffs, NJ, 1962.

[2] A. L. Bertozzi, A. Münch, and M. Shearer, *Undercompressive shocks in thin film flows*, Phys. D, 134 (1999), pp. 431–464.

[3] M. S. Borgas and J. B. Grotberg, *Monolayer flow on a thin film*, J. Fluid Mech., 193 (1988), pp. 151–170.

[4] A. M. Cazabat, F. Heslot, S. M. Troian, and P. Carles, *Fingering instability of thin spreading films driven by temperature-gradients*, Nature, 346 (1990), pp. 824–826.

[5] A. Dominguez, A. Fernandez, N. Gonzalez, E. Iglesias, and L. Montenegro, *Determination of critical micelle concentration of some surfactants by three techniques*, J. Chem. Educ., 74 (1996), pp. 1227–1232.

[6] B. D. Edmonstone, R. V. Craster, and O. K. Matar, *Flow of surfactant-laden thin films down an inclined plane*, J. Engrg. Math., 50 (2004), pp. 141–156.

[7] B. D. Edmonstone, R. V. Craster, and O. K. Matar, *Surfactant-induced fingering phenomena in thin film flow down an inclined plane*, Phys. D, 209 (2005), pp. 62–79.
[8] X. Fanton, A. M. Cazabat, and D. Quere, *Thickness and shape of films driven by a Marangoni flow*, Langmuir, 12 (1996), pp. 5875–5880.
[9] B. J. Fischer and S. M. Troian, *Thinning and disturbance growth in liquid films mobilized by continuous surfactant delivery*, Phys. Fluids, 15 (2003), pp. 3837–3845.
[10] D. Halpern, J. L. Bull, and J. B. Grotberg, *The effect of airway wall motion on surfactant delivery*, J. Biomech. Eng., 126 (2004), pp. 410–419.
[11] D. Halpern, O. E. Jensen, and J. B. Grotberg, *A theoretical study of surfactant and liquid delivery into the lung*, J. Appl. Physiol., 1 (1998), pp. 333–352.
[12] O. E. Jensen, *Self-similar, surfactant-driven flows*, Phys. Fluids, 6 (1994), pp. 1084–1094.
[13] O. E. Jensen and J. B. Grotberg, *Insoluble surfactant spreading on a thin viscous film: Shock evolution and film rupture*, J. Fluid Mech., 240 (1992), pp. 259–288.
[14] R. Levy, *Partial Differential Equations of Thin Liquid Films: Analysis and Numerical Simulation*, Ph.D. thesis, North Carolina State University, Raleigh, NC, 2005.
[15] R. Levy and M. Shearer, *Kinetics and nucleation for driven thin film flow*, Phys. D, 209 (2005), pp. 145–163.
[16] O. K. Matar and S. M. Troian, *Spreading of a surfactant monolayer on a thin liquid film: Onset and evolution of digitated structures*, Chaos, 9 (1999), pp. 141–153.
[17] M. Renardy, *On an equation describing the spreading of surfactants on thin films*, Nonlinear Anal., 26 (1996), pp. 1207–1219.
[18] M. Renardy, *A singularly perturbed problem related to surfactant spreading on thin films*, Nonlinear Anal., 27 (1996), pp. 287–296.
[19] H. Stone, *A simple derivation of the time-dependent convective-diffusion equation for surfactant transport along a deforming interface*, Phys. Fluids A, 2 (1990), pp. 111–112.
[20] S. M. Troian, E. Herbolzheimer, S. A. Safran, and J. F. Joanny, *Fingering instabilities of driven spreading films*, Europhys. Lett., 10 (1989), pp. 25–30.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1610–1631

INTERACTION OF ADVANCING FRONTS AND MENISCUS PROFILES FORMED BY SURFACE-TENSION-GRADIENT-DRIVEN LIQUID FILMS*

P. L. EVANS† AND ANDREAS MÜNCH‡

Abstract. On a tilted heated substrate, surface tension gradients can draw liquid up out of a reservoir. The resulting film thickness profile is controlled by two parameters, which depend on the tilt of the substrate, the imposed temperature gradient, and the thickness of a postulated thin precursor layer. The evolution of this film in time is studied using a lubrication model. A number of distinct behaviors are possible as the substrate tilt angle and other parameters are varied. Recent results for the multiple stationary profiles possible near the meniscus are used, and the interaction of these profiles with the advancing front is examined. We demonstrate how to systematically determine the evolution of the entire film profile from the meniscus to the apparent contact line. This allows a categorization of the range of behaviors for a transversely uniform profile in a two-dimensional parameter space. In addition to capillary fronts and double shock structures, new combinations that arise for certain ranges of large substrate tilt and precursor thickness are described. These include profiles involving rarefaction fans, connecting to either an undercompressive or a classical wave at the advancing front.

Key words. lubrication theory, Marangoni shear stress, capillarity, traveling waves

AMS subject classifications. 76D08, 35Q35, 76A20, 76D45, 35G25, 34E05

DOI. 10.1137/050625242

1. Introduction.

In this paper we consider the time-dependent behavior of a thin liquid film on a tilted heated substrate. Such a film is produced when a temperature difference is imposed along a substrate with one end immersed into a reservoir containing a liquid such as silicone oil, giving rise to a surface tension gradient. The resulting surface shear stress drags liquid up from the reservoir, while gravitational forces act to return liquid to the reservoir. This may give rise to a film of liquid which climbs up the substrate. The evolution of the film above the reservoir has received considerable attention in recent years [11, 12, 16]. An understanding of thin-film flows driven by surface tension gradients is of importance, for instance, in "Marangoni drying" [9] and controlling flows in microfluidic applications [6].

Experiments (e.g., by Schneemilch and Cazabat [14, 15]) reveal that the film tends to advance with a steep front at a contact line, where the liquid-air interface meets the substrate. Previous studies have considered the film behavior in the vicinity of the meniscus and at the advancing front independently of each other. The advancing front can be a simple compressive one or an undercompressive shock as part of a double wave structure [2, 11]. Furthermore, it is known that at the meniscus, multiple film profiles are possible. For a fixed combination of substrate inclination and shear stress, the meniscus can settle into either of basically two different profiles [13]. In

*Received by the editors February 24, 2005; accepted for publication (in revised form) February 23, 2006; published electronically June 9, 2006. This work was supported by the DFG Research Center MATHEON (Project C10) in Berlin.

http://www.siam.org/journals/siap/66-5/62524.html

†Humboldt University of Berlin, Institute of Mathematics, Unter den Linden 6, D-10099 Berlin, Germany (pevans@mathematik.hu-berlin.de).

‡Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Mohrenstraße 39, D-10117 Berlin, Germany (muench@mathematik.hu-berlin.de). This author was supported by a Heisenberg Fellowship.

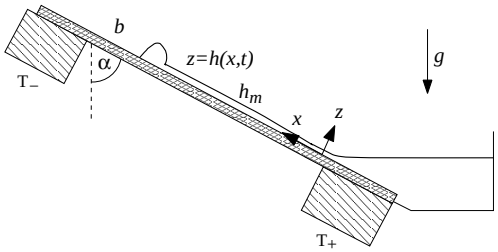


FIG. 1.1. *The thin film on a heated tilted substrate rises from a meniscus. Two heaters hold the temperature at the ends of the substrate at temperatures $T_+ > T_-$. The resulting Marangoni shear stress drives a thin liquid film of thickness $h(x,t)$ up the substrate. At the front, the film advances over a thin precursor layer of thickness b (not shown).*

this paper we determine which meniscus profiles are linked with which traveling wave profiles, and how this link occurs, i.e., which are selected. In the sense that this is a composite description, this work is in the spirit of earlier work by Hocking [7] on the connection between a moving contact line and the meniscus during withdrawal of a moving substrate, which is closely related to the present work. The picture that emerges here for the Marangoni-driven film is, however, more complicated, since we have to include the additional possibility arising from structures involving nonclassical waves in our investigation.

As in our earlier work [13], we consider the arrangement shown in Figure 1.1. The substrate is held at an angle α measured from the vertical, and it is heated so as to impose a uniform temperature gradient $(\mathrm{d}T/\mathrm{d}x) = \gamma < 0$ along the substrate. The film surface tension is σ at some reference temperature, and $\sigma_T = (\mathrm{d}\sigma/\mathrm{d}T) < 0$ is the sensitivity of the surface tension to temperature changes. The shear stress is then $\tau = \gamma\sigma_T$. The film density is ρ, and g is the acceleration due to gravity. We begin section 2 with a statement of the equations governing the film evolution. We then review the evolution of the film at the meniscus and advancing front, considered individually. The former area has been the subject of extensive investigations by us [11, 13] and others, and the latter by Münch, Bertozzi, and others [2, 10].

The dimensionless parameter D was introduced by Bertozzi, Münch, and Shearer [2]. It is a measure of both the substrate angle α and the strength of the surface shear stress driving the flow. Thus

$$D = \left(\frac{3\delta}{\cos^2\alpha}\right)^{2/3} \sin\alpha, \quad \text{where } \delta = \frac{\tau}{2\sqrt{\sigma\rho g}}. \tag{1.1}$$

Bertozzi, Münch, and Shearer [2] define their substrate angle as the complement of α, but their definition of D is equivalent to (1.1). For large inclinations from the vertical (and fixed shear stress) the parameter D is relatively large, and in this case the normal component of gravity is important. A distinct separation between the meniscus and a structure at the advancing front occurs in most circumstances. We build on earlier work, including our preliminary investigation [13], by systematically considering the possible interactions between these structures, and we describe what structures can arise, and the connections between them, in section 3. The combined picture we present is confirmed by dynamical simulations. We also extend previous work by considering the large D limit in section 4. In this case we find that the film can no longer be thought of as separate meniscus and front structures; instead the

film smoothly varies from one to the other without a flat region. The result of our investigations in sections 3 and 4 is a coherent picture of the film behavior as D and the thickness of a presumed precursor layer are varied, which we present in section 5. Finally section 6 summarizes our work.

2. Preliminaries.

2.1. Formulation. We denote the time-dependent film thickness profile by $h(x,t)$, where x measures distance up the substrate and t is time. Using ideas from singular perturbation theory, an evolution equation governing $h(x,t)$ may be obtained [11]. This governing equation is

$$h_t + \Omega_x (h^2)_x - (h^3)_x = -(h^3 \kappa_x)_x + D(h^3 h_x)_x, \tag{2.1}$$

where $\kappa = h_{xx}(1+\epsilon^2 h_x)^{-3/2}$ is the nonlinear expression for the curvature of the free surface. Here $\epsilon = H/L$ (where H and L are given below) is a small parameter, and so $\epsilon \ll 1$. In (2.1), Ω is a dimensionless temperature profile for which $\Omega_x = 1$ except near the heaters, where Ω becomes constant, cutting off surface tension gradients there. Equation (2.1) is obtained by scaling h, x, and t by

$$H = \frac{3\tau}{2\rho g \cos\alpha}, \quad L = \left(\frac{3\sigma\tau}{2\rho^2 g^2 \cos^2\alpha}\right)^{1/3}, \quad T_0 = 2\mu\left(\frac{4\sigma\rho g\cos\alpha}{9\tau^5}\right)^{1/3},$$

respectively. The terms in h^2 and h^3 on the left-hand side of (2.1) account for the competing effects of the imposed shear stress and drainage due to the component of gravity parallel to the substrate. The first term on the right-hand side is due to surface tension, which is supposed to not differ appreciably from its reference value, except inasmuch as it provides the driving shear stress. The second arises from the leveling effect of the component of gravity normal to the substrate. It is useful to define the flux function, $f(h) = h^2 - h^3$, which represents the flux of liquid up the substrate in the absence of the second- and fourth-order smoothing terms in (2.1).

As explained in our earlier work [13], when $\epsilon = H/L = (9\delta^2/\cos\alpha)^{1/3} \ll 1$, it is appropriate to replace κ by the approximate expression h_{xx} in the thin film region away from the reservoir. In addition, when $\epsilon \ll D$ it follows that $\epsilon|h_x| \ll 1$ in the vicinity of the reservoir, and approximating κ by h_{xx} is also appropriate there. We set $\Omega_x \equiv 1$, requiring that as α is increased the position of the heater is moved further into the reservoir, i.e., towards large negative x values. In this way a uniform temperature gradient, and hence uniform shear stress, is imposed. Equation (2.1) then reduces to

$$h_t + (h^2 - h^3)_x = -(h^3 h_{xxx})_x + D(h^3 h_x)_x. \tag{2.2}$$

An appropriate boundary condition at the meniscus is that the film profile flattens out to meet the undisturbed reservoir, so $\partial h^*/\partial x^* \sim -\cot\alpha$ in dimensional units. After rescaling, this yields

$$h \sim -x/D \qquad \text{as } x \to -\infty. \tag{2.3}$$

To avoid the singularity associated with a moving contact line, we adopt a precursor model so

$$h \to b \qquad \text{as } x \to \infty \tag{2.4}$$

and define the apparent contact line to be the point where the film thickness first becomes approximately b.

We are concerned with the behavior of solutions of (2.2) subject to (2.3) and (2.4). Solutions of (2.2)–(2.4) are potentially influenced by just two parameters, D defined by (1.1) above and the precursor layer thickness, b. A particular combination of these parameters determines the structure of the climbing film, including the meniscus. In the following sections, we enumerate and describe the possible film structures.

Solutions of (2.2) typically have two distinct parts, a meniscus and a wave structure consisting of one or two advancing waves. Near the reservoir the meniscus part settles into an equilibrium solution with thickness approaching some value h_m. Immediately behind the moving apparent contact line, where the film abruptly decreases to the precursor, is a traveling wave, which we refer to as the "advancing front." Behind it may be additional waves that, together with the advancing front, make up the moving wave structure.

A good guide to the possible behavior of the complete film comes from considering what happens in the two parts independently. In the remainder of this section, we first describe the waves near the contact line. Here the resulting wave structure is determined by a left thickness h_w, together with D and b. We then (in section 2.3) summarize what limiting meniscus thicknesses h_m are possible for a given D. We assume that the precursor thickness b is the smallest thickness scale that appears in the film profile, and so we consider only cases where $h_w > b$.

2.2. Advancing front behavior. The wave dynamics that arise at and behind the rising contact line have been investigated in detail in recent years [1, 2, 10] by considering solutions of (2.2) with initial data that connect flat left and right states where the film profile thicknesses are h_w and b, respectively. To capture the large scale structure of the emerging waves, we rescale $x = x^*/\lambda$ and $t = t^*/\lambda$ in (2.2) and let $\lambda \to 0$. We find that (2.2) is a nonlinear perturbation of the scalar conservation law (dropping the asterisks),

$$h_t + [f(h)]_x = 0, \qquad f(h) \equiv h^2 - h^3. \tag{2.5}$$

We first discuss the situation when (2.5) is considered as the limit of the problem with exclusively nonlinear second-order diffusion, i.e., of the equation

$$h_t + [f(h)]_x = (h^3 h_x)_x. \tag{2.6}$$

One quickly finds that traveling wave solutions of (2.6) with left and right far-field conditions $h_w > b$ and b, respectively, exist precisely if $h_w \le h_{\rm int} \equiv (1-b)/2$. In the limit of vanishing second-order diffusion, these traveling waves correspond to shock solutions of (2.5), i.e., jump discontinuities

$$h(x,t) = \begin{cases} h_w & \text{if } x < st, \\ b & \text{if } x > st, \end{cases} \tag{2.7}$$

that move according to the Rankine–Hugoniot condition,

$$s = s(h_w, b) = (f(h_w) - f(b))/(h_w - b). \tag{2.8}$$

If, however, $h_w > h_{\rm int}$, a double wave structure forms, which corresponds to a rarefaction-shock solution of (2.5),

$$h(x,t) = \begin{cases} h_w & \text{if } x \le f'(h_w)t, \\ f'^{-1}(x/t) & \text{if } f'(h_w)t \le x \le f'(h_{\rm int})t, \\ b & \text{if } x \ge st, \end{cases} \tag{2.9}$$

where the speed $s = s(h_{\rm int}, b)$ of the shock between the intermediate height $h_{\rm int}$ and b also satisfies $s = f'(h_{\rm int})$.

This solution structure is consistent with the classical theory of conservation laws, which admits only Lax (also called compressive) shocks in solutions of (2.5), i.e., shocks that satisfy the Lax entropy condition

$$f'(b) < s < f'(h_w), \tag{2.10}$$

which stipulates that characteristics from both states must cross the shock trajectory, or at most the limits of Lax shocks, sometimes called *generalized Lax* shocks, where one of the inequalities is replaced by an equality. The shock in (2.9) is an example of a generalized Lax shock.

Returning to the situation where (2.5) is considered to be the limit problem of (2.2), we find that new combinations of waves that include nonclassical undercompressive shocks which violate (2.10) arise for $D = 0$, i.e., pure fourth-order diffusion, and for a range of $D > 0$. These waves have been studied in detail for $D = 0$ by Bertozzi, Münch, and Shearer [2] and for general D by Münch [10]. This investigation was carried out via numerical simulations of (2.2) and by systematic determination of the traveling wave solutions $h(x,t) = h(\xi)$, $\xi = x - st$ with far-field states h_- as $\xi \to -\infty$ and b as $\xi \to \infty$. Inserting this ansatz into (2.2), we obtain, after integrating once and using the far-field condition, the third-order ODE

$$h''' = -Dh' + f(h) - f(b) - s(h - b), \qquad ' \equiv {\rm d}/{\rm d}\xi, \tag{2.11}$$

with the wave speed $s = s(h_-, b)$ given by (2.8). This ODE was systematically investigated via phase space methods [2, 3, 10]. Traveling wave profiles appear as heteroclinic connections between the equilibria of (2.11), formed by the intersections of invariant manifolds. These can be computed numerically by following individual trajectories on the manifolds, and the intersections of invariant manifolds can be tracked (as b is varied) in Poincaré sections.

We now summarize the results of these studies on the solution of (2.2) connecting h_w and b. For $D = 0$ and a fixed value for $b < 1/3$, four different situations arise as $h_w > b$ is increased. For $b < h_w < h_{ii}(b)$, with a b-dependent upper bound, a compressive wave arises. In contrast to the case of second-order diffusion, which smooths the shock to a monotonic profile, the fourth-order diffusion induces a capillary ridge (see Figure 2.1, lower left corner). Furthermore, in a thin region $h_i(b) < h_w < h_{ii}(b)$, multiple traveling wave solutions of (2.2) exist that correspond to the same compressive shock. These waves differ by the width of their capillary ridge. Furthermore, for $1 - h_{\rm uc} - b < h_w < h_{ii}(b)$, double shocks composed of a Lax and an undercompressive shock are also possible; $h_{\rm uc}$ and undercompressive shocks are explained below. However, for $h_w < h_{ii}(b)$, monotonic initial data for (2.2) typically gives rise to the compressive wave with the smallest capillary ridge. For $h_{ii}(b) < h_w < h_{\rm uc}(b)$ a double shock results,

$$h(x,t) = \begin{cases} h_w & \text{if } x \le s(h_w, h_{\rm uc})t, \\ h_{\rm uc} & \text{if } s(h_w, h_{\rm uc})t \le x \le s(h_{\rm uc}, b)t, \\ b & \text{if } x \ge s(h_{\rm uc}, b)t, \end{cases} \tag{2.12}$$

where the trailing shock that connects the states h_w and $h_{\rm uc}$ is a Lax shock, and the leading shock from $h_{\rm uc}$ to b is undercompressive; i.e., it violates the Lax entropy condition (2.10) in that characteristics only enter from the right. The undercompressive

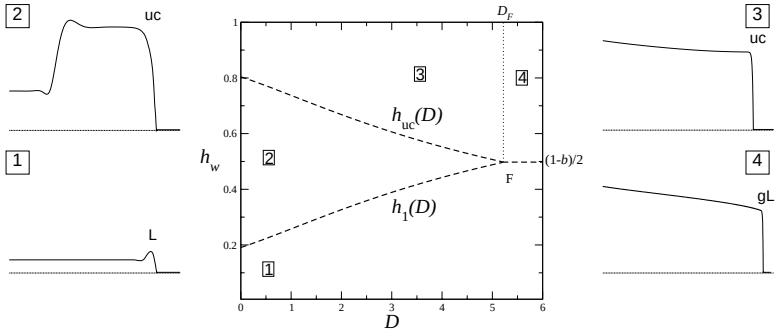


FIG. 2.1. *The front wedge diagram for $b = 0.005$ showing the four possible types of advancing front behavior, depending on D and the left state value h_w. Surrounding subfigures show qualitatively the behavior observed in each of regions 1 to 4: region 1—compressive front; 2—double shock structure (compressive and undercompressive waves); 3—rarefaction fan and undercompressive front; 4—rarefaction fan and generalized Lax shock. These regions are described further in the text. Labels indicate the type of front: L—Lax shock; uc—undercompressive shock; gL—generalized Lax shock.*

wave arises as a codimension one intersection of invariant manifolds for (2.11) and for a given b (and D) appears only for a specific value of the left state, $h_{\rm uc}$. (In contrast, compressive waves arise as codimension zero intersections.) In practice, a shooting method allows easy computation of $h_{\rm uc}(D, b)$ and hence the shape of the wedge. The trailing wave moves at a slower speed, so that the width of the plateau separating the two waves grows in time. The profile of a numerical solution of (2.2) corresponding to such a double shock is shown in the upper left corner of Figure 2.1. The emergence of nonclassical undercompressive shocks as one passes a certain threshold for h_w has been interpreted as "nucleation" [8].

For $h_w > h_{\rm uc}$, we get a rarefaction-undercompressive shock wave combination,

$$h(x,t) = \begin{cases} h_w & \text{if } x \le f'(h_w)t, \\ f'^{-1}(x/t) & \text{if } f'(h_w)t \le x \le f'(h_{\rm uc})t, \\ h_{\rm uc} & \text{if } f'(h_{\rm uc})t \le x \le s(h_{\rm uc}, b)t, \\ b & \text{if } x \ge s(h_{\rm uc}, b)t. \end{cases} \tag{2.13}$$

Note that since the shock is undercompressive, i.e., $f'(h_{\rm uc}) < s(h_{\rm uc}, b)$, it separates from the leading edge of the trailing rarefaction wave. The upper right corner of Figure 2.1 shows the corresponding solution of the full PDE (2.2).

The results for general values of D as described by Münch [10] may be summarized in a two-dimensional diagram that displays, for each D, the values of left states h_w where the different types of wave or waves connecting this state to the precursor right state b exist. In the (D, h_w) plane, the boundaries between these different ranges essentially form a wedge-like shape, shown in Figure 2.1 for the precursor thickness $b = 0.005$. This shape is defined by the graphs of $h_{\rm uc}(D, b)$ (at the upper edge) and $h_1(D, b) = 1 - h_{\rm uc} - b$ (lower edge). We refer to this shape as the "front wedge."

Because $h_{\rm uc}$ depends on b, the shape and position of the front wedge on a (D, h_w) diagram also depends on b. The apex of the wedge (shown as F in Figure 2.1) is located at $(D_F, (1-b)/2))$, where D_F is itself a monotonically decreasing function of b. The significance of this will become apparent in section 3. The two sides of the wedge, together with the extension of the wedge's apex $\{(D, h) : D > D_F,\ h = (1-b)/2\}$ and the line $\{(D, h) : D = D_F,\ h > (1-b)/2\}$ divide the plane into four regions,

labeled 1–4 in the figure. Which wave structure results from a particular (D, h_w) pair is indicated by the region in which the pair belongs, as follows:

1. For relatively small values of h_w, corresponding to region 1, a simple compressive wave arises. This wave has a capillary ridge that diminishes as D increases.
2. Within the wedge (region 2), one obtains the compressive-undercompressive double shock structure described above.
3. For large values of h_w, one obtains double wave structures involving a rarefaction fan; these are either a rarefaction-undercompressive wave when $D < D_F$ and $h_w > h_{\rm uc}$ (region 3), or
4. a rarefaction-shock wave (with a leading generalized Lax shock), for larger D (region 4).

The sketches for regions 3 and 4 in Figure 2.1 differ in the profile behind the leading undercompressive wave. In the former, the rarefaction wave separates from the leading shock, giving rise to a convex portion behind the front, while the latter profile is concave right up to the shock. In region 3, the rarefaction wave separates from the leading shock, but not in region 4. To the right of the apex ($D > D_F$) there are no structures involving undercompressive waves: above $(1 - b)/2$ there is a rarefaction-shock wave combination, while below this line a simple compressive wave occurs. Existence of undercompressive shock solutions for D less than some finite upper bound, and nonexistence for sufficiently large D, was proved by Bertozzi and Shearer [3].

We have limited the above discussion to values of $h_w > b$. In this study, the initial film profiles used have thickness $h \geq b$ everywhere, and so the thickness at later times is never much smaller than b.

In addition, we have simplified matters by neglecting the presence of a thin region of thicknesses h_w located around the lower side of the wedge. This exists for a range of D close to zero, where multiple wave structures that connect h_w to b are possible, either one of a number of compressive waves or the double shock structure. This range ends towards the right at a value D (say D_1) below two [10]. Which wave structure is selected in a numerical simulation depends on the initial profile; for monotonic initial data connecting h_w and b, a single compressive wave with the smallest capillary ridge typically arises [1, 2]. Hence for such initial data and $D < D_1$, the range of h_w for which we get a single compressive front is slightly increased above the lower edge of the wedge.

2.3. Meniscus structures. At the meniscus, the profile should become stationary after the contact line has moved away, up the inclined substrate. Hence, of primary interest here are the stationary solutions of (2.2) with far-field condition (2.3) as $x \to -\infty$ and a flat right state $h \to h_m$ for $x \to \infty$. Letting $h_t = 0$ and integrating once yields the ODE

$$h_{xxx} = Dh_x - (f(h) - f(h_m))/h^3. \tag{2.14}$$

Note that the constant of integration $Q = f(h_m)$ represents the total flux through the flat film. In earlier work [13] we investigated this boundary value problem by numerically exploring the relevant invariant manifolds in the three-dimensional phase space associated with (2.14). Solutions satisfying the far-field conditions as $x \to -\infty$ form a two-dimensional invariant manifold, while those that tend to a constant film thickness h_m as $x \to \infty$ form either a one- or a two-dimensional (stable) invariant manifold

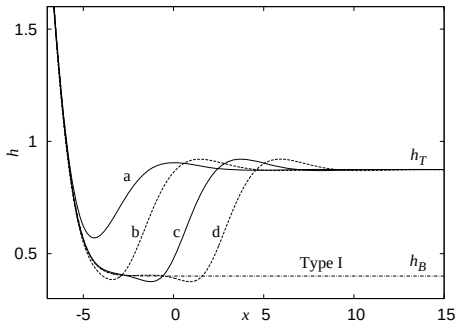


FIG. 2.2. *The first two stable and first two unstable Type* II *steady meniscus profiles, for* $D = 0.322$. *The profiles labeled a and c are stable, while b and d are unstable. Here* $h_m = h_T(0.322) = 0.8744$, *so there are an infinite number of these profiles. Also shown is the Type* I *meniscus profile for this value of D; it has* $h_m = h_B(0.322) = 0.4002$.

when either $h_m < 2/3$ or $h_m > 2/3$, respectively. Hence the stationary meniscus profiles correspond to either codimension one or codimension zero intersections of the invariant manifolds, respectively, and we call these Type I meniscus solutions in the former, and Type II meniscus solutions in the latter, case. While Type I solutions monotonically decrease in thickness as they approach h_B, Type II solutions typically have a characteristic depression or dimple in the meniscus region, particularly for sufficiently small values of D, followed by damped oscillations as $x \to \infty$ (Figure 2.2). The dimple and the oscillations can be suppressed by the normal component of gravity, i.e., for larger values of D.

The Type I solutions are structurally unstable and, for given D, exist only for discrete values of h_m. In fact, for a range of values $D < D_M = 0.8008$, numerical exploration of the phase space [13] shows that a unique Type I solution exists for a single thickness $h_m = h_B(D)$, while there are none at all for $D > D_M$. In the range $D > D_M$, Type II meniscus profiles exist for all $h_m > 2/3$, while for $D < D_M$, the situation is more complicated. We define $h_T(D)$ to be the larger positive root h of the cubic equation $f(h) = f(h_B)$, so that $h_T > 2/3$. Then Type II solutions can always be found for $h_m \geq h_T(D)$. In addition, for $D < D_B = 0.7142$, Type II solutions exist for a range of thicknesses h_m slightly below $h_T(D)$, down to a value $h_*(D)$. Moreover, when h_m is near $h_T(D)$, in the range $h_*(D) < h_m < h_{**}(D)$, with a D-dependent value $h_{**}(D) > h_T(D)$, multiple Type II solutions appear with the same thickness h_m; their profiles differ by the depth and width of the dimple, as shown in Figure 2.2. The profiles of four of the multiple Type II menisci are shown in Figure 2.2 for $D = 0.322$ and $h_m = h_T(0.322) \approx 0.8744$. (The corresponding Type I profile for this value of D is also shown.) The question of which, if any, of the Type II menisci are stable to in-plane disturbances then arises. Numerical simulations of the time-dependent PDE (2.2) revealed that these meniscus solutions are alternately stable and unstable. These simulations were initialized using the Type II profiles computed by a shooting method [13]. The unstable solutions to the ODE do not occur as solutions of the PDE at long times; instead, initial conditions which are close to these evolve toward the stable solutions. In Figure 2.2 the stable and unstable solutions are shown as solid and dashed lines, respectively.

Figure 2.3 summarizes the solution structure, showing the thicknesses $h_B(D)$ and $h_T(D)$ controlling Type I and Type II solutions in the (D, h) plane. These values

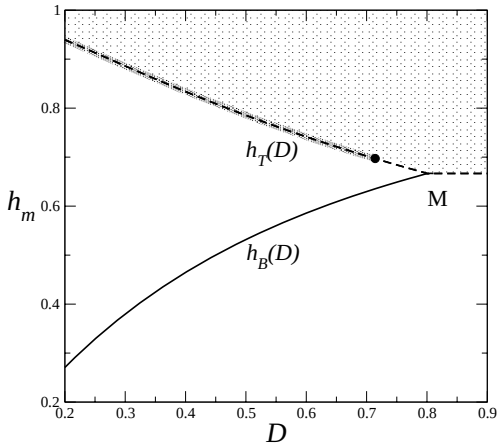


FIG. 2.3. *The meniscus wedge diagram. Lines show the allowable values for the right state of the meniscus, h_m. For $D < D_M$, the meniscus may approach a thickness h_m, where either $h_m = h_B(D)$ is given by the lower branch (a Type* I *solution) or h_m lies above the upper branch, i.e., $h_m > h_T(D)$ (a Type* II *solution). The dark shaded line indicates where multiple Type* II *menisci are possible for a range of values h_m above and below $h_T(D)$. For $D > D_M$, Type* I *solutions are not possible, but Type* II *solutions exist for all h_m above* 2/3. *The light shaded region shows where one or more Type* II *menisci are available.*

merge with $h = h_T = h_B = 2/3$ when D approaches D_M, at the point labeled M in the figure. We call the structure formed by the graphs of $h_T(D)$ and $h_B(D)$ the "meniscus wedge." Note that it does not depend on b. One or more Type II solutions exist at points in the light shaded region. The range of thicknesses with multiple Type II solutions is not precisely shown in Figure 2.3, but is indicated there by the dark shaded line along $h_T(D)$, ending in a solid dot in the figure at $D = D_B$.

3. Interaction of meniscus and front dynamics. The information summarized in section 2.2 and encapsulated in the front wedge diagram gives a fairly complete picture of which wave, or combination of waves, arises near the contact line if, for given b and D, the leftmost value of the film thickness is set at some specific value, h_w. The question of how this value is selected then arises; it is evident that the meniscus plays an important role here. If at long times, when the contact line has traveled far from the reservoir, the meniscus profile approaches a steady state, then the value of h_w must be equal to an h_m for which either a Type I or Type II solution exists. This information is found in the "meniscus wedge" diagram in section 2.3. An overview of the possible combinations of the different types of meniscus and wave structures can be obtained by superimposing the two wedge diagrams. In many cases, this suggests more than one possible outcome for a given D and b. For example, if $D < D_M$, wave structures can be found to connect either to a Type I meniscus or to a whole range of Type II menisci. However, the only situations which can arise dynamically are those for which the wave part next to the meniscus has a nonnegative speed. (If its speed were negative, such a wave part could never emerge from the meniscus.) Rarefaction waves, or parts of rarefaction waves, move with a wave speed given by characteristics, namely $f'(h)$, where the film thickness is h. For shock profiles, with left and right states h_- and h_+ the wave speed s is given by the Rankine–Hugoniot condition, $s = (f(h_+) - f(h_-))/(h_+ - h_-)$.

In this section we use such considerations to determine which meniscus and film

profile eventually evolves from monotonic initial data representing a thin precursor layer on a substrate which is partially immersed into the reservoir. The approach outlined above nearly always allows us to single out one possible scenario. The exceptions will be pointed out further below. We then verify our predictions using time-dependent simulations of (2.2). To obtain numerical solutions we use finite difference schemes on a finite spatial domain, $[0, L]$. At the left-hand boundary, we specify $h = H_0$ to be a large constant (typically 20–50) and impose $h_{xxx} = 0$. At the right-hand boundary, $h_x = h_{xxx} = 0$. Solutions are advanced in time using an implicit Euler scheme. The time step is controlled using a step-doubling approach.

Simulations began from an initial profile $h(x, 0) = h_0(x)$. The form used for h_0 was generally

$$h_0(x) = \begin{cases} D^{-3/2}\left(\exp(D^{1/2}x) - D^{1/2}x - 1\right) + b & \text{for } x \le 0, \\ b & \text{for } x > 0. \end{cases} \tag{3.1}$$

This represents the static meniscus which arises through the balance of mean surface tension and gravity for $x < 0$, and joins smoothly to the precursor layer at $x = 0$. Other initial profiles, including the function

$$h_0(x) = \frac{\log(2\cosh(ay)) - ay}{2a} + b, \tag{3.2}$$

where $y = x/D - 20$ and $a = 0.4$, were also used. This has slope -1 for $x \to -\infty$, and has $h_0 \to b$ as $x \to \infty$. We also used

$$h_0(x) = \max\left(-x/D, b\right).$$

The particular choice did not alter the qualitative behavior of the film.

As explained in section 2.2, the apex of the front wedge at $D = D_F$ moves towards smaller values of D as b is increased. Following Bertozzi, Münch, and Shearer [2], b is restricted to be less than $1/3$. Depending on b, the front wedge and meniscus wedge can therefore overlap in four characteristic arrangements, leading to four different cases. In increasing order of b, these are as follows:

A. The most important case is for small b, i.e., a very thin precursor layer. In this case D_F is large, and $h = (1-b)/2$ is close to its maximum value of $1/2$. The upper part of the front wedge $h_{\rm uc}(D)$ makes intersections with both $h_T(D)$ and $h = 2/3$. This is the arrangement which results when $b = 0.005$; it is shown in Figure 3.1(a). Here a line marked with circles indicates the left state of the advancing front h_f, while the right state of the meniscus $h_m = h_w$ is shown by crosses, for each value of D. It continues until $h_{\rm uc}(D_M, b) = 2/3$; this happens for $b = 0.0202$ (to four decimal places).

B. For larger b, the line $h_{\rm uc}(D)$ makes only one intersection with the meniscus wedge, and this is now along the lower branch $h_B(D)$. Figure 3.1(b) shows the situation for $b = 0.05$. At $b = 0.1484$, the apex of the meniscus wedge and that of the front wedge are at the same value of D, i.e., $D_F = D_M$.

C. In the third case (for $0.1484 < b < 0.2338$ to four decimal places), $D_F < D_M$, but $h_B(D_F) > (1-b)/2$, so the graph of $h_{\rm uc}(D)$ still intersects $h_B(D)$. The significance of this is explained below. Figure 3.1(c) shows this case when $b = 0.16$.

D. For the largest b ($b > 0.2338$ to four decimal places) D_F is small enough that the merger at $D = D_F$ happens with $h = (1-b)/2 > h_B(D)$, i.e., above and to the left of the line $h_B(D)$ (see Figure 3.1(d) for $b = 0.25$).

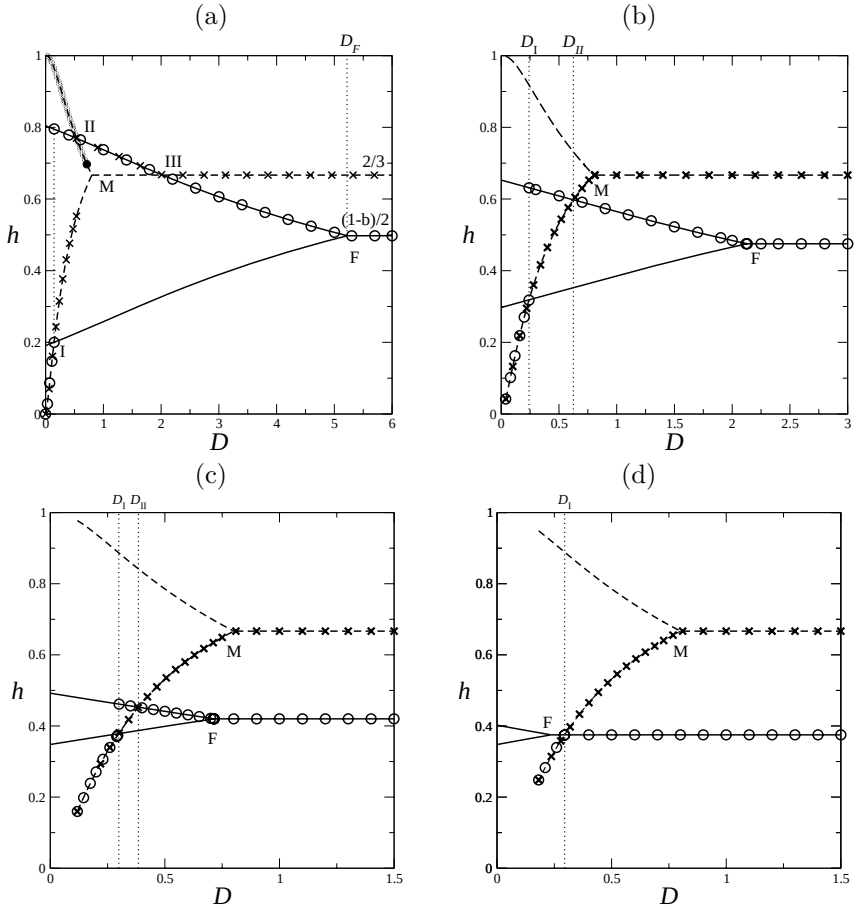


FIG. 3.1. *The (D, h) diagrams, showing the front wedge (solid lines) and meniscus wedge (dashed lines). A line marked with circles indicates the left state, h_f, of the advancing front. A line with crosses indicates the right state of the meniscus, h_m. When these lines coincide, there is a flat film region directly connecting the meniscus to the advancing front.* (a) *Case A, when $b = 0.005$;* (b) *Case B for $b = 0.05$. In* (b)*, a new profile featuring a Type* I *meniscus connected to a rarefaction fan becomes possible for D between $D_{\rm II}$ and D_M.* (c) *Case C, $b = 0.16$;* (d) *Case D, $b = 0.25$. In Case C, $D_F < D_M$, requiring a connection between h_B and $(1 - b)/2$ for $D_F < D < D_M$. In Case D, D_F is so small that the front wedge does not intersect h_B at all.*

3.1. Case A: Thin precursor layer. We first define $D_{\rm I}(b)$ to be the solutions of

$$h_B(D) = h_1(D, b) = 1 - b - h_{\rm uc}(D,\, b);$$

thus, for a given value of b, $(D_{\rm I}, h_B(D_{\rm I}))$ is the point in the (D, h) diagram where the lower sides of the two wedges intersect. We also define $D_{\rm II}(b)$ and $D_{\rm III}(b)$, denoting the solutions of

$$h_T(D) = h_{\rm uc}(D,\, b) \quad \text{and} \quad h_{\rm uc}(D, b) = \frac{2}{3},$$

respectively. These define the intersection of the upper side of the front wedge with the upper side of the meniscus wedge or the line $h = 2/3$, respectively, and are

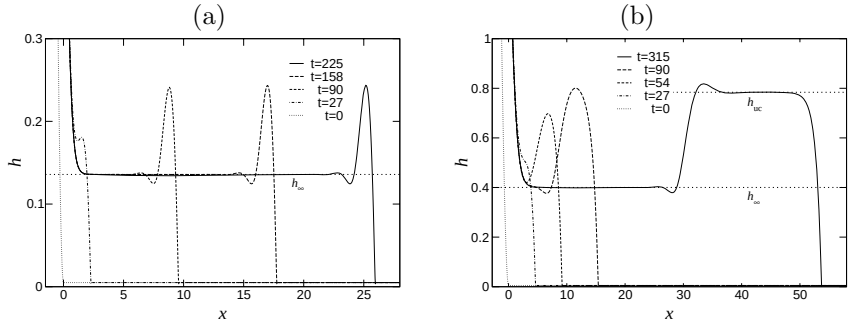


FIG. 3.2. *Numerical solutions of* (2.2) *when* (a) $D = 0.1021$ *and* (b) $D = 0.322$. *Film profiles are shown at dimensionless times given in the legends. In* (a) *there is a single compressive wave, while in* (b) *an undercompressive-Lax double shock develops and travels up the substrate. There is a thin precursor film of thickness* $b = 0.005$ *in both cases.*

shown as points II and III on Figure 3.1(a). When $b = 0.005$, the numerical value for $D_{\rm I} = 0.1461$. The other special values of D given in this section are also for $b = 0.005$.

We begin with the smallest D (but sufficiently large that $h_B > b$) and argue that the meniscus must be of Type I there. First, consider values of $D < D_{\rm I}$. Suppose that the meniscus is of Type II, with a right state thickness $h_m \geq h_T$. In this range of D, h_T in turn is larger than $h_{\rm uc}$. A connection from h_m to the precursor would involve a rarefaction fan followed by an undercompressive wave joining to b. However, the left part of the rarefaction fan would have negative speed, and therefore would fall back into the meniscus. Hence such a solution cannot persist.

On the other hand, if the meniscus is of Type I, then $h_m = h_B < h_1$, and a simple compressive connection to the precursor is possible. This is connected by a flat film to a steadily advancing front. The left state of the advancing front and the right state of the meniscus are identical in this case, since the two are directly connected. Our dynamical simulations for $D = 0.1021 < D_{\rm I}$ and $b = 0.005$ (Example 1 in [13], also shown in Figure 3.2(a)) confirm that this combination of Type I meniscus and a simple compressive wave occurs. The flat region thickness in this case is controlled by the meniscus.

When D is increased above $D_{\rm I}$, the graph of $h_B(D)$ enters the undercompressive region of the front wedge. In place of a simple compressive connection, a Type I meniscus must now connect to a double shock structure. A Type II meniscus is still not possible, for the same reason as in the previous case, namely that $h_T > h_{\rm uc}$ while $D < D_{\rm II}$. Now the left state of the advancing front is the undercompressive wave height $h_{\rm uc}$. The flat region ahead of the meniscus, with thickness h_B, is connected to $h_{\rm uc}$ by the trailing compressive part of the double wave structure. This trailing shock moves upwards, but somewhat slower than the advancing front. In the (D, h) diagram, Figure 3.1(a), the line marked by circles jumps to $h_{\rm uc}$ at $D = D_{\rm I}$, separating from the line portion emphasized by crosses.

In section 2.2 it was noted that when D is small, compressive waves exist for $h_w < h_{ii}$, where h_{ii} is slightly larger than h_1. For $h_1 < h_w < h_{ii}$, the meniscus can be connected to either a double shock structure or to one of several compressive waves. These all have positive wave speeds, and which one is selected depends on the initial data. Due to the experience with monotonic (jump) initial data, we expect that the simplest compressive wave is selected if $h_B < h_{ii}$. The net effect of this is that, for the initial profiles considered in this paper, the transition from a compressive front to

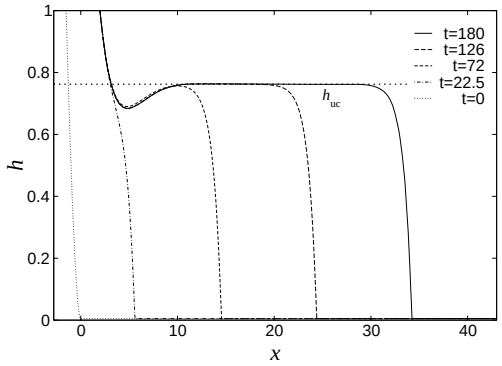


FIG. 3.3. *Evolution of* (2.2) *for* $D = 0.6424$ *at dimensionless times given in the legends. Only the undercompressive shock propagates away from the meniscus, which evolves into a Type* II *meniscus. Again the precursor layer thickness is* $b = 0.005$.

a double shock is delayed and occurs for a value D_I' slightly larger than D_I. Indeed, for $b = 0.005$ we find that this transition occurs for D between 0.16 and 0.18, instead of precisely at $D_{\mathrm{I}} = 0.1461$.

This behavior continues until $D = D_{\mathrm{II}}$, the value of D at which $h_{\mathrm{uc}} = h_T$ and the upper sides of the two wedges cross. For $b = 0.005$, D_{II} equals 0.535. A double shock structure moving up the substrate is shown for $D = 0.322$ in Example 2 of [13], and also in Figure 3.2(b) just as we described it here.

For $D_{\mathrm{II}} < D < D_{\mathrm{III}}$, h_{uc} is larger than h_T, and so it is in the region where Type II meniscus solutions are possible. Hence a direct undercompressive shock connection from a Type II meniscus to the precursor is possible. (The existence of multiple Type II solutions for h near h_T means that a Type II solution is available for matching to h_{uc} via a direct connection at values of D slightly below D_{II}.) These continue until $D = D_{\mathrm{III}}$; for $b = 0.005$, $D_{\mathrm{III}} = 2.025$. For $D < D_{\mathrm{III}}$, h_{uc} is larger than 2/3. Thus we can rule out connections involving intermediate waves as follows. Only shocks can connect to h_{uc} from below (since characteristics for the left and right state would cross, ruling out a rarefaction fan), and these would have a negative speed. Similarly, any wave connection from above must be a rarefaction fan, all parts of which would also have a negative speed.

As a result, the only structure possible is a Type II meniscus connecting directly to a flat state with thickness h_{uc}. This flat state is the left state of an undercompressive shock connection to the precursor. The right state of the meniscus and the left state of the advancing front are again identical, and in Figure 3.1(a) the lines marked by crosses and circles coincide. It is notable that in this range of D the thickness of the flat region, h_{uc}, is determined by the precursor thickness, not by the meniscus. We therefore refer to these structures as "front controlled." This situation in this range is exactly what is observed for Example 3 from [13] and in Figure 3.3, where $D = 0.6424$.

Once again, this description has to be slightly amended. The reason is that, for $D < D_B = 0.7142$, Type II menisci exist even for h_m below h_T (down to the value h_* introduced in section 2.3). Hence, in principle, they can arise and connect directly to an advancing front for values slightly smaller than D_{II}, as an alternative to a double shock structure rising from a Type I meniscus. Furthermore, in the range from h_* to $h_{**} > h_T$, mentioned in section 2.3, different Type II meniscus solutions sharing the

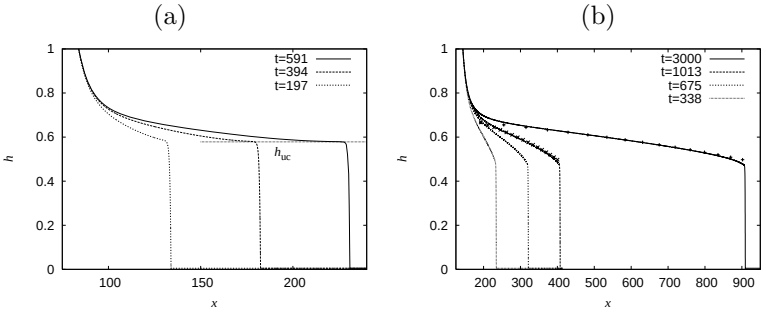


FIG. 3.4. (a) *Evolution of* (2.2) *when* $D = 3.5$ *and* $b = 0.005$*, so* $D < D_F$*. The front is undercompressive. A small flat region of thickness* $h_{\rm uc}$ *(indicated by a dotted line) gradually develops behind the front.* (b) *With* D *increased to* 6*, larger than* D_F*, the front is a generalized Lax shock. The rarefaction fan is also shown at* $t = 3000$ (+ *symbols*) *and* $t = 1013$ (×)*; it was obtained by solving the scalar hyperbolic conservation law as described in detail in section* 4.

same value for h_m exist. Hence, in the thin range of values of D for which $h_{\rm uc}(D, b)$ lies between the graphs of $h_*(D)$ and $h_{**}(D)$, the combination of the wedge diagrams and the criterion for the wave speeds is not sufficient to predict the film behavior. Instead behavior can be determined by, e.g., numerical simulations. We will not elaborate further on this subtlety.

For the small values of D considered so far, there is in fact a distinct flat region between the meniscus and the wave structure. For larger D, this is not so. As D increases beyond $D_{\rm III}$, the height of the left state for an undercompressive front $h_{\rm uc}$ drops below $h = 2/3$. Now there can be no direct connection between meniscus and front. All available meniscus profiles have h_m larger than $h_{\rm uc}$, so an intermediate wave is needed to span the gap of thicknesses. From the front wedge part in Figure 2.1, we see that the resulting wave structure must be a rarefaction-undercompressive wave combination.

No flat film can emerge between the meniscus and the rarefaction wave of thickness $h_m = h_w$ strictly larger than $2/3$, since the portions of the rarefaction wave larger than $2/3$ would have a negative characteristic speed. Instead, the meniscus evolves into a shape that is the limiting profile of all the Type II menisci, while the portions of the film between $2/3$ and $h_{\rm uc}$ tend to the profile of a rarefaction wave with left state $2/3$. Since the characteristic speed at $h = 2/3$ is exactly zero, the rarefaction wave never completely separates from the meniscus, but as it gets increasingly stretched, the film thickness at any fixed position x in front of the meniscus eventually tends to $2/3$. We call the emerging limiting meniscus profile with thickness $h_m = 2/3$ a *generalized Type II meniscus*, in analogy to the terminology for Lax shocks, to reflect the fact that $f'(2/3) = 0$.

This situation is indicated in Figure 3.1(a), where for $D > D_{\rm III}$ the crossed and circled lines part again. The former lies at the boundary of the Type II regime, while the latter follows the upper edge of the front wedge. Dynamical simulations with $D = 3.5$ confirm our picture. In Figure 3.4(a) we show the evolution of the film from the initial condition (3.2), for $D = 3.5$. At long times, the film left of the advancing front forms a flat plateau with thickness equal to $h_{\rm uc} = 0.5783$ (the value was obtained by solving the traveling wave ODE as in [2, 10]). At an increasing distance from the advancing front, the film profile slightly steepens to a rarefaction wave, which blends over into the meniscus.

When D increases further, $h_{\rm uc}$ decreases, and the difference between the speed of the undercompressive wave and the left characteristic speed of this wave also decreases. They become equal when $D = D_F$ and $h_{\rm uc} = (1-b)/2$ at the apex of the front wedge. For $b = 0.005$, D_F is 5.227. For the largest D, in excess of D_F, the possible wave structures are those that are permitted according to classical shock theory.

Again, the meniscus profile tends to a generalized Type II meniscus, and it must connect to a rarefaction fan with left state 2/3. The rarefaction wave now connects directly to the advancing front, which connects in turn to b. The characteristic speed of the thickness $(1-b)/2$ where the two structures connect is identical to the shock speed. The leading shock is therefore a generalized Lax shock, which is not undercompressive, and there is neither the flat region of thickness $h_{\rm uc}$ nor the steep shock front which were visible when $D = 3.5$. This is seen in a dynamical simulation for $D = 6$ in Figure 3.4(b). Instead, the rarefaction fan expands over time, always stretching from the meniscus to the advancing front. The front is a generalized Lax shock and connects to the rarefaction fan via a rounded corner at thickness $h = (1-b)/2 = 0.4975$.

3.2. Case B: Thicker precursor layer, $0.0202 < b < 0.1484$. For larger precursor thicknesses, the expected film configuration is generally similar to the small b case described above. However, some new configurations do appear for a range of D values, while the front controlled profiles with a Type II meniscus and flat region thicker than 2/3 no longer occur.

The intersections occurring at $D = D_{\rm II}$ and $D = D_{\rm III}$ both happen at $D = D_M$ when $b = 0.0202$. For larger b, the upper branch $h_{\rm uc}$ of the front wedge intersects only the lower branch of the meniscus wedge. We call the value of D for the remaining intersection $D_{\rm II}$, and so

$$h_B(D_{\rm II}) = h_{\rm uc}(D_{\rm II}, b)$$

for this range of b. The two wedges are shown for $b = 0.05$ in Figure 3.1(b). We describe below the profiles which result as D is increased.

For $D < D_{\rm II}$, the film behaves as for the first two cases in Case A (section 3.1). For the smallest D, profiles continue to be controlled by the meniscus, with a Type I meniscus, a flat region of thickness h_B, and simple compressive front, with a capillary ridge connecting the flat region to the precursor layer. When $D_{\rm I} < D < D_{\rm II}$, there are again a Type I meniscus and flat region with thickness h_B, but at the advancing front there is a double shock structure. Both these behaviors have been seen in our dynamic simulations with $b = 0.05$.

Because $h_{\rm uc}(D)$ is smaller than in Case A for the larger values of b considered here, it never exceeds h_T, the threshold for a Type II meniscus. (See Figure 3.1(b).) Thus for this range of b there are no front controlled profiles. Instead, for $D_{\rm II} < D < D_M$, a new configuration is possible. Now both h_B and h_T are greater than $h_{\rm uc}$, so any connection from the meniscus must be via a rarefaction fan. This cannot connect to a Type II meniscus, since $h_T > 2/3$, and so the left part of the rarefaction would have negative speed. Thus here there is a Type I meniscus, connected via a rarefaction fan to a flat region of height $h_{\rm uc}$ where the film is thinner. This in turn is connected via an undercompressive shock to b. All parts of the rarefaction wave have positive speed, so it must gradually move away from the meniscus, leaving a flat region of thickness h_B behind it. Furthermore, the leading edge of the rarefaction fan is slower than the undercompressive wave, so the length of the flat region between the rarefaction fan leading edge and the advancing front, with thickness $h_{\rm uc}$, will increase with time. Dynamical simulations with $D = 0.7$ and $b = 0.05$, shown in Figure 3.5(a), confirm

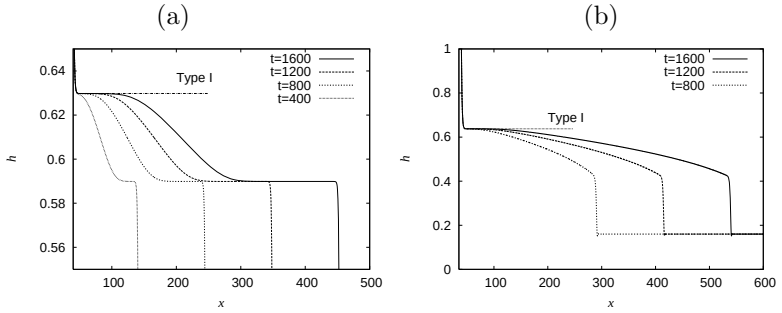


FIG. 3.5. (a) *Evolution of the film when* $D = 0.7$ *and* $b = 0.05$ *(so* $D_{\rm II} < D < D_M$*). A Type* I *meniscus, which tends towards* $h_B = 0.6298$*, is connected via a rarefaction fan to a flat region of height* $h_{\rm uc}$*, which connects to the precursor thickness* b *via an undercompressive shock. This configuration is possible only when* $b > 0.0202$*.* (b) *Evolution of the film when* $D = 0.72$*. Here* $b = 0.16$*. The front is now a generalized Lax shock, unlike that seen in the left figure. The meniscus is still Type* I*, approaching a value* $h_B = 0.6376 < 2/3$*. A knee develops close to* $h = (1-b)/2 = 0.42$*.*

that two flat regions develop. The first has thickness close to $h_B = 0.6298$, as expected for a Type I meniscus solution for this value of D, and extends up to the rarefaction fan. A second develops between the rarefaction fan and the advancing front, with a thickness close to the expected undercompressive region height $h_{\rm uc} = 0.5907$.

For $D > D_M$, the behavior is similar to that for $D > D_{\rm III}$ for small b (section 3.1). There is a rarefaction fan extending from a generalized Type II meniscus at $h = 2/3$ to $h_{\rm uc}$. Provided that $D < D_F$, a flat region of thickness $h_{\rm uc}$ exists before an undercompressive shock, while for $D > D_F$ there is no flat region, and the connection to the precursor is a generalized Lax shock.

3.3. Case C: Even thicker precursor layer, $0.1484 < b < 0.2338$. As b is further increased, D_F reduces, so that when it reaches the critical value $b = 0.1484$, D_F and D_M are equal. For somewhat larger b (so that $D_F < D_M$, but is not too small) the apex of the front wedge still lies outside the meniscus wedge. This arrangement of the wedges is shown in Figure 3.1(c) when $b = 0.16$.

With this arrangement, the film behaves as for Cases A and B while $D < D_{\rm II}$, forming a Type I meniscus with a compressive shock for $D < D_{\rm I}$, and then a double shock for $D_{\rm I} < D < D_{\rm II}$. For $D > D_{\rm II}$ but less than D_F, the Type I meniscus still exists and connects to the precursor via a flat region of thickness h_B, then a rarefaction fan and undercompressive shock, as in Case B. In dynamic simulations the flat $h_{\rm uc}$ region takes a long time to develop.

For $D_F < D < D_M$ another new behavior occurs. Type I menisci are still possible, but they must connect to the precursor by a rarefaction fan with right state $(1-b)/2$, followed by a classical generalized Lax shock, since undercompressive connections do not exist for $D > D_F$. Such behavior is seen in Figure 3.5(b), for $D = 0.72$ and $b = 0.16$. (Note that $D_F = 0.7153$ for $b = 0.16$.) Here $h_B = 0.6376$, and it is apparent that in the meniscus region the film approaches this thickness before entering the rarefaction fan region and dropping to $(1-b)/2 = 0.42$. Since h_B is quite close to $2/3$ here, characteristics have a slow speed, and the left edge of the rarefaction fan takes a long time to move away from the meniscus. Finally, for $D > D_M$, behavior is again as for the largest D values in Cases A and B: a generalized Type II meniscus connects to a rarefaction fan and from there to a generalized Lax shock, as in Figure 3.4(b).

3.4. Case D: Thickest precursor layers: $b > 0.2338$. With this arrangement of the wedges, $D_F < D_{\mathrm{I}}$, and the front wedge lies entirely within the meniscus wedge. Figure 3.1(d) shows the case $b = 0.25$. Three types of behavior are possible; these are similar to those of Case C.

For small $D < D_{\mathrm{I}}$ (but sufficiently large that $h_B > b$), the film continues to display meniscus-controlled behavior. There is a Type I meniscus, followed by a flat region with thickness $h_B(D)$, which is connected to the precursor layer by a compressive advancing front. This is the case, regardless of whether D is larger or smaller than D_F.

When $D > D_{\mathrm{I}}$, the connection to the precursor must be via a classical structure (since $D_{\mathrm{I}} > D_F$). Since the preferred right state of the meniscus exceeds $(1 - b)/2$, there is a rarefaction fan, connecting to a classical generalized Lax shock. The two behaviors possible for $D > D_{\mathrm{I}}$ differ near the meniscus: for $D_{\mathrm{I}} < D < D_M$, it is a Type I meniscus, which connects to the rarefaction fan as in section 3.3. Once D exceeds D_M, there is a generalized Type II meniscus that joins to the rarefaction fan, as for the largest D values in the previous cases.

4. Nearly horizontal substrate: Large D. For a substrate which is nearly horizontal, the leveling effects of the normal component of gravity become important. Here we consider steady state profiles in the limit $D \to \infty$. For studying this regime, we adopt a scaling in which surface tension is neglected but both components of gravity are retained. Smoothing of discontinuities in the film is now provided by the normal component of gravity instead of predominantly by surface tension. Rescaling by defining $\tilde{x}$ and $\tilde{t}$ by

$$D\tilde{x} = x, \qquad D\tilde{t} = t, \tag{4.1}$$

and letting $D \to \infty$ causes the governing PDE (2.2) to reduce from fourth to second order:

$$h_{\tilde{t}} + \left(h^2 - h^3\right)_{\tilde{x}} = \left(h^3 h_{\tilde{x}}\right)_{\tilde{x}}. \tag{4.2}$$

Steady state solutions of (4.2), which represent feasible meniscus profiles, must satisfy boundary conditions far upstream and downstream. The film must match onto the reservoir, so

$$\frac{\mathrm{d}h}{\mathrm{d}\tilde{x}} \to -1 \qquad \text{as } \tilde{x} \to -\infty$$

(which is simply the rescaled form of (2.3)), and its thickness must approach a constant value h_m far downstream:

$$h \to h_m \qquad \text{as } \tilde{x} \to \infty.$$

Setting $h_{\tilde{t}}$ to zero, (4.2) can be integrated with respect to $\tilde{x}$ to yield a first-order (nonlinear) ODE:

$$h_{\tilde{x}} = \frac{h^2 - h^3 - c}{h^3}. \tag{4.3}$$

For large h, equation (4.3) has $h_{\tilde{x}} \to -1$, and by setting the constant of integration $c = (h_m^2 - h_m^3)$, both boundary conditions are satisfied. Note that c is the total flux of liquid flowing through the flat film in front of the meniscus; physically meaningful

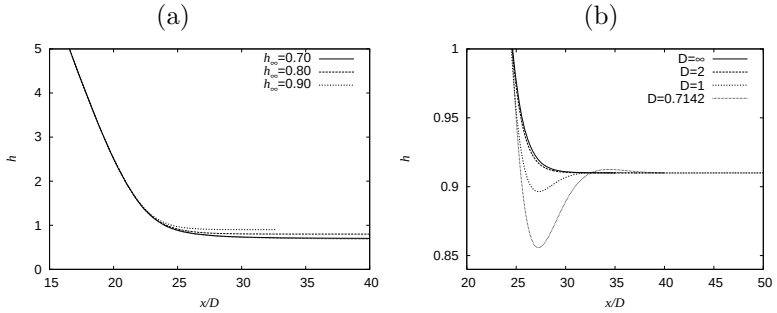


FIG. 4.1. (a) *Solutions of* (4.3) *in a parameter regime where surface tension is unimportant. Shown are* $h_m = 0.7, 0.8, 0.9$. *On this scale, the solution with* $h_m = 0.7$ *is barely distinguishable from the solution of* (4.5) *with* $h_m = 2/3$, *shown as a dotted line.* (b) *Steady solutions of* (2.2), *which have* $h_m = 0.91$, *rescaled according to* (4.1). *As* D *is increased, these approach the large* D *limit, shown as a solid line. The dip which is a feature of Type* II *solutions for small* D *is gone for* $D = 2$.

values for climbing films are $0 < c \leq 4/27$. A one-parameter family of solutions is generated by varying h_m, or alternatively, c. For $c < 4/27$, the positive-valued fixed points of (4.3) are $h = h_B < 2/3$ and $h = h_T > 2/3$, the same as those of the steady form of (2.2).

When h_B and h_T are distinct, (4.3) shows that $h_{\tilde{x}}$ is negative for $h > h_T$ and $h < h_B$, and positive for $h_B < h < h_T$; h_T is therefore a stable fixed point, while h_B is unstable. Any solution which becomes infinite as $\tilde{x} \to -\infty$ must be monotonic, decreasing, and have $h > h_T > 2/3$ everywhere, with $h \to h_T$ as $\tilde{x} \to \infty$. In other words, the meniscus profiles for $c < 4/27$ connect to a thickness $h_m > 2/3$, and so are Type II profiles. The exact solution to (4.3), up to translation in $\tilde{x}$, is given implicitly by

$$\tilde{x} = -h + \sum_{i=1}^{3} \frac{h_i^3}{(h_i - h_j)(h_k - h_i)} \log|h - h_i|, \qquad j \neq i,\ k \neq i, j, \tag{4.4}$$

where the summation is over the three distinct roots h_B, h_T, and $(1 - h_B - h_T)$ of the cubic equation $h^2 - h^3 - c = 0$. Solutions of the form of (4.4) are shown in Figure 4.1(a) for three values of $h_m > 2/3$. The solution with $h_m = 0.91$ is compared to meniscus profiles with finite D in Figure 4.1(b). As the influence of surface tension diminishes, the dip disappears, and the meniscus profiles become monotonic, even though they are of Type II.

As h_m approaches $2/3$, the fixed points h_B and h_T also approach $2/3$. When $h_m = 2/3$, $c = 4/27$, and there is a repeated root of $h^2 - h^3 - c = 0$. The solution to (4.3) is then given by

$$\tilde{x} = -h + \frac{8}{27}\frac{1}{h - 2/3} - \frac{28}{27}\log\left|h - \frac{2}{3}\right| + \frac{1}{27}\log\left|h + \frac{1}{3}\right|. \tag{4.5}$$

The $(h - 2/3)^{-1}$ term rapidly blows up, indicating that the film requires a very long distance to reach its limiting value h_m. The solution for $h_m = 2/3$ is shown in Figure 4.1(a) as a dotted line.

At this point, we have established essentially the same picture for the possible meniscus structure as for the large (but finite) D case discussed in section 3. On

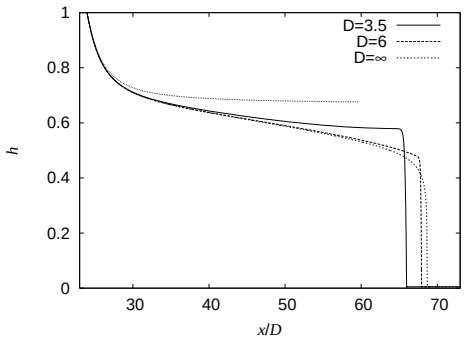


FIG. 4.2. *Advancing fronts with $b = 0.005$ and different D, shown at the same scaled time $\tilde{t} = 169$. Also shown is the long-time solution for $h_m = 2/3$ and $D = \infty$ (dotted line), which is approached by the solutions for large D, and when the front has moved far from the meniscus.*

the other hand, since (4.2) is a second-order equation, the classical theory of shocks only allows a rarefaction fan and generalized Lax shock combination for connecting a thickness larger than $(1-b)/2$ with a precursor of thickness b. Therefore in numerical simulations of (4.2) a generalized Type II meniscus profile should emerge connected to such a combination, and indeed, this is seen in the results of the previous section for sufficiently large D.

We compared the profiles obtained from our dynamic model including the surface tension terms (2.2) for two moderately large values of D with those of the time-dependent model (4.2) valid when $D \to \infty$. Simulations were performed with $b = 0.005$, for which $D_F \approx 5.227$. The same initial condition (3.2) was used for each case.

As described in section 2.2, when D is larger than D_F, the presence of the surface tension terms is expected to give rise to a classical front, with a rarefaction fan connected to a compressive shock. Numerical simulations with $D = 6$ confirm this. This is shown in Figure 3.4(b), where the front advances with a rounded corner, typical of a generalized Lax shock. In contrast, when $D = 3.5 < D_F$ the advancing front is undercompressive (Figure 3.4(a)). It separates from the rarefaction wave and has a markedly higher left state. At the same rescaled time $\tilde{t} = 169$, the rescaled profiles for the meniscus region and much of the film are very similar to the large D result, i.e., using (4.2), for both finite D values. This is shown in Figure 4.2. However, at the advancing front, the difference between the undercompressive and generalized Lax fronts is evident, while the transition from generalized Lax shock to the precursor is slightly more rounded for $D = 6$ than for $D = \infty$.

Finally, we demonstrate that the portion behind the advancing front seen in Figure 3.4(b) (and also Figure 3.5(b)) is indeed a rarefaction fan, by comparing it directly to solutions of the first-order equation resulting from neglecting all second- and fourth-order smoothing terms in (2.2). Rarefaction wave solutions of $h_t + f(h)_x = 0$ may be found, subject to appropriate initial data, using the method of characteristics. Thus within the rarefaction fan delimited by left and right states h_- and h_+,

$$h(x,t) = h_R(\xi) = (f')^{-1}(\xi), \quad \text{where } \xi = \frac{x - x_0}{t - t_0} \tag{4.6}$$

for some x_0 and t_0. The function $h_R(\xi)$ is given implicitly by $\xi = f'(h) = 2h - 3h^2$. The unknowns x_0 and t_0 may be estimated as follows. For the situation shown in Figure 3.4(b), we take h_+ to be $(1-b)/2 = 0.4975$. At a given time, t_2 say, we observe

where $h(x, t_2) = h_+$, at $x = x_+$ say, and estimate a value for t_0. We then compute

$$x_0 = x_+ + (t_2 - t_0)\xi_+,$$

where $\xi_+ = f'(h_+)$. The shape of the rarefaction fan may be constructed at any time t using

$$x = (t - t_0)\xi(h) + x_0 \quad \text{for } h_+ < h < 2/3. \tag{4.7}$$

We vary t_0 until (4.7) provides a good fit for $h(x, t_2)$ within the rarefaction fan.

We demonstrate this by computing the rarefaction fan constants x_0 and t_0 using our result at $t = 3000$ (shown using "+" symbols in Figure 3.4(b)) and then confirming these by comparison with (4.7) at $t = 1013$ ("×" symbols in Figure 3.4(b)). The agreement is satisfactory in the interior of the rarefaction fan. At the ends, the higher-order terms in (2.2) are important and smooth the profile.

5. Summary of behavior: A catalog. The previous sections' observations can be summarized by considering regions of (D, b) parameter space in which distinct behaviors arise. These are shown in Figure 5.1. The graphs of $D_i(b)$ $(i = \mathrm{I}, \mathrm{II}, \mathrm{III})$ and $D_F(b)$ divide the parameter space into several regions. These regions are indicated by the labels in the figure. The descriptions given for Cases A to D in section 3 correspond to moving along horizontal lines (constant b) in this figure.

A number of features are apparent. For small D (though with $D \gg \epsilon$), a Type I meniscus (labeled "T1+L") results for all choices of b. At the other extreme, for the largest $D > \max(D_F, D_M)$, the advancing front is a generalized Lax shock, and the connection to the meniscus is via a rarefaction fan (labeled "2/3+rf+gL"). States in

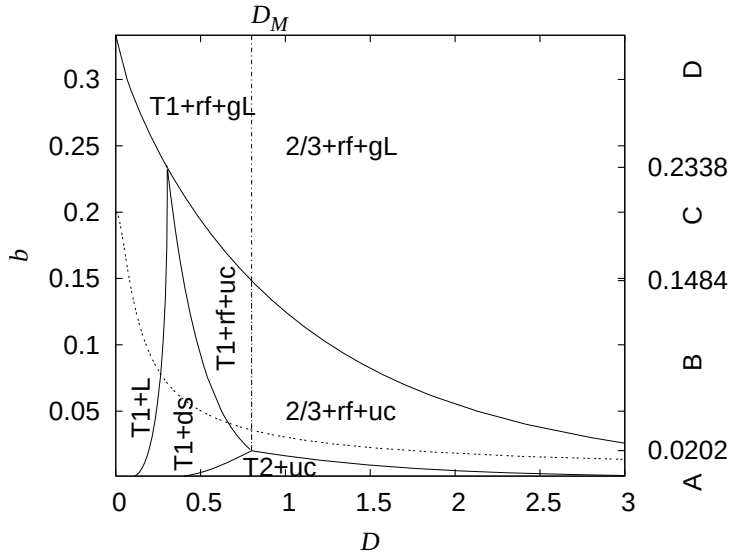


FIG. 5.1. *Behavior of a thin surface-tension-gradient-driven film emerging from a meniscus. The thickness b of a thin precursor layer and an inclination parameter D control the behavior. The following abbreviations are used, and explained in the text: T1, T2, and 2/3 denote Type* I, *Type* II, *and 2/3 (generalized Type* II*) menisci, respectively; the other labels are: L—(compressive) Lax shock, ds—double shock, gL—generalized Lax shock, rf—rarefaction fan, uc—undercompressive shock. The dotted line indicates points which differ only in tilt angle* α.

which there is only a flat region between the meniscus and front—the Type I meniscus with a Lax front or with a double shock structure, and the Type II meniscus with an undercompressive front—exist in the lower left part of the diagram, shown by labels "T1+L," "T1+ds," and "T2+uc." These are states are familiar from previous work, e.g., [13]. For a fixed value of D, one of these three configurations can arise only when a sufficiently thin precursor is present. As the precursor layer is thickened these give way to film profiles with rarefaction fans, shown by "rf" in the figure labels. These new kinds of behavior were described in section 3. In particular, the combination of a Type I meniscus and rarefaction fan seen in Figure 3.5(a) occurs only for $b > 0.0202$, while a Type II meniscus with an extended flat region and undercompressive front occurs only for b smaller than this value. In the upper right part of the diagram lie structures with rarefaction fans, for which there is no clear separation between the meniscus and the front.

It should be noted that in a set of experiments, the dimensional precursor thickness b^* (or equivalently the wetting behavior) is likely to be fixed. If the substrate angle α is varied while other parameters including b^* are fixed, then this corresponds to moving along a curved path in Figure 5.1, given parametrically by $((3\delta)^{2/3} \sin\alpha/\cos^{4/3}\alpha,\, b_0 \cos\alpha)$, where b_0 is the dimensionless precursor thickness for $\alpha = 0$. One such curve is shown in the figure for $\delta = 0.00782$, corresponding to experiments by Schneemilch and Cazabat [14], and $b_0 = 0.2$, as a dotted line. This value is likely to be larger than in their experiments, but demonstrates how the film changes from a Type I meniscus and compressive front to more complex behavior as α is increased. For this value of δ, values of $b_0 > 0.11$ result in not entering the "T2+uc" region for any α.

6. Concluding remarks. In this paper, understanding of the film on a heated tilted substrate near a meniscus and at an advancing front has been combined to generate an understanding of the possible behavior for the entire film. This is graphically summarized by Figure 5.1. This analysis is based on approximating the curvature, which is appropriate for relatively tilted substrates, for which $\tan\alpha \gg 1$. This is equivalent to requiring that $D \gg \epsilon$.

When D is of order ϵ or smaller, the curvature can no longer be approximated by h_{xx} everywhere. In this limit h_B approaches a finite nonzero value [4, 5, 11, 13]. For practical values of the shear stress, the most significant effect is to move the bottom edge of the meniscus wedge so that the graph of $1 - b - h_{\rm uc}$ no longer intersects $h_B(D)$. This situation occurs more easily for smaller values of b. In that case, instead of a Type I meniscus with a compressive front for the smallest values of D, either a double shock profile or a Type II meniscus with undercompressive front may occur. In principle, this means that undercompressive advancing fronts are possible for nearly vertical substrates, provided that the precursor layer is sufficiently thin. Retaining full curvature would modify the results shown in Figure 5.1 for D near 0.

Our results indicate that when the control parameter D is sufficiently large there is no extended flat region, but rather a rarefaction fan links the meniscus to the advancing front. An interesting observation is that when a Type II meniscus arises, its flat region thickness is not controlled by conditions at the meniscus, as for the Type I meniscus. Rather, it is the precursor thickness which determines $h_{\rm uc}$ and so the thickness of the flat region.

Similarly the interesting question of what would happen if the film were to advance over a substrate for which wetting is imperfect, i.e., for which there is a nonzero contact angle, has not been addressed. For the drag-out problem, there is a minimum

withdrawal speed required to draw out a film, if the contact angle is prescribed [7].

Despite these limitations, we expect that the guide presented here will be a useful tool for experimentalists. We look forward to experimental confirmation of these results.

REFERENCES

[1] A. L. Bertozzi, A. Münch, X. Fanton, and A. M. Cazabat, *Contact line stability and 'undercompressive shocks' in driven thin film flow*, Phys. Rev. Lett., 81 (1998), pp. 5169–5172.

[2] A. L. Bertozzi, A. Münch, and M. Shearer, *Undercompressive waves in driven thin film flow*, Phys. D, 134 (1999), pp. 431–464.

[3] A. L. Bertozzi and M. Shearer, *Existence of undercompressive traveling waves in thin film equations*, SIAM J. Math. Anal., 32 (2000), pp. 194–213.

[4] P. Carles and A.-M. Cazabat, *The thickness of surface-tension-gradient-driven spreading films*, J. Colloid Interface Sci., 157 (1993), pp. 196–201.

[5] X. Fanton, A.-M. Cazabat, and D. Quéré, *Thickness and shape of films driven by a Marangoni flow*, Langmuir, 12 (1996), pp. 5875–5880.

[6] R. P. Haskett, T. P. Witelski, and J. Sur, *Localized Marangoni forcing in driven thin films*, Phys. D, 209 (2005), pp. 117–134.

[7] L. M. Hocking, *Meniscus draw-up and draining*, European J. Appl. Math., 12 (2001), pp. 195–208.

[8] R. Levy and M. Shearer, *Kinetics and nucleation for driven thin film flow*, Phys. D, 209 (2005), pp. 145–163.

[9] O. K. Matar and R. V. Craster, *Models for Marangoni drying*, Phys. Fluids, 13 (2001), pp. 1869–1883.

[10] A. Münch, *Shock transitions in Marangoni-gravity driven thin film flow*, Nonlinearity, 13 (2000), pp. 731–746.

[11] A. Münch, *The thickness of a Marangoni-driven thin liquid film emerging from a meniscus*, SIAM J. Appl. Math., 62 (2002), pp. 2045–2063.

[12] A. Münch, *Pinch-off transition in Marangoni-driven thin films*, Phys. Rev. Lett., 91 (2003), paper 016105.

[13] A. Münch and P. L. Evans, *Marangoni-driven liquid films rising out of a meniscus onto a nearly horizontal substrate*, Phys. D, 209 (2005), pp. 164–177.

[14] M. Schneemilch and A. M. Cazabat, *Shock separation in wetting films driven by thermal gradients*, Langmuir, 16 (2000), pp. 9850–9856.

[15] M. Schneemilch and A. M. Cazabat, *Wetting films in thermal gradients*, Langmuir, 16 (2000), pp. 8796–8801.

[16] L. W. Schwartz, *On the asymptotic analysis of surface-stress-driven thin-layer flow*, J. Engrg. Math., 39 (2001), pp. 171–188.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1632–1648

ALGORITHMS FOR FINDING GLOBAL MINIMIZERS OF IMAGE SEGMENTATION AND DENOISING MODELS*

TONY F. CHAN†, SELIM ESEDOĞLU‡, AND MILA NIKOLOVA§

Abstract. We show how certain nonconvex optimization problems that arise in image processing and computer vision can be restated as convex minimization problems. This allows, in particular, the finding of global minimizers via standard convex minimization schemes.

Key words. denoising, segmentation

AMS subject classifications. 94A08, 65K10

DOI. 10.1137/040615286

1. Introduction.

Image denoising and segmentation are two related, fundamental problems of computer vision. The goal of denoising is to remove noise and/or spurious details from a given, possibly corrupted, digital picture while maintaining essential features such as edges. The goal of segmentation is to divide the image into regions that belong to distinct objects in the depicted scene.

Approaches to denoising and segmentation based on the calculus of variations and partial differential equations (PDEs) have had great success. One important reason for their success is that these models are particularly well suited to imposing geometric constraints (such as regularity) on the solutions sought. Among the best known and most influential examples are the Rudin–Osher–Fatemi (ROF) total variation–based image denoising model [22] and the Mumford–Shah image segmentation model [17].

Denoising models such as the ROF model can be easily adapted to different situations. An interesting scenario is the denoising of shapes: Here, the given image is binary (representing the characteristic function of the given shape), and the noise is in the geometry of the shape: Its boundary might be very rough, and the user might be interested in smoothing out its boundary, and perhaps removing small, unnecessary connected components of the shape. This task is a common first step in many object detection and recognition algorithms.

A common difficulty with many variational image processing models is that the energy functional to be minimized has local minima (which are not global minima). This is a much more serious drawback than nonuniqueness of global minimizers (which is also a common phenomenon) because local minima of segmentation and denoising models often have completely wrong levels of detail and scale: whereas global minimizers of a given model are usually all reasonable solutions, the local minima tend to be blatantly false. Many solution techniques for variational models are based on gradient descent, and are therefore prone to getting stuck in such local minima. This

*Received by the editors September 17, 2004; accepted for publication (in revised form) November 21, 2005; published electronically June 19, 2006.
http://www.siam.org/journals/siap/66-5/61528.html

†Mathematics Department, UCLA, Los Angeles, CA 90095 (TonyC@college.ucla.edu). The research of this author was supported in part by NSF contract DMS-9973341, NSF contract ACI-0072112, ONR contract N00014-03-1-0888, and NIH contract P20 MH65166.

‡Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 (esedoglu@umich.edu). The research of this author was supported in part by NSF award DMS-0410085.

§Centre de Mathematiques et de Leurs Applications, ENS de Cachan, 61 av. du President Wilson, 94235 Cachan Cedex, France (nikolova@cmla.ens-cachan.fr).

makes the initial guess for gradient descent–based algorithms sometimes critically important for obtaining satisfactory results.

In this paper we propose algorithms which are guaranteed to find global minimizers of certain denoising and segmentation models that are known to have local minima. As a common feature, the models we consider involve minimizing functionals over characteristic functions of sets, which is a nonconvex collection; this feature is responsible for the presence of local minima. Our approach, which is based on observations of Strang in [23, 24], is to extend the functionals and their minimization to all functions in such a way that the minimizers of the extended functionals can be subsequently transformed into minimizers for the original models by simple thresholding. This allows, among other things, computing global minimizers for the original nonconvex variational models by carrying out standard convex minimization schemes.

Our first example is binary image denoising, which we briefly discuss as a precursor to and a motivation for the more general segmentation problem that we subsequently consider. Here the given noisy image for the ROF model is taken to be binary, and the solution is also sought among binary images. This problem has many applications where smoothing of geometric shapes is relevant. Some examples are the denoising of fax documents, and the fairing of surfaces in computer graphics. Because the space of binary functions is nonconvex, the minimization problem involved is actually harder than minimizing the original ROF model. In section 2, we recall results from [6] that show how this model can be written as a convex optimization problem, and we exhibit its use as a numerical algorithm. In this section, we also analytically verify that the energy concerned possesses local minimizers that are not global minimizers, which can trap standard minimization procedures.

In section 3, we extend some of the results of [6] to the more important problem of *image segmentation*. In particular, we consider the two-phase, piecewise constant Mumford–Shah segmentation functional proposed by Chan and Vese in [7] and show that part of the minimization involved can be given a convex formulation. This model has become a very popular tool in segmentation and related image processing tasks (also see [26] for its multiphase version). The convex formulation we obtain turns out to be closely related to the algorithm of Chan and Vese presented in [7]. Our observations indicate why this algorithm is successful in finding interior contours and other hard-to-get features in images.

2. Previous work. The results and the approach of this paper follow very closely the observations of Strang in [23, 24]. In those papers, optimization problems of the following form, among others, are studied:

$$\inf_{\{u:\int f u\,dx=1\}} \int |\nabla u|, \tag{1}$$

where $f(x)$ is a given function. It is shown in particular that the minimizers of (1) turn out to be characteristic functions of sets. The main idea involved is to express the functional to be minimized and the constraint in terms of the super level sets of the functions $u(x)$ and $f(x)$. The coarea formula of Fleming, Rishel, and Rishel [11] is the primary tool.

In this paper, the idea of expressing functionals in terms of level sets is applied to some simple image processing models. For instance, in section 4, where we study the piecewise constant Mumford–Shah energy, we show that the relevant energy, which is originally formulated in terms of *sets*, can be reformulated as an optimization problem over *functions* in such a way that the resulting *convex* energy turns out to be almost

the same as (1). After this reformulation, following Strang's work, we are also able to express the resulting convex variational problem in terms of super level sets of the unknown functions. That in turn allows us to extract a minimizer of the original nonconvex model from a minimizer of the convex functional by simple thresholding.

We should point out that our emphasis in this paper is in some sense opposite that of [23, 24]. Indeed, in those works the main point is that some energies of interest that need to be minimized over all functions turn out to have minimizers that take only two values. In our case, we start with a variational problem that is to be minimized over only functions that take two values (i.e., characteristic functions of sets) but show that we may instead minimize over all functions (that are allowed to take intermediate vales), i.e., we may ignore the nonconvex constraint. This allows us to end up with a convex formulation of the original nonconvex problem.

3. The ROF model. Rudin, Osher, and Fatemi's total variation image denoising model [22] is one of the best known and successful of PDE-based image denoising models. Indeed, being convex it is one of the simplest denoising techniques that has the all-important edge preserving property.

Let $D \subset \mathbf{R}^N$ denote the image domain. In practice, D is simply a rectangle, modeling the computer screen. Therefore, mathematically, it is natural to assume that D is a bounded domain with Lipschitz boundary. However, for the convenience of not dealing with boundaries, in this section we will take D to be the entire space $\mathbf{R}^N$. This simplifying assumption has no bearing on the essential ideas discussed below.

Let $f(x) : \mathbf{R}^N \to [0,1]$ denote the given (grayscale) possibly corrupted (noisy) image. The energy to be minimized in the standard ROF model is then given by

$$E_2(u,\lambda) = \int_{\mathbf{R}^N} |\nabla u| + \lambda \int_{\mathbf{R}^N} \Big(u(x) - f(x)\Big)^2 \, dx. \tag{2}$$

The appropriate value of the parameter $\lambda > 0$ in the model (2) can be determined if the noise level is known; an algorithm for doing so is given in [22]. If information about noise level is not available, then λ needs to be chosen by the user. This choice can be facilitated by the observation that λ acts as a scale parameter [25]: Its value determines in some sense the smallest image feature that will be maintained in the reconstructed image. Energy (2) is often minimized via gradient descent; however, see [5, 9] for an alternative approach in the $\lambda = 0$ case.

An interesting application of the ROF model described above is to *binary* image denoising. This situation arises when the given image $f(x)$ is binary (i.e., $f(x) \in \{0,1\}$ for all $x \in \mathbf{R}^N$) and is known to be the corrupted version of another binary image $u : \mathbf{R}^N \to \{0,1\}$ that needs to be estimated. Naturally, $f(x)$ can then be expressed as

$$f(x) = \mathbf{1}_\Omega(x),$$

where Ω is an arbitrary bounded measurable subset of $\mathbf{R}^N$. In this case, the noise is in the *geometry*; for example, the boundary $\partial\Omega$ of Ω might have spurious oscillations, or Ω might have small connected components (due to presence of noise) that need to be eliminated. The ROF model (2) can easily be specialized to this scenario by restricting the unknown $u(x)$ to have the form $u(x) = \mathbf{1}_\Sigma(x)$, where Σ is a subset of

$\mathbf{R}^N$. One then obtains the following optimization problem:

$$\min_{\substack{\Sigma\subset\mathbf{R}^N\\ u(x)=\mathbf{1}_\Sigma(x)}} \int_{\mathbf{R}^N} |\nabla u| + \lambda \int_{\mathbf{R}^N} \Big(u(x) - \mathbf{1}_\Omega(x)\Big)^2\, dx. \tag{3}$$

Problem (3) is nonconvex because the minimization is carried out over a nonconvex set of functions. Recalling that the total variation of the characteristic function of a set is its perimeter (see, e.g., [10, 12] for such basic facts), and noticing that the fidelity term in this case simplifies, we write (3) as the following geometry problem:

$$\min_{\Sigma\subset\mathbf{R}^N} \mathrm{Per}(\Sigma) + \lambda|\Sigma \triangle \Omega|, \tag{4}$$

where $\mathrm{Per}(\cdot)$ denotes the perimeter, $|\cdot|$ is the N-dimensional Lebesgue measure, and $S_1 \triangle S_2$ denotes the symmetric difference between the two sets S_1 and S_2.

Usual techniques for approximating the solution. A very successful method of solving problems of the type (4) has been via some *curve evolution* process, sometimes referred to as *active contours*. Indeed, the unknown set Σ can be described by its boundary $\partial\Sigma$. The boundary $\partial\Sigma$ is then updated iteratively, usually according to gradient flow for the energy involved.

Numerically, there are several ways of representing $\partial\Sigma$. For the applications mentioned above, explicit curve representations as in Kass, Witkin, and Terzopoulos [15] are not appropriate, since such methods do not allow changes in curve topology (and have a number of other drawbacks). Instead, the most successful algorithms are those based on either the level set method of Osher and Sethian [21, 20] or on the variational approximation approach known as Gamma convergence theory [8].

In the level set formulation, the unknown boundary $\partial\Sigma$ is represented as the 0-level set of a (Lipschitz) function $\phi : \mathbf{R}^N \to \mathbf{R}$:

$$\Sigma = \big\{x \in \mathbf{R}^N : \phi(x) > 0\big\},$$

so that $\partial\Sigma = \{x \in \mathbf{R}^N : \phi(x) = 0\}$. The functional to be minimized in (3), which we called $E_2(\cdot, \lambda)$, can then be expressed in terms of the function $\phi(x)$ as follows:

$$\int_{\mathbf{R}^N} |\nabla H(\phi(x))|\, dx + \lambda \int_{\mathbf{R}^N} \Big(H(\phi(x)) - \mathbf{1}_\Omega(x)\Big)^2\, dx. \tag{5}$$

Here, the function $H(x) : \mathbf{R} \to \mathbf{R}$ is the Heaviside function:

$$H(\xi) = \begin{cases} 0 & \text{if } x < 0,\\ 1 & \text{if } x \geq 0. \end{cases}$$

In practice, one takes a smooth (or at least Lipschitz) approximation to $H(x)$, which we shall call $H_\varepsilon(\xi)$, where $H_\varepsilon(\xi) \to H(\xi)$ in some manner as $\varepsilon \to 0$.

The Euler–Lagrange equation for (5) is easy to obtain. It leads to the following gradient flow:

$$\phi_t(x,t) = H'_\varepsilon(\phi)\left\{\mathrm{div}\left(\frac{\nabla\phi}{|\nabla\phi|}\right) + 2\lambda\Big(\mathbf{1}_\Omega(x) - H_\varepsilon(\phi)\Big)\right\}. \tag{6}$$

When (6) is simulated using reinitialization for the level set function $\phi(x)$ and a compactly supported approximation $H_\varepsilon(x)$ to $H(x)$, it is observed to define a continuous

evolution (with respect to, say, the L^1-norm) for the unknown function $u(x) = \mathbf{1}_\Sigma(x)$ and decreases the objective energy (3) through binary images. It is analogous to the gradient descent equation in [7], which is natural since (3) is the restriction to binary images of also the energy considered in that work, namely, the two-phase, piecewise constant Mumford–Shah segmentation energy. In section 4, we will consider this energy for general (not necessarily binary) images.

Another representation technique for the unknown set Σ in (4) is, as we mentioned, based on the Gamma convergence ideas. Here, the given energy is replaced by a sequence of approximate energies with a small parameter $\varepsilon > 0$ in them. The sequence converges to the original energy as $\varepsilon \to 0$. The approximations have the form

$$E_\varepsilon(u, \lambda) = \int_{\mathbf{R}^N} \varepsilon|\nabla u|^2 + \frac{1}{\varepsilon}W(u) + \lambda\Big\{u^2\big(c_1 - f\big)^2 + (1-u)^2\big(c_2 - f\big)^2\Big\}\, dx.$$

In this energy, $W(\xi)$ is a double-well potential with equidepth wells at 0 and 1; for instance, a simple choice is $W(\xi) = \xi^2(1-\xi)^2$. The term $\frac{1}{\varepsilon}W(u)$ can be thought of as a penalty term that forces the function u to look like the characteristic function of a set: u is forced to be approximately 0 or 1 on most of $\mathbf{R}^N$. The term $\varepsilon|\nabla u|^2$, on the other hand, puts a penalty on the transitions of u between 0 and 1. Taken together, it turns out that these terms both impose the constraint that u should be a characteristic function and approximate its total variation. Precise versions of these statements have been proved in [16]. The remaining terms in E_ε are simply the fidelity term written in terms of u. This approach was extended to the full Mumford–Shah functional in [4].

We now argue, with the help of a very simple example, that these techniques will get stuck in local minima in general, possibly leading to resultant images with the wrong level of detail. This fact is already quite familiar to researchers working with these techniques from practical numerical experience.

Example. Consider the two-dimensional case, where the observed binary image $f(x)$ to be denoised is the characteristic function of a ball $B_R(0)$ of radius R, which is centered at the origin. In other words, we take $\Omega = B_R(0)$. Implementing the gradient descent algorithm defined by (6) requires the choice of an initial guess for the interface $\phi(x)$ (or, equivalently, an initial guess for the set Σ that is represented by $\phi(x)$). A common choice in practical applications is to take the observed image itself as the initial guess. In our case, that means initially we set $\Sigma = B_R(0)$.

Now, one can see without much trouble that the evolution defined by (6) will maintain radial symmetry of $\phi(x)$. That means, at any given time $t \geq 0$, the set (i.e., the candidate for minimization) represented by $\phi(x)$ is of the form

$$\Big\{x \in \mathbf{R}^2 : \phi(x) > 0\Big\} = B_r(0)$$

for some choice of the radius $r \geq 0$. We can write the energy of $u(x) = \mathbf{1}_{B_r(0)}(x)$ in terms of r, as follows:

$$E(r) := E_2(\mathbf{1}_{B_r(0)}(x), \lambda) = 2\pi r + \lambda\pi|R^2 - r^2|.$$

A simple calculation shows that if $\lambda < \frac{2}{R}$, then the minimum of this function is at $r = 0$. Hence, if we fix $\lambda > 0$, then the denoising model prefers to remove disks of radius smaller than the critical value $\frac{2}{R}$.

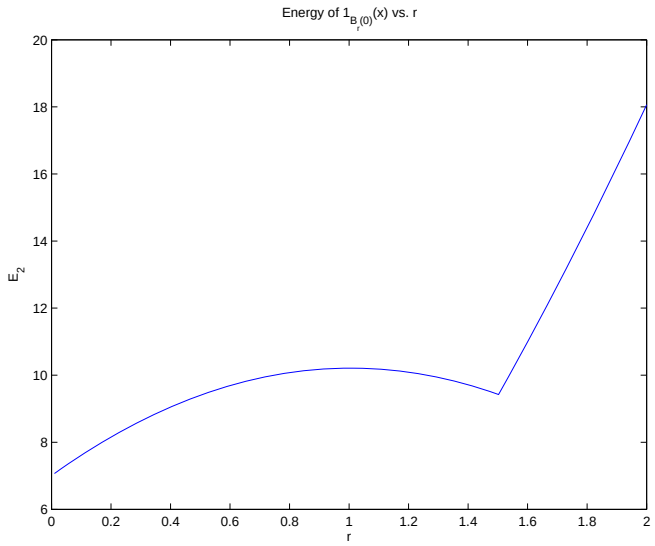


FIG. 1. *Energy* (3) *of* $u(x) = \mathbf{1}_{B_r(0)}(x)$ *as a function of* $r \in [0, 2]$ *when the observed image is given by* $f(x) = \mathbf{1}_{B_R(0)}(x)$. *Here,* $R = \frac{3}{2}$ *and the parameter* λ *was chosen to be* $\lambda = 1$. *There is clearly a local minimum, corresponding to* $r = R = \frac{3}{2}$.

But now, once again an easy calculation shows that if $R > \frac{1}{\lambda}$, then $E(r)$ has a local maximum at $r_{max}(\lambda) = \frac{1}{\lambda}$. See Figure 1 for the plot of $E(r)$ in such a case. Thus the energy minimization procedure described by (6) cannot shrink disks of radius $R \in (\frac{1}{\lambda}, \frac{2}{\lambda})$ to a point, even though the global minimum of the energy for an original image given by such a disk is at $u(x) \equiv 0$.

We can easily say a bit more: There exists $\delta > 0$ such that if $\Sigma \subset \mathbf{R}^N$ satisfies $|\Sigma \triangle B_R(0)| < \delta$, then $E_2(\mathbf{1}_\Sigma(x), \lambda) > E_2(\mathbf{1}_{B_R(0)}(x), \lambda)$. In other words, all binary images close to, but not identical with, the observed image $\mathbf{1}_{B_R(0)}(x)$ have strictly higher energy. This can be seen simply by noting that the energy of any region that is *not* a disk is strictly larger than the energy of the disk having the same area as the given region and its center at the origin.

To summarize: If $f(x) = \mathbf{1}_{B_R(0)}(x)$ with $R \in (\frac{1}{\lambda}, \frac{2}{\lambda})$, and if the initial guess for the continuous curve evolution–based minimization procedure (6) is taken to be the observed image $f(x)$ itself, then the procedure gets stuck in the local minimizer $u(x) = f(x)$. The unique global minimizer is actually $u(x) \equiv 0$.

Our example highlights the following caveat of using continuous curve evolution–based gradient descent algorithms in practice: There are many situations in which the user should be able to choose the value of λ that appears in the model in such a way that all image features smaller than the one implied by this choice of parameter are eliminated from the final result. (Such a need might arise, for instance, in the denoising of printed text, where the noise can consist of small ink blots.) With continuous curve evolution techniques, whether this goal will be achieved depends on the initial guess (our example above exhibits an unfortunate initial guess). It is clearly of interest to find an algorithm that does not have this dependence on initial conditions.

Proposed method for finding the global minimum. We now turn to an alternative way of carrying out the constrained, nonconvex minimization problem (3) that is guaranteed to yield a global minimum.

The crux of our approach is to consider minimization of the following convex energy, defined for any given observed image $f(x) \in L^1(\mathbf{R}^N)$ and $\lambda \geq 0$:

$$E_1(u(x), \lambda) := \int_{\mathbf{R}^N} |\nabla u| + \lambda \int_{\mathbf{R}^N} |u(x) - f(x)|\, dx. \tag{7}$$

This energy differs from the standard ROF model only in the fidelity term: The L^2-norm square of the original model is replaced by the L^1-norm as a measure of fidelity. It was previously introduced and studied in signal and image processing applications in [1, 2, 3, 18, 19, 6]. This variant of the ROF model has many interesting properties and uses; the point we'd like to make in this section is that it also turns out to solve our geometry denoising problem (4).

First, let us state the obvious fact that energies (2) and (7) agree on binary images (i.e., when both u and f are characteristic functions of sets). On the other hand, energy (7) is convex, but unlike energy (2), it is not strictly so. Accordingly, its global minimizers are not unique in general. Nevertheless, being convex, it does not have any local minima that are not global minima, unlike the constrained minimization (3). We therefore adopt the following notation: For any $\lambda \geq 0$, we let $M(\lambda)$ denote the set of all minimizers of $E_1(\cdot, \lambda)$. It is easy to show that for each $\lambda \geq 0$ the set $M(\lambda)$ is nonempty, closed, and convex.

The relevance of energy (7) for our purposes is established in Theorem 5.2 of [6], where additional geometric properties of it are noted. The proof is based on the following proposition, taken from [6], that expresses energy (7) in terms of the super level sets of u and f.

PROPOSITION 1. *The energy $E_1(u, \lambda)$ can be rewritten as follows:*

$$E_1(u, \lambda) = \int_{-\infty}^{\infty} \mathrm{Per}\big(\{x : u(x) > \mu\}\big) + \lambda \Big| \{x : u(x) > \mu\} \triangle \{x : f(x) > \mu\} \Big|\, d\mu. \tag{8}$$

Proof. The proof can be found in [6] (Proposition 5.1). □

We now recall also Theorem 5.2 of [6].

THEOREM 1. *If the observed image $f(x)$ is the characteristic function of a bounded domain $\Omega \subset \mathbf{R}^N$, then for any $\lambda \geq 0$ there is a minimizer of $E_1(\cdot, \lambda)$ that is also the characteristic function of a (possibly different) domain. In other words, when the observed image is binary, then for each $\lambda \geq 0$ there is at least one $u(x) \in M(\lambda)$ which is also binary.*

In fact, if $u_\lambda(x) \in M(\lambda)$ is any minimizer of $E_1(\cdot, \lambda)$, then for almost every $\gamma \in [0, 1]$ we have that the binary function

$$\mathbf{1}_{\{x : u_\lambda > \gamma\}}(x)$$

is also a minimizer of $E_1(\cdot, \lambda)$.

Proof. The proof can be found in [6] (Theorem 5.2). □

The proposition and its consequence, the theorem cited above from [6] (which are related to observations in [23, 24]), lead to a guaranteed algorithm for solving the binary image denoising problem (3), which we now state.

ALGORITHM 1. *To find a solution (i.e., a global minimizer) $u(x)$ of the nonconvex variational problem (3), it is sufficient to carry out the following three steps:*

1. *Find any minimizer of the* **convex** *energy (7); call it $v(x)$.*
2. *Let $\Sigma = \{x \in \mathbf{R}^N : v(x) > \mu\}$ for some $\mu \in (0, 1)$.*
3. *Set $u(x) = \mathbf{1}_\Sigma(x)$.*

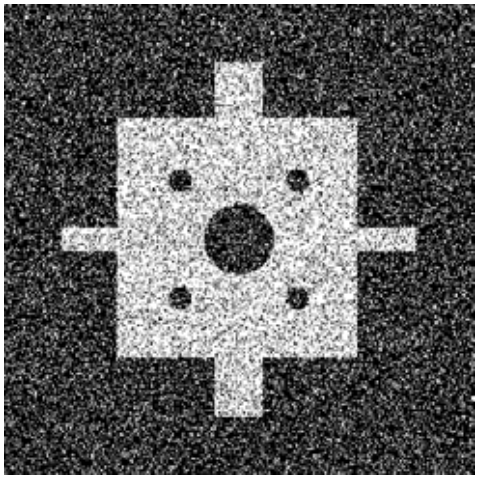


FIG. 2. *Original noisy binary image used in the numerical experiment of section* 3.

Then $u(x)$ *is a global minimizer of* (3) *for almost every choice of* μ.

Algorithm 1 reduces the *shape optimization* problem (4) to the *image denoising* problem (7). In section 4, we will obtain the analogue of Proposition 1 for the piecewise constant Mumford–Shah segmentation model of Chan and Vese, which will lead to a guaranteed algorithm for finding global minimizers of the more general shape optimization problem involved in that model, just like Algorithm 1 did for model (3); this is the content of Theorem 2 in that section. Our convex formulation will once again reduce the Chan–Vese shape optimization to a variant of the image denoising model (7).

The most involved step in the solution procedure described in Algorithm 1 is finding a minimizer of (7). One can approach this problem in many ways; for instance, one possibility is to simply carry out gradient descent.

Numerical example. The synthetic image of Figure 2 represents the given binary image $f(x)$, which is a simple geometric shape covered with random (binary) noise. The initial guess was an image composed of all 1's (an all white image). In the computation, the parameter λ was chosen to be quite moderate, so that in particular the small circular holes in the shape should be removed while the larger one should be kept. The result of the minimization is shown in Figure 3; in this case the minimizer is automatically very close to being binary, and hence the thresholding step of Algorithm 1 is almost unnecessary.

Figure 4 shows the histograms of intermediate steps during the gradient descent based minimization. As can be seen, the intermediate steps themselves are very far from being binary. The histogram in the lower right-hand corner belongs to the final result shown in Figure 3. Thus the gradient flow goes through nonbinary images, but in the end reaches another binary one. Although this is not implied by Proposition 1, Theorem 1, or Algorithm 1, it seems to hold in practice.

4. Piecewise constant segmentation. In this section, we extend the discussion of section 3 to the two-phase, piecewise constant Mumford–Shah segmentation model [17] of Chan and Vese [7]. Unlike in the previous section, this time we let the corrupted image $f(x)$ be nonbinary: it is merely assumed to be some measurable

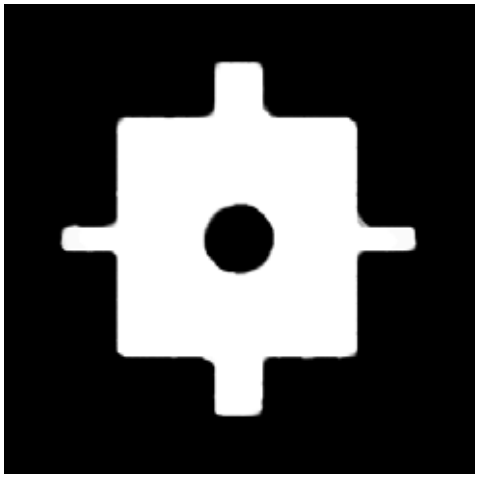


FIG. 3. *Final result found using the algorithm proposed in section* 3*, by minimizing* (7)*. Algorithm* 1 *says that global minimizers of the binary image denoising problem can be obtained by simply thresholding this result. In this case, the minimizer of energy* (7) *turns out to be very close to being binary itself, so there is no need to threshold. In the experiment, the value of* λ *was chosen small enough so that small holes in the original shape should get filled in, but also large enough so that the large hole in the middle should be maintained.*

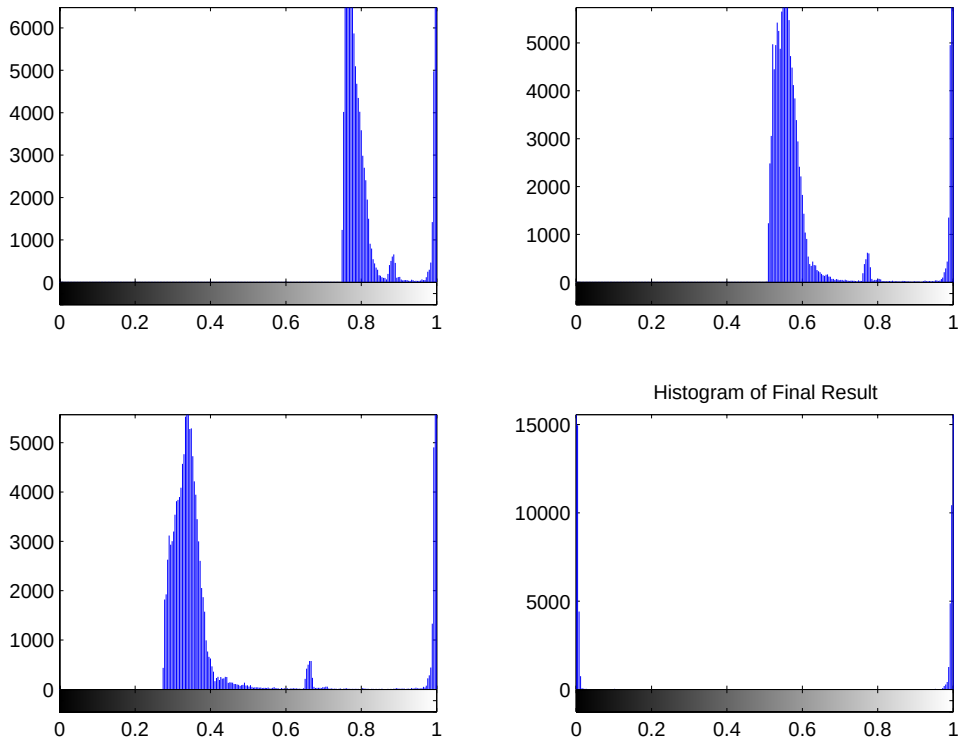


FIG. 4. *Histograms for intermediate images as the gradient descent proceeds. As can be seen, the intermediate images themselves are not binary; however, by the time the evolution reaches steady state, we are back to a binary image.*

function that takes its values in the unit interval. Thus, the discussion of this section supersedes that of the previous. Also, from now on we will assume that the image domain D is a bounded subset of $\mathbf{R}^N$ with Lipschitz boundary. The segmentation energy, which we will call MS, can then be written as

$$MS(\Sigma, c_1, c_2) := \mathrm{Per}(\Sigma; D) + \lambda \int_{\Sigma} (c_1 - f(x))^2 \, dx + \lambda \int_{D \setminus \Sigma} (c_2 - f(x))^2 \, dx. \tag{9}$$

The model says we should solve

$$\min_{\substack{c_1, c_2 \in \mathbf{R} \\ \Sigma \subset D}} MS(\Sigma, c_1, c_2). \tag{10}$$

This optimization problem can be interpreted to be looking for the best approximation in the L^2 sense to the given image $f(x)$ among all functions that take only two values. These values, denoted c_1, c_2, and where each is taken, namely, Σ and $D \setminus \Sigma$, are unknowns of the problem. As before, there is a penalty on the geometric complexity of the interface $\partial\Sigma$ that separates the regions where the two values c_1 and c_2 are taken. Functional (9) is nonconvex and can have more than one minimizer. Existence of at least one minimizer follows easily from standard arguments. Notice that if Σ is fixed, the values of c_1 and c_2 that minimize $MS(\Sigma, \cdot, \cdot)$ read

$$c_1 = \frac{1}{|\Sigma|} \int_{\Sigma} f(x) \, dx \quad \text{and} \quad c_2 = \frac{1}{|D \setminus \Sigma|} \int_{D \setminus \Sigma} f(x) \, dx. \tag{11}$$

A natural way to approximate the solution is a two-step scheme where in the first step one computes c_1 and c_2 according to these formulae, and in the second step updates the shape Σ. Even the minimization of $MS(\cdot, c_1, c_2)$ is a difficult problem since this functional is nonconvex. In what follows we focus on the minimization of $MS(\cdot, c_1, c_2)$.

We point out that if the two constants c_1 and c_2 are fixed to be 1 and 0, respectively, and if the given image $f(x)$ in (9) is taken to be the characteristic function $\mathbf{1}_{\Omega}(x)$ of a set Ω, then the minimization problem (10) reduces to the geometry problem (4); it is in this sense that this section's problem is a generalization of the previous section's.

Chan–Vese algorithm. In [7] Chan and Vese proposed a level set–based algorithm for solving the optimization problem (10). The idea is to represent the boundary $\partial\Sigma$ with the 0-level set of the function $\phi : D \to \mathbf{R}^N$. Energy (9) can then be written in terms of the level set function ϕ; it turns out to be

$$\begin{aligned} CV(\phi, c_1, c_2) = &\int_D |\nabla H_\varepsilon(\phi)| \\ &+ \lambda \int_D H_\varepsilon(\phi)(c_1 - f(x))^2 + (1 - H_\varepsilon(\phi))(c_2 - f(x))^2 \, dx. \end{aligned} \tag{12}$$

The function H_ε is, as before, a regularization of the Heaviside function. *The precise choice of the regularization H_ε of H is a crucial ingredient of the Chan–Vese algorithm.* We will return to this topic.

Variations of energy (12) with respect to the level set function ϕ lead to the following gradient descent scheme:

$$\phi_t = H'_\varepsilon(\phi) \left\{ \mathrm{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) - \lambda \left((c_1 - f(x))^2 - (c_2 - f(x))^2 \right) \right\}.$$

The optimal choice for the constants c_1, c_2 is easily determined in terms of the function ϕ.

The proposed algorithm. The Chan–Vese algorithm chooses a noncompactly supported, smooth approximation H_ε for H. As a result, the gradient descent equation given above and the following one have the same stationary solutions:

$$\phi_t = \operatorname{div}\left(\frac{\nabla\phi}{|\nabla\phi|}\right) - \lambda\left((c_1 - f(x))^2 - (c_2 - f(x))^2\right),$$

where we simply omitted the approximate Heaviside function altogether. This equation, in turn, is gradient descent for the following energy:

$$\int_D |\nabla\phi| + \lambda \int_D \left((c_1 - f(x))^2 - (c_2 - f(x))^2\right)\phi\, dx. \tag{13}$$

This energy is homogeneous of degree 1 in ϕ. As a result, it does not have a minimizer in general. In other words, the gradient descent written above does not have a stationary state: If the evolution is carried out for a long time, the level set function ϕ would tend to $+\infty$ wherever it is positive, and to $-\infty$ wherever it is negative. This issue is related to the nonuniqueness of representation with level sets and is easy to fix: one can simply restrict minimization to ϕ such that $0 \le \phi(x) \le 1$ for all $x \in D$. With this fix, and following [23, 24], we arrive at the statement below.

THEOREM 2. *For any given fixed $c_1, c_2 \in \mathbf{R}$, a global minimizer for $MS(\cdot, c_1, c_2)$ can be found by carrying out the following convex minimization:*

$$\min_{0\le u\le 1} \underbrace{\int_D |\nabla u| + \lambda \int_D \left\{(c_1 - f(x))^2 - (c_2 - f(x))^2\right\} u(x)\, dx}_{:=\tilde{E}(u,c_1,c_2)}$$

and then setting $\Sigma = \{x : u(x) \ge \mu\}$ for a.e. $\mu \in [0,1]$.

Proof. We once again rely on the coarea formula; since u takes its values in $[0,1]$, we have

$$\int_D |\nabla u| = \int_0^1 \operatorname{Per}(\{x : u(x) > \mu\}; D)\, d\mu.$$

For the other terms that constitute the fidelity term, we proceed as follows:

$$\begin{aligned}\int_D (c_1 - f(x))^2 u(x)\, dx &= \int_D (c_1 - f(x))^2 \int_0^1 \mathbf{1}_{[0,u(x)]}(\mu)\, d\mu\, dx \\ &= \int_0^1 \int_D (c_1 - f(x))^2 \mathbf{1}_{[0,u(x)]}(\mu)\, dx\, d\mu \\ &= \int_0^1 \int_{D\cap\{x:u(x)>\mu\}} (c_1 - f(x))^2\, dx\, d\mu.\end{aligned}$$

Also, we have

$$\begin{aligned}\int_D (c_2 - f(x))^2 u(x)\, dx &= \int_0^1 \int_{D\cap\{x:u(x)>\mu\}} (c_2 - f(x))^2\, dx\, d\mu \\ &= C - \int_0^1 \int_{D\cap\{x:u(x)>\mu\}^c} (c_2 - f(x))^2\, dx\, d\mu,\end{aligned}$$

where $C = \int_D (c_2 - f)^2\,dx$ is independent of u. Putting it all together, and setting $\Sigma(\mu) := \{x : u(x) > \mu\}$, we get the following formula that is valid for any $u(x) \in L^2(D)$ such that $0 \le u(x) \le 1$ for a.e. $x \in D$:

$$\begin{aligned}\tilde{E}(u, c_1, c_2) &= \int_0^1 \left\{ \mathrm{Per}(\Sigma(\mu); D) + \lambda \int_{\Sigma(\mu)} (c_1 - f(x))^2\,dx \right. \\ &\qquad \left. + \lambda \int_{D\setminus\Sigma(\mu)} (c_2 - f(x))^2\,dx \right\} d\mu - C \\ &= \int_0^1 MS\big(\Sigma(\mu), c_1, c_2\big)\,d\mu - C.\end{aligned}$$

It follows that if $u(x)$ is a minimizer of the convex problem, then for a.e. $\mu \in [0, 1]$ the set $\Sigma(\mu)$ has to be a minimizer of the original functional $MS(\cdot, c_1, c_2)$. ∎

Remark. The optimization problem that forms the content of Theorem 2 can be interpreted as follows: The level set formulation of the two-phase model depends on the level set function ϕ only through the term $H(\phi)$. The term $H(\phi)$ represents a parametrization of binary functions (since, for any given function ϕ, the function $H(\phi)$ is binary). So the minimization of (12) is thus a minimization over binary functions. Minimization of (13), on the other hand, corresponds to removing the nonconvex constraint of being binary; instead we minimize over functions that are allowed to take intermediate values. The content of the theorem above is that the minimizers (essentially) automatically satisfy the more stringent constraint.

We now turn to the question of how to minimize the convex problem stated in the theorem. In that connection, we have the following claim.

Claim 1. *Let $s(x) \in L^\infty(D)$. Then the convex, constrained minimization problem*

$$\min_{0\le u\le 1} \int_D |\nabla u| + \lambda \int_D s(x)u\,dx$$

has the same set of minimizers as the following convex, unconstrained *minimization problem:*

$$\min_u \int_D |\nabla u| + \int_D \alpha\nu(u) + \lambda s(x)u\,dx,$$

where $\nu(\xi) := \max\{0, 2|\xi - \frac{1}{2}| - 1\}$, provided that $\alpha > \frac{\lambda}{2}\|s(x)\|_{L^\infty(D)}$.

Proof. The term $\alpha\nu(u)$ that appears in the second, unconstrained minimization problem given in the claim is an *exact penalty* term [13, 14]; see Figure 5 for a plot of its graph. Indeed, the two energies agree for $\{u \in L^\infty(D) : 0 \le u(x) \le 1\ \forall x\}$. So we only need to show that any minimizer of the unconstrained problem automatically satisfies the constraint $0 \le u \le 1$. This is immediate: If $\alpha > \frac{\lambda}{2}\|s(x)\|_{L^\infty}$, then

$$|\lambda s(x)| \max\{|u(x)|, |u(x) - 1|\} < \alpha\nu(u(x)) \text{ whenever } u(x) \in [0, 1]^c,$$

which means that the transformation $u \to \min\{\max\{0, u\}, 1\}$ always decreases the energy of the unconstrained problem (strictly if $u(x) \in [0, 1]^c$ on a set of positive measure). That leads to the desired conclusion. ∎

Numerical examples. Here we detail how we obtained the numerical results pertaining to the two-phase piecewise constant segmentation models that are presented in this paper. Given c_1, c_2, the "exact penalty" formulation of the equivalent

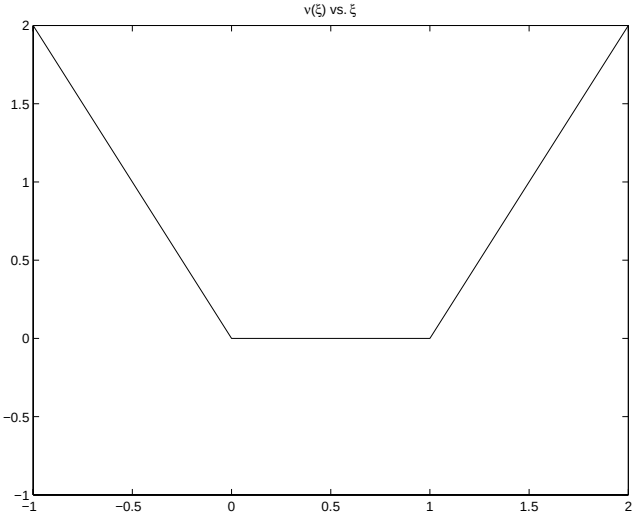


FIG. 5. *The function $\nu(\xi)$ is used for exact penalization as a method to impose the constraint $0 \le u \le 1$ in the minimization of Claim* 1.

minimization problem described above leads to the following Euler–Lagrange equation:

$$\operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right) - \lambda s(x) - \alpha \nu'(u) = 0,$$

where $s(x) = (c_1 - f(x))^2 - (c_2 - f(x))^2$. The following explicit gradient descent scheme was used to solve the last equation:

$$\frac{u^{n+1} - u^n}{\delta t} = D_x^- \left(\frac{D_x^+ u^n}{\sqrt{(D_x^+ u^n)^2 + (D_y^+ u^n)^2 + \varepsilon_1}} \right) + D_y^- \left(\frac{D_y^+ u^n}{\sqrt{(D_x^+ u^n)^2 + (D_y^+ u^n)^2 + \varepsilon_1}} \right) - \lambda s(x) - \alpha \nu'_{\varepsilon_2}(u^n), \tag{14}$$

where $\varepsilon_1, \varepsilon_2 > 0$ are small constants, and $\nu_{\varepsilon_2}(\xi)$ is a regularized version of $\nu(\xi)$ that smooths the latter's kinks at 0 and 1.

The image shown in the Figure 6 is not piecewise constant with two regions; in fact it is not very well approximated by any image that takes only two values. This makes it a challenging test case for the two-phase segmentation problem (images that are already approximately two-valued are easily and very quickly segmented by these algorithms, and thus are easier examples).

Figure 7 shows the result found (i.e., the function u) using (14) to update the unknown function $u(x)$ that represents the two phases, when the given image $f(x)$ is the one shown in Figure 6. The two constants c_1 and c_2 were initially chosen to be 1 and 0, and updated occasionally according to (11); they eventually converged to 0.4666 and 0.0605, respectively. Although the considerations above (in particular, Theorem 2) do not imply that the minimizers of the objective functional turn out to be binary themselves (which would make the thresholding step in the algorithm

FIG. 6. *The given image $f(x)$ used in the two-phase segmentation numerical results discussed in section* 4.

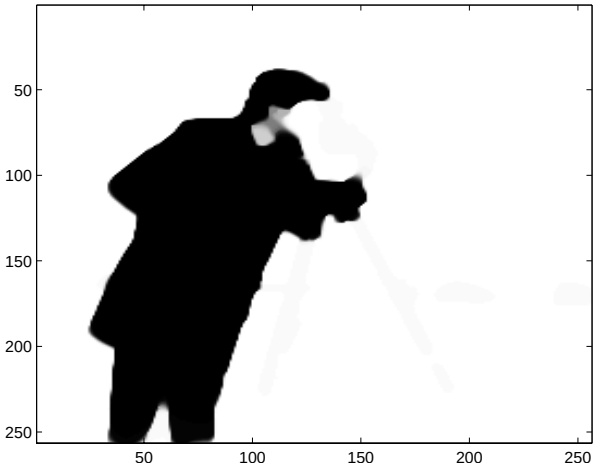


FIG. 7. *The solution $u(x)$ obtained by the numerical scheme of section* 4. *Although our claims do not imply the solution itself turns out to be binary, this seems to be always the case in practice. As can be seen, the computed solution is very close to being binary.*

of Theorem 2 unnecessary), in practice they seem to. Indeed, the image of Figure 7 is very close to being binary. Furthermore, it gets even closer to being binary if the computation is repeated using smaller values of the regularization parameters ε_1 and ε_2 that appear in scheme (14). On the other hand, it might be possible to cook up special given images f and special values λ for which there are nonbinary minimizers; for instance, in the case of the convex formulation (7) of the related geometry problem (4), the simple example of a disk as the given shape leads to nonbinary solutions for a specific choice of the parameter λ, as shown in [6].

Figure 8 displays the histograms of $u(x)$ at intermediate stages of the gradient descent computation. During the evolution, the function u certainly takes a continuum of values; however, as the steady state approaches, the values that u can take accumulate at the extreme ends of its allowed range. In this case, the extreme values

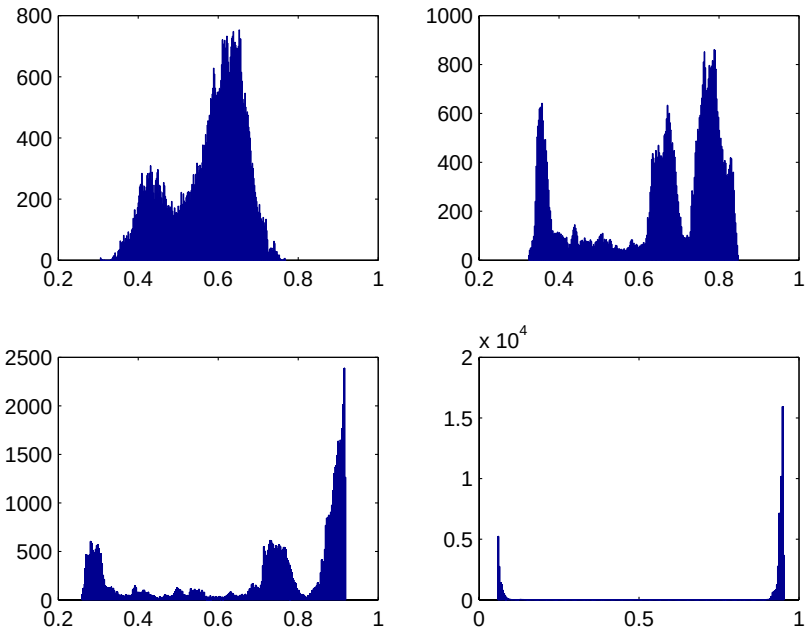


FIG. 8. *Histograms of intermediate solutions* $u^n(x)$ *of the flow in* (14)*, for the example of Figures* 6 *and* 7.

seem to be about 0.04 at the low end and 0.97 at the high end. They are not 0 and 1 because the exact penalty function ν that appears in (14) is regularized.

The theory above says that for fixed c_1 and c_2, all level sets of the function $u(x)$ are minimizers. The table below shows the value of energy (9) computed by taking $\Sigma = \{x : u(x) = \mu\}$ for several different values of μ, where $u(x)$ is the numerical example of Figures 6, 7, and 8, and the constants c_1 and c_2 have the values quoted above.

μ	Energy
0.2	17.7055
0.4	17.6458
0.5	17.6696
0.6	17.6655
0.8	17.6740

For comparison, we note that the energy of a disk centered at the middle of the image with radius a quarter of a side of the image domain has energy of about 112. Thus, even though there is some minor variation among different level sets of the function $u(x)$ (see Figure 9 for a plot of several level contours) and their corresponding energies (due to some of the approximations, such as in the penalty function μ, that were made to get a practical numerical algorithm), the difference in energy between them is quite small; they are all almost minimizers.

Acknowledgment. The authors would like to thank Prof. Robert V. Kohn, from whom they learned the observations of Strang in [23, 24].

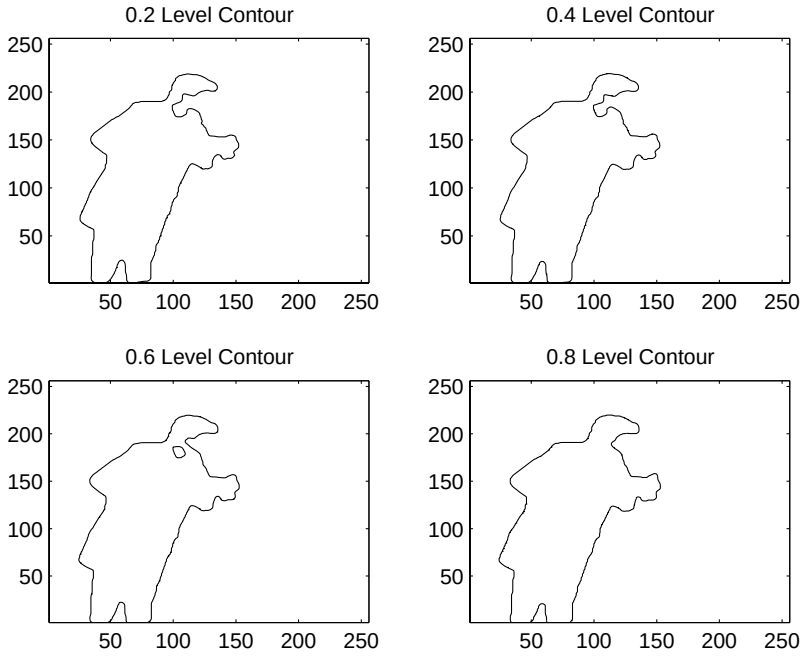


FIG. 9. *Plot of several level contours of the solution obtained. They are all very close to each other.*

REFERENCES

[1] S. ALLINEY, *Digital filters as absolute norm regularizers*, IEEE Trans. Signal Process., 40 (1992), pp. 1548–1562.

[2] S. ALLINEY, *Recursive median filters of increasing order: A variational approach*, IEEE Trans. Signal Process., 44 (1996), pp. 1346–1354.

[3] S. ALLINEY, *A property of the minimum vectors of a regularizing functional defined by means of the absolute norm*, IEEE Trans. Signal Process., 45 (1997), pp. 913–917.

[4] L. AMBROSIO AND V. M. TORTORELLI, *Approximation of functionals depending on jumps by elliptic functionals via Gamma convergence,* Comm. Pure Appl. Math., 43 (1990), pp. 999–1036.

[5] F. CATTÉ, F. DIBOS, AND G. KOEPFLER, *A morphological scheme for mean curvature motion and applications to anisotropic diffusion and motion of level sets*, SIAM J. Numer. Anal., 32 (1995), pp. 1895–1909.

[6] T. F. CHAN AND S. ESEDOḠLU, *Aspects of total variation regularized L^1 function approximation*, SIAM J. Appl. Math., 65 (2005), pp. 1817–1837.

[7] T. F. CHAN AND L. A. VESE, *Active contours without edges*, IEEE Trans. Image Process., 10 (2001), pp. 266–277.

[8] G. DAL MASO, *An Introduction to Γ-convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkhäuser Boston, Boston, MA, 1993.

[9] F. DIBOS AND G. KOEPFLER, *Global total variation minimization*, SIAM J. Numer. Anal., 37 (2000), pp. 646–664.

[10] L. C. EVANS AND R. F. GARIEPY, *Measure Theory and Fine Properties of Functions*, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1992.

[11] W. FLEMING, W. RISHEL, AND R. RISHEL, *An integral formula for total gradient variation*, Arch. Math., 11 (1960), pp. 218–222.

[12] E. GIUSTI, *Minimal Surfaces and Functions of Bounded Variation*, Monogr. Math. 80, Birkhäuser Verlag, Basel, 1984.

[13] J.-B. HIRIART-URRUTY AND C. LEMARECHAL, *Convex Analysis and Minimization Algorithms. I. Fundamentals*, Grundlehren Math. Wiss. 305, Springer-Verlag, New York, 1993.

[14] J.-B. HIRIART-URRUTY AND C. LEMARECHAL, *Convex Analysis and Minimization Algorithms.*

II. *Advanced Theory and Bundle Methods*, Grundlehren Math. Wiss. 306, Springer-Verlag, New York, 1993.

[15] M. Kass, A. Witkin, and D. Terzopoulos, *Snakes: Active contour models*, Internat. J. Comput. Vision, 1 (1987), pp. 321–331.

[16] L. Modica and S. Mortola, *Un esempio di Gamma-convergenza*, Boll. Un. Mat. Ital. B, 14 (1977), pp. 285–299.

[17] D. Mumford and J. Shah, *Optimal approximations by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math., 42 (1989), pp. 577–685.

[18] M. Nikolova, *A variational approach to remove outliers and impulse noise*, J. Math. Imaging and Vision, 20 (2004), pp. 99–120.

[19] M. Nikolova, *Minimizers of cost-functions involving nonsmooth data-fidelity terms. Application to the processing of outliers*, SIAM J. Numer. Anal., 40 (2002), pp. 965–994.

[20] S. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, Appl. Math. Sci., 153, Springer-Verlag, New York, 2003.

[21] S. Osher and J. Sethian, *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.

[22] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.

[23] G. Strang, *L^1 and L^∞ approximation of vector fields in the plane*, in Nonlinear Partial Differential Equations in Applied Science (Tokyo, 1982), North-Holland Math. Stud. 81, North-Holland, Amsterdam, 1983, pp. 273–288.

[24] G. Strang, *Maximal flow through a domain,* Math. Programming, 26 (1983), pp. 123–143.

[25] D. Strong and T. F. Chan, *Edge-preserving and scale-dependent properties of total variation regularization*, Inverse Problems, 19 (2003), pp. S165–S187.

[26] L. A. Vese and T. F. Chan, *A multiphase level set framework for image segmentation using the Mumford and Shah model*, Internat. J. Comput. Vision, 50 (2002), pp. 271–293.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1649–1668

MULTIPLE SCATTERING BY MULTIPLE SPHERES: A NEW PROOF OF THE LLOYD–BERRY FORMULA FOR THE EFFECTIVE WAVENUMBER*

C. M. LINTON† AND P. A. MARTIN‡

Abstract. We provide the first classical derivation of the Lloyd–Berry formula for the effective wavenumber of an acoustic medium filled with a sparse random array of identical small scatterers. Our approach clarifies the assumptions under which the Lloyd–Berry formula is valid. More precisely, we derive an expression for the effective wavenumber which assumes the validity of Lax's quasi-crystalline approximation but makes no further assumptions about scatterer size, and then we show that the Lloyd–Berry formula is obtained in the limit as the scatterer size tends to zero.

Key words. multiple scattering, effective wavenumber, random media, acoustics

AMS subject classifications. 74J20, 78A45, 78A48

DOI. 10.1137/050636401

1. Introduction.

Suppose that we are interested in the scattering of sound by many small scatterers; for example, we might be interested in using ultrasound to determine the quality of certain composites [15], fresh mortar [2], or food products such as mayonnaise [24]. If we knew the shape, size, and location of every scatterer, we could solve the multiple-scattering problem by solving a boundary integral equation, for example [22]. However, usually we do not have this information. Thus, it is common to regard the volume containing the scatterers as a random medium, with certain average (homogenized) properties. Here, we are concerned with finding an *effective wavenumber*, K, that can be used for modeling wave propagation through the scattering volume. This is a classical topic with a large literature: we cite well-known papers by Foldy [7], Lax [18, 19], Waterman and Truell [28], Twersky [26], and Fikioris and Waterman [6], and we refer to the book by Tsang et al. [25] for more information.

A typical problem is the following. The region $z < 0$ is filled with a homogeneous compressible fluid of density ρ and sound-speed c. The region $z > 0$ contains the same fluid and many scatterers; to fix ideas, suppose that the scatterers are identical spheres. Then, a time-harmonic plane wave with wavenumber $k = \omega/c$ (ω is the angular frequency) is incident on the scatterers. The scattered field may be computed exactly for any given configuration (ensemble) of N spheres, but the cost increases as N increases. If the computation can be done, it may be repeated for other configurations, and then the average reflected field could be computed (this is the Monte Carlo approach). Instead of doing this, we shall do some ensemble averaging in order to calculate the average (coherent) field. One result of this is a formula for K.

Foldy [7] considered isotropic point scatterers; this is an appropriate model for

*Received by the editors July 19, 2005; accepted for publication (in revised form) February 13, 2006; published electronically June 19, 2006. This work was partially supported by EPSRC (UK) grant GR/S35585/01.

http://www.siam.org/journals/siap/66-5/63640.html

†Department of Mathematical Sciences, Loughborough University, Leicestershire LE11 3TU, UK (c.m.linton@lboro.ac.uk).

‡Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401 (pamartin@mines.edu).

small sound-soft scatterers. He obtained the formula

$$K^2 = k^2 - 4\pi \mathrm{i} g n_0/k, \tag{1.1}$$

where n_0 is the number of spheres per unit volume and g is the scattering coefficient for an isolated scatterer. The formula (1.1) assumes that the scatterers are independent and that n_0 is small. We are interested in calculating the correction to (1.1) (a term proportional to n_0^2), and this will require saying more about the distribution of the scatterers; specifically, we shall use pair correlations. Thus, our goal is a formula of the form

$$K^2 = k^2 + \delta_1 n_0 + \delta_2 n_0^2, \tag{1.2}$$

with computable expressions for δ_1 and δ_2. Moreover, we do not want to restrict our formula only to sound-soft scatterers.

There is some controversy over the proper value for δ_2. In order to state one such formula, we introduce the *far-field pattern* f. For scattering by one sphere, we have $u_{\mathrm{in}} = \exp(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r})$ for the incident plane wave, where $\boldsymbol{k} = k\hat{\boldsymbol{k}}$, $\boldsymbol{r} = r\hat{\boldsymbol{r}}$, $k = |\boldsymbol{k}|$, and $r = |\boldsymbol{r}|$; the angle of incidence, θ_{in}, is defined by $\cos\theta_{\mathrm{in}} = \hat{\boldsymbol{k}}\cdot\hat{\mathbf{z}}$, where $\hat{\mathbf{z}} = (0,0,1)$ is a unit vector in the z-direction. Then the scattered waves satisfy

$$u_{\mathrm{sc}} \sim (\mathrm{i}kr)^{-1}\mathrm{e}^{\mathrm{i}kr} f(\Theta) \quad \text{as } kr \to \infty, \tag{1.3}$$

where $\cos\Theta = \hat{\boldsymbol{r}}\cdot\hat{\boldsymbol{k}}$. Then, Twersky [26] has obtained (1.2) with

$$\delta_1 = -(4\pi\mathrm{i}/k) f(0) \quad \text{and} \quad \delta_2 = (4\pi^2/k^4)\sec^2\theta_{\mathrm{in}} \left\{ [f(\pi - 2\theta_{\mathrm{in}})]^2 - [f(0)]^2 \right\}. \tag{1.4}$$

The formula for δ_2 involves θ_{in}, so that it gives a different effective wavenumber for different incident fields. The same formulas but with $\theta_{\mathrm{in}} = 0$ (normal incidence) were given by Urick and Ament [27] and by Waterman and Truell [28]:

$$\delta_1 = -(4\pi\mathrm{i}/k) f(0) \quad \text{and} \quad \delta_2 = (4\pi^2/k^4) \left\{ [f(\pi)]^2 - [f(0)]^2 \right\}. \tag{1.5}$$

Other formulas were obtained more recently [14, 30].

In 1967, Lloyd and Berry [21] showed that the formula for δ_2 should be

$$\delta_2 = \frac{4\pi^2}{k^4} \left\{ -[f(\pi)]^2 + [f(0)]^2 + \int_0^\pi \frac{1}{\sin(\theta/2)} \frac{\mathrm{d}}{\mathrm{d}\theta} [f(\theta)]^2 \, \mathrm{d}\theta \right\}, \tag{1.6}$$

with no dependence on θ_{in}. They used methods and language coming from nuclear physics. Thus, in their approach, which they

> call the "resummation method," a point source of waves is considered to be situated in an infinite medium. The scattering series is then written out completely, giving what Lax has called the "expanded" representation. In this expanded representation the ensemble average may be taken exactly [but then] the coherent wave does not exist; the series must be resummed in order to obtain any result at all.

The main purpose of the present paper is to demonstrate that a proper analysis of the semi-infinite model problem (with arbitrary angle of incidence) leads to the Lloyd–Berry formula. Our analysis does not involve "resumming" series or divergent integrals. It builds on a conventional approach, in the spirit of the paper by Fikioris and Waterman [6].

There are two good reasons for giving a new derivation of the Lloyd–Berry formula. First, our analysis clarifies the assumptions that lead to (1.6). Second, erroneous formulas (such as (1.4) or (1.5)) continue to be used widely, perhaps because they are simpler than (1.6) or perhaps because the original derivation in [21] seems suspect. For some representative applications, see [15, 2, 23] and [24, Chapter 4].

The paper begins with a brief summary of some elementary probability theory. The pair-correlation function is introduced, including the notion of "hole correction," which ensures that spheres do not overlap during the averaging process. In section 3, we consider isotropic scatterers and derive the integral equations of Foldy (independent scatterers, no hole correction) and of Lax (hole correction included). Foldy's equation is solved exactly. A method is developed in section 3.2 for obtaining an expression for K which does not require an exact solution of the integral equation, merely an assumption that an effective wavenumber can be used at some distance from the "interface" at $z = 0$ between the homogeneous region ($z < 0$) and the region occupied by many small scatterers ($z > 0$). The virtue of this method is that it succeeds when the governing integral equation cannot be solved exactly. Thus, in section 3.3, we obtain an expression for K from Lax's integral equation; Foldy's approximation is recovered when the hole correction is removed. The same method is used in section 4 but without the restriction to isotropic scatterers. We start with an exact, deterministic theory for acoustic scattering by N spheres; the spheres can be soft, hard, or penetrable. We combine multipole solutions in spherical polar coordinates with an appropriate addition theorem. This method is well known; for some recent applications, see [17, 10, 12]. The exact system of equations is then subjected to ensemble averaging in section 4.3; Lax's "quasi-crystalline approximation" [19] is invoked. This leads to a homogeneous infinite system of linear algebraic equations; the existence of a nontrivial solution determines K. We solve the system for small n_0 and recover the Lloyd–Berry formula.

An analogous theory can be developed in two dimensions and leads to a result that is reminiscent of the Lloyd–Berry formula [20]. However, the three-dimensional calculations described below are much more complicated, as they involve addition theorems for spherical wavefunctions and properties of spherical harmonics. Nevertheless, the final results are rather simple.

2. Some probability theory. In this section, we give a very brief summary of the probability theory needed. For more information, see [7], [18], or Chapter 14 of [13].

Suppose that we have N scatterers located at the points $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N$; denote the configuration of points by $\Lambda_N = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N\}$. Then, the ensemble (or configurational) average of any quantity $F(\boldsymbol{r}|\Lambda_N)$ is defined by

$$\langle F(\boldsymbol{r})\rangle = \int\cdots\int p(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N)\, F(\boldsymbol{r}|\Lambda_N)\, \mathrm{d}V_1 \cdots \mathrm{d}V_N, \tag{2.1}$$

where the integration is over N copies of the volume B_N containing N scatterers. Here, $p(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N)\, \mathrm{d}V_1\, \mathrm{d}V_2 \cdots \mathrm{d}V_N$ is the probability of finding the scatterers in a configuration in which the first scatterer is in the volume element $\mathrm{d}V_1$ about $\boldsymbol{r}_1$, the second scatterer is in the volume element $\mathrm{d}V_2$ about $\boldsymbol{r}_2$, and so on, up to $\boldsymbol{r}_N$. The joint probability distribution $p(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N)$ is normalized so that $\langle 1\rangle = 1$. Similarly, the average of $F(\boldsymbol{r}|\Lambda_N)$ over all configurations for which the first scatterer is fixed

at $\boldsymbol{r}_1$ is given by

$$\langle F(\boldsymbol{r})\rangle_1 = \int\cdots\int p(\boldsymbol{r}_2,\ldots,\boldsymbol{r}_N|\boldsymbol{r}_1)\,F(\boldsymbol{r}|\Lambda_N)\,\mathrm{d}V_2\cdots\mathrm{d}V_N, \tag{2.2}$$

where the conditional probability $p(\boldsymbol{r}_2,\ldots,\boldsymbol{r}_N|\boldsymbol{r}_1)$ is defined by $p(\boldsymbol{r}_1,\boldsymbol{r}_2,\ldots,\boldsymbol{r}_N) = p(\boldsymbol{r}_1)\,p(\boldsymbol{r}_2,\ldots,\boldsymbol{r}_N|\boldsymbol{r}_1)$. If two scatterers are fixed, say the first and the second, we can define

$$\langle F(\boldsymbol{r})\rangle_{12} = \int\cdots\int p(\boldsymbol{r}_3,\ldots,\boldsymbol{r}_N|\boldsymbol{r}_1,\boldsymbol{r}_2)\,F(\boldsymbol{r}|\Lambda_N)\,\mathrm{d}V_3\cdots\mathrm{d}V_N, \tag{2.3}$$

where $p(\boldsymbol{r}_2,\ldots,\boldsymbol{r}_N|\boldsymbol{r}_1) = p(\boldsymbol{r}_2|\boldsymbol{r}_1)\,p(\boldsymbol{r}_3,\ldots,\boldsymbol{r}_N|\boldsymbol{r}_1,\boldsymbol{r}_2)$.

As each of the N scatterers is equally likely to occupy $\mathrm{d}V_1$, the density of scatterers at $\boldsymbol{r}_1$ is $Np(\boldsymbol{r}_1) = n_0$, the (constant) number of scatterers per unit volume. Thus

$$p(\boldsymbol{r}) = n_0/N = |B_N|^{-1}, \tag{2.4}$$

where $|B_N|$ is the volume of B_N. For spheres of radius a, the simplest sensible choice for the pair-correlation function is

$$p(\boldsymbol{r}_2|\boldsymbol{r}_1) = (n_0/N)H(R_{12}-b), \quad \text{where } R_{12} = |\boldsymbol{r}_1-\boldsymbol{r}_2| \tag{2.5}$$

and H is the Heaviside unit function: $H(x) = 1$ for $x > 0$, and $H(x) = 0$ for $x < 0$. The parameter b (the "hole radius") satisfies $b \geq 2a$ so that spheres are not allowed to overlap.

3. Foldy–Lax theory: Isotropic scatterers. Foldy's theory [7] begins with a simplified deterministic model for scattering by N identical scatterers, each of which is supposed to scatter isotropically. Thus, the total field is assumed to be given by the incident field plus a point source at each scattering center, $\boldsymbol{r}_j$:

$$u(\boldsymbol{r}|\Lambda_N) = u_{\mathrm{in}}(\boldsymbol{r}) + g\sum_{j=1}^{N} u_{\mathrm{ex}}(\boldsymbol{r}_j;\boldsymbol{r}_j|\Lambda_N)\,h_0(k|\boldsymbol{r}-\boldsymbol{r}_j|). \tag{3.1}$$

Here, $h_n(w) \equiv h_n^{(1)}(w)$ is a spherical Hankel function, g is the (assumed known) scattering coefficient, and the exciting field u_{ex} is given by

$$u_{\mathrm{ex}}(\boldsymbol{r};\boldsymbol{r}_n|\Lambda_N) = u_{\mathrm{in}}(\boldsymbol{r}) + g\sum_{\substack{j=1\\ j\neq n}}^{N} u_{\mathrm{ex}}(\boldsymbol{r}_j;\boldsymbol{r}_j|\Lambda_N)\,h_0(k|\boldsymbol{r}-\boldsymbol{r}_j|); \tag{3.2}$$

the N numbers $u_{\mathrm{ex}}(\boldsymbol{r}_j;\boldsymbol{r}_j|\Lambda_N)$ $(j = 1,2,\ldots,N)$ required in (3.1) are to be determined by solving the linear system obtained by evaluating (3.2) at $\boldsymbol{r} = \boldsymbol{r}_n$.

If we try to compute the ensemble average of u, using (3.1) and (2.1), we obtain

$$\langle u(\boldsymbol{r})\rangle = u_{\mathrm{in}}(\boldsymbol{r}) + gn_0\int_{B_N}\langle u_{\mathrm{ex}}(\boldsymbol{r}_1)\rangle_1\,h_0(k|\boldsymbol{r}-\boldsymbol{r}_1|)\,\mathrm{d}V_1, \tag{3.3}$$

where we have used (2.2), (2.4), and the indistinguishability of the scatterers. For $\langle u_{\mathrm{ex}}(\boldsymbol{r}_1)\rangle_1$, we obtain

$$\langle u_{\mathrm{ex}}(\boldsymbol{r})\rangle_1 = u_{\mathrm{in}}(\boldsymbol{r}) + g(N-1)\int_{B_N} p(\boldsymbol{r}_2|\boldsymbol{r}_1)\,\langle u_{\mathrm{ex}}(\boldsymbol{r}_2)\rangle_{12}\,h_0(k|\boldsymbol{r}-\boldsymbol{r}_2|)\,\mathrm{d}V_2, \tag{3.4}$$

where we have used (2.3) and (3.2). Equations (3.3) and (3.4) are the first two in a hierarchy, involving more and more complicated information on the statistics of the scatterer distribution. In practice, the hierarchy is broken using an additional assumption. At the lowest level, we have Foldy's assumption,

$$\langle u_{\rm ex}(\boldsymbol{r})\rangle_1 \simeq \langle u(\boldsymbol{r})\rangle, \tag{3.5}$$

at least in the neighborhood of $\boldsymbol{r}_1$. When this is used in (3.3), we obtain

$$\langle u(\boldsymbol{r})\rangle = u_{\rm in}(\boldsymbol{r}) + gn_0 \int_{B_N} \langle u(\boldsymbol{r}_1)\rangle\, h_0(k|\boldsymbol{r}-\boldsymbol{r}_1|)\,{\rm d}V_1, \quad \boldsymbol{r}\in B_N. \tag{3.6}$$

We call this *Foldy's integral equation* for $\langle u\rangle$. The integral on the right-hand side is an acoustic volume potential. Hence, an application of (∇^2+k^2) to (3.6) eliminates the incident field and shows that $(\nabla^2+K^2)\langle u\rangle = 0$ in B_N, where K^2 is given by Foldy's formula, (1.1).

At the next level, we have Lax's quasi-crystalline assumption (QCA) [19],

$$\langle u_{\rm ex}(\boldsymbol{r})\rangle_{12} \simeq \langle u_{\rm ex}(\boldsymbol{r})\rangle_2. \tag{3.7}$$

When this is used in (3.4) evaluated at $\boldsymbol{r}=\boldsymbol{r}_1$, we obtain

$$v(\boldsymbol{r}) = u_{\rm in}(\boldsymbol{r}) + g(N-1)\int_{B_N} p(\boldsymbol{r}_1|\boldsymbol{r})\, v(\boldsymbol{r}_1)\, h_0(k|\boldsymbol{r}-\boldsymbol{r}_1|)\,{\rm d}V_1, \quad \boldsymbol{r}\in B_N, \tag{3.8}$$

where $v(\boldsymbol{r}) = \langle u_{\rm ex}(\boldsymbol{r})\rangle_1$. We call this *Lax's integral equation*.

In what follows, we let $N\to\infty$ so that $B_N\to B_\infty$, a semi-infinite region, $z>0$.

3.1. Foldy's integral equation: Exact treatment. Consider a plane wave at oblique incidence, so that

$$u_{\rm in} = \exp({\rm i}\boldsymbol{k}\cdot\boldsymbol{r}) = {\rm e}^{{\rm i}\alpha z}\exp({\rm i}\boldsymbol{k}_T\cdot\boldsymbol{q}), \tag{3.9}$$

where $\boldsymbol{r}=(x,y,z)$, $\boldsymbol{q}=(x,y,0)$, $\boldsymbol{k}=\boldsymbol{k}_T+\alpha\hat{\mathbf{z}}$, $\hat{\mathbf{z}}=(0,0,1)$, the wavenumber vector $\boldsymbol{k}$ is given in spherical polar coordinates by

$$\boldsymbol{k} = k\hat{\boldsymbol{k}} \quad \text{with} \quad \hat{\boldsymbol{k}} = (\sin\theta_{\rm in}\,\cos\phi_{\rm in},\ \sin\theta_{\rm in}\,\sin\phi_{\rm in},\ \cos\theta_{\rm in}), \quad 0\le\theta_{\rm in}<\pi/2,$$

$\alpha = k\cos\theta_{\rm in}$, and $\boldsymbol{k}_T$ is the transverse wavenumber vector, satisfying $\boldsymbol{k}_T\cdot\hat{\mathbf{z}}=0$.

For a semi-infinite domain B_∞ ($z>0$), Foldy's integral equation (3.6) becomes

$$\langle u(x,y,z)\rangle = u_{\rm in}(x,y,z) + gn_0\int_0^\infty\int_{-\infty}^\infty\int_{-\infty}^\infty \langle u(x+X,y+Y,z_1)\rangle\, h_0(k\varrho_1)\,{\rm d}X\,{\rm d}Y\,{\rm d}z_1,$$

for $0\le|x|<\infty$, $0\le|y|<\infty$, and $z>0$, where $\varrho_1=\sqrt{X^2+Y^2+(z-z_1)^2}$. This equation can be solved exactly. Thus, writing

$$\langle u(x,y,z)\rangle = U(z)\,\exp({\rm i}\boldsymbol{k}_T\cdot\boldsymbol{q}), \quad 0\le|\boldsymbol{q}|<\infty, \quad z>0, \tag{3.10}$$

we obtain

$$U(z) = {\rm e}^{{\rm i}\alpha z} + gn_0\int_0^\infty\int_{-\infty}^\infty\int_{-\infty}^\infty U(z_1)\,h_0(k\varrho_1)\,\exp({\rm i}\boldsymbol{k}_T\cdot\mathbf{Q})\,{\rm d}X\,{\rm d}Y\,{\rm d}z_1 \tag{3.11}$$

for $z > 0$, where $\mathbf{Q} = (X, Y, 0)$.

In Appendix B, it is shown that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_0(k\varrho_1)\,\exp\left(\mathrm{i}\boldsymbol{k}_T\cdot\mathbf{Q}\right)\mathrm{d}X\,\mathrm{d}Y = \frac{2\pi}{k\alpha}\,\mathrm{e}^{\mathrm{i}\alpha|z-z_1|}. \tag{3.12}$$

Thus, we see that U solves

$$U(z) = \mathrm{e}^{\mathrm{i}\alpha z} + \frac{2\pi g n_0}{k\alpha}\int_0^{\infty} U(z_1)\,\mathrm{e}^{\mathrm{i}\alpha|z-z_1|}\,\mathrm{d}z_1, \quad z > 0. \tag{3.13}$$

Now, set $U(z) = U_0\,\mathrm{e}^{\mathrm{i}\lambda z}$, so that (3.13) gives

$$U_0\,\mathrm{e}^{\mathrm{i}\lambda z} - \mathrm{e}^{\mathrm{i}\alpha z} = \frac{2\pi g n_0}{\mathrm{i}k\alpha}U_0\left(\frac{2\alpha\mathrm{e}^{\mathrm{i}\lambda z}}{\lambda^2-\alpha^2} - \frac{\mathrm{e}^{\mathrm{i}\alpha z}}{\lambda-\alpha}\right),$$

where we have assumed that $\operatorname{Im}\lambda > 0$. If we compare the coefficients of $\mathrm{e}^{\mathrm{i}\lambda z}$, we see that U_0 cancels, leaving

$$\lambda^2 - \alpha^2 = -4\pi\mathrm{i}g n_0/k, \tag{3.14}$$

which determines λ. Then, the coefficients of $\mathrm{e}^{\mathrm{i}\alpha z}$ give $U_0 = 2\alpha/(\lambda+\alpha)$. A similar method can be used to find $\langle u\rangle$ when B_∞ is a slab of finite thickness, $0 < z < h$.

It is natural to define an effective wavenumber vector by

$$\begin{aligned}\mathbf{K} &= K(\sin\vartheta\cos\varphi,\ \sin\vartheta\sin\varphi,\ \cos\vartheta) = K\hat{\mathbf{K}}\\ &= (k\sin\theta_{\mathrm{in}}\cos\phi_{\mathrm{in}},\ k\sin\theta_{\mathrm{in}}\sin\phi_{\mathrm{in}},\ \lambda),\end{aligned} \tag{3.15}$$

whence

$$\lambda = K\cos\vartheta \quad\text{and}\quad K\sin\vartheta = k\sin\theta_{\mathrm{in}}. \tag{3.16}$$

The last equality is recognized as Snell's law, even though K and ϑ are complex, with $\operatorname{Im}K > 0$. Hence, we see that

$$\lambda^2 - \alpha^2 = K^2 - k^2, \tag{3.17}$$

whence (3.14) reduces to Foldy's formula (1.1).

3.2. Foldy's integral equation: Alternative treatment. We have seen that Foldy's integral equation can be solved exactly, and that the solution process has two parts: first find λ (and hence the effective wavenumber) and then find U_0. In fact, λ can be found without finding the complete solution; the reason for pursuing this is that we cannot usually find exact solutions. Thus, consider (3.13), and suppose that

$$U(z) = U_0\mathrm{e}^{\mathrm{i}\lambda z} \quad \text{for } z > \ell,$$

where U_0, λ, and ℓ are unknown. To proceed, we need say nothing about the solution U in the "boundary layer" $0 < z < \ell$. Now, evaluate the integral equation for $z > \ell$; we find that

$$\begin{aligned}U_0\,\mathrm{e}^{\mathrm{i}\lambda z} - \mathrm{e}^{\mathrm{i}\alpha z} &= \frac{2\pi g n_0}{k\alpha}\,\mathrm{e}^{\mathrm{i}\alpha z}\int_0^{\ell} U(t)\,\mathrm{e}^{-\mathrm{i}\alpha t}\,\mathrm{d}t + \frac{2\pi g n_0}{k\alpha}\int_{\ell}^{\infty} U(t)\,\mathrm{e}^{\mathrm{i}\alpha|z-t|}\,\mathrm{d}t\\ &= \mathcal{A}\,\mathrm{e}^{\mathrm{i}\lambda z} + \mathcal{B}\,\mathrm{e}^{\mathrm{i}\alpha z} \quad \text{for } z > \ell,\end{aligned}$$

where $\mathcal{A} = -4\pi \mathrm{i} g n_0 U_0/[k(\lambda^2 - \alpha^2)]$ and

$$\mathcal{B} = \frac{2\pi g n_0}{k\alpha} \int_0^\ell U(t)\, \mathrm{e}^{-\mathrm{i}\alpha t}\, \mathrm{d}t + \frac{2\pi \mathrm{i} g n_0 U_0}{k\alpha(\lambda - \alpha)}\, \mathrm{e}^{\mathrm{i}(\lambda-\alpha)\ell}.$$

Then, setting $U_0 = \mathcal{A}$ gives (3.14) again, without knowing the solution U everywhere. This basic method will be used again below.

3.3. Lax's integral equation. Using (2.5) for $p(\boldsymbol{r}_1|\boldsymbol{r})$ in (3.8) gives

$$v(\boldsymbol{r}) = u_{\mathrm{in}}(\boldsymbol{r}) + g n_0 \frac{N-1}{N} \int_{B_N^b} v(\boldsymbol{r}_1)\, h_0(kR_1)\, \mathrm{d}\boldsymbol{r}_1, \quad \boldsymbol{r} \in B_N, \tag{3.18}$$

where $B_N^b(\boldsymbol{r}) = \{\boldsymbol{r}_1 \in B_N : R_1 = |\boldsymbol{r} - \boldsymbol{r}_1| > b\}$, which is B_N with a (possibly incomplete) ball excluded.

Let $N \to \infty$ and take an incident plane wave, (3.9), giving

$$v(x,y,z) = \mathrm{e}^{\mathrm{i}\alpha z} \exp(\mathrm{i}\boldsymbol{k}_T \cdot \boldsymbol{q}) + g n_0 \int_{z_1>0,\ \varrho_1>b} v(x+X, y+Y, z_1)\, h_0(k\varrho_1)\, \mathrm{d}X\, \mathrm{d}Y\, \mathrm{d}z_1,$$

for $0 \le |\boldsymbol{q}| < \infty$ and $z > 0$. As in section 3.1, we write

$$v(x,y,z) = V(z)\, \exp(\mathrm{i}\boldsymbol{k}_T \cdot \boldsymbol{q}), \quad 0 \le |\boldsymbol{q}| < \infty, \quad z > 0, \tag{3.19}$$

giving

$$V(z) = \mathrm{e}^{\mathrm{i}\alpha z} + g n_0 \int_{z_1>0,\ \varrho_1>b} V(z_1)\, h_0(k\varrho_1)\, \exp(\mathrm{i}\boldsymbol{k}_T \cdot \mathbf{Q})\, \mathrm{d}X\, \mathrm{d}Y\, \mathrm{d}z_1 \tag{3.20}$$

for $0 \le |\boldsymbol{q}| < \infty$ and $z > 0$, where $\mathbf{Q} = (X, Y, 0)$. Then, using (3.12), we see that V solves

$$V(z) = \mathrm{e}^{\mathrm{i}\alpha z} + g n_0 \int_0^\infty V(z_1)\, \mathcal{L}(z - z_1)\, \mathrm{d}z_1, \quad z > 0, \tag{3.21}$$

where the kernel, $\mathcal{L}(z - z_1)$, is given by

$$\begin{aligned} \mathcal{L}(Z) &= \frac{2\pi}{k\alpha} \mathrm{e}^{\mathrm{i}\alpha|Z|} - \int_0^{c(Z)} \int_0^{2\pi} h_0(k\sqrt{Q^2+Z^2}) \mathrm{e}^{\mathrm{i}kQ \sin\theta_{\mathrm{in}} \cos(\Phi - \phi_{\mathrm{in}})} Q\, \mathrm{d}\Phi\, \mathrm{d}Q \\ &= \frac{2\pi}{k\alpha}\, \mathrm{e}^{\mathrm{i}\alpha|Z|} - 2\pi \int_0^{c(Z)} h_0(k\sqrt{Q^2+Z^2})\, J_0(kQ\sin\theta_{\mathrm{in}})\, Q\, \mathrm{d}Q \end{aligned} \tag{3.22}$$

with $c(Z) = \sqrt{b^2 - Z^2}\, H(b - |Z|)$; here, J_n is a Bessel function, and we have written the double integral over X and Y in (3.20) as an integral over all X and Y minus an integral through the cross section of the ball at z, if necessary.

We have been unable to solve (3.21) exactly. However, the alternative method described in section 3.2 can be used. Thus, let us suppose that

$$V(z) = V_0\, \mathrm{e}^{\mathrm{i}\lambda z} \quad \text{for } z > \ell, \tag{3.23}$$

where V_0, λ, and ℓ are unknown. Then, consider (3.21) for $z > \ell + b$, so that the interval $|z - z_1| < b$ is entirely within the range $z_1 > \ell$. Using (3.22), (3.21) gives

$$\begin{aligned} \frac{V_0\, \mathrm{e}^{\mathrm{i}\lambda z} - \mathrm{e}^{\mathrm{i}\alpha z}}{g n_0} &= \frac{2\pi}{k\alpha}\, \mathrm{e}^{\mathrm{i}\alpha z} \int_0^\ell V(t)\, \mathrm{e}^{-\mathrm{i}\alpha t}\, \mathrm{d}t + \frac{2\pi}{k\alpha} \int_\ell^\infty V(t)\, \mathrm{e}^{\mathrm{i}\alpha|z-t|}\, \mathrm{d}t \\ &\quad - 2\pi \int_{z-b}^{z+b} V(t) \int_0^{c(z-t)} h_0(k\sqrt{Q^2 + (z-t)^2})\, J_0(kQ\sin\theta_{\mathrm{in}})\, Q\, \mathrm{d}Q\, \mathrm{d}t \end{aligned} \tag{3.24}$$

for $z > \ell + b$. Equation (3.23) can be used in the second and third integrals. The second integral is elementary, and has the value

$$\frac{2\pi \mathrm{i} V_0}{k\alpha}\left\{\frac{\mathrm{e}^{\mathrm{i}(\lambda-\alpha)\ell}}{\lambda-\alpha}\,\mathrm{e}^{\mathrm{i}\alpha z} - \frac{2\alpha}{\lambda^2-\alpha^2}\,\mathrm{e}^{\mathrm{i}\lambda z}\right\}.$$

Denote the third integral in (3.24) by I_3; we have

$$\begin{aligned} I_3 &= -2\pi V_0 \int_{-b}^{b} \mathrm{e}^{\mathrm{i}\lambda(z+\xi)} \int_0^{\sqrt{b^2-\xi^2}} h_0(k\sqrt{Q^2+\xi^2})\, J_0(kQ\sin\theta_{\mathrm{in}})\, Q\,\mathrm{d}Q\,\mathrm{d}\xi \\ &= -2\pi V_0\,\mathrm{e}^{\mathrm{i}\lambda z} \int_0^{\pi}\int_0^{b} \mathrm{e}^{\mathrm{i}\lambda r\cos\theta}\, h_0(kr)\, J_0(kr\sin\theta\sin\theta_{\mathrm{in}})\, r^2\sin\theta\,\mathrm{d}r\,\mathrm{d}\theta \\ &= -V_0\,\mathrm{e}^{\mathrm{i}\lambda z} \int_0^{2\pi}\int_0^{\pi}\int_0^{b} \mathrm{e}^{\mathrm{i}r[\lambda\cos\theta + k\sin\theta\sin\theta_{\mathrm{in}}\cos(\phi-\phi_{\mathrm{in}})]}\, h_0(kr)\, r^2\sin\theta\,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\phi. \end{aligned}$$

Using (3.16), the exponent simplifies to $\mathbf{K}\cdot\boldsymbol{r}$, whence

$$\begin{aligned} I_3 &= -V_0\,\mathrm{e}^{\mathrm{i}\lambda z} \int_{r<b} \exp(\mathrm{i}\mathbf{K}\cdot\boldsymbol{r}) h_0(k|\boldsymbol{r}|)\,\mathrm{d}V(\boldsymbol{r}) \\ &= -2\pi V_0\,\mathrm{e}^{\mathrm{i}\lambda z} \int_0^{b}\int_0^{\pi} \frac{\mathrm{e}^{\mathrm{i}kr}}{\mathrm{i}kr}\,\mathrm{e}^{\mathrm{i}Kr\cos\theta}\, r^2\sin\theta\,\mathrm{d}\theta\,\mathrm{d}r \\ &= \frac{2\pi V_0}{kK}\,\mathrm{e}^{\mathrm{i}\lambda z}\int_0^{b}\mathrm{e}^{\mathrm{i}kr}\left(\mathrm{e}^{\mathrm{i}Kr}-\mathrm{e}^{-\mathrm{i}Kr}\right)\mathrm{d}r = \frac{4\pi\mathrm{i}V_0}{k(K^2-k^2)}\,\mathrm{e}^{\mathrm{i}\lambda z}\{1-\mathcal{N}_0(Kb)\}, \end{aligned}$$

where $\mathcal{N}_0(x) = \mathrm{e}^{\mathrm{i}kb}\{\cos x - \mathrm{i}(kb/x)\sin x\}$. Using these results in (3.24) and noting (3.17), we obtain

$$V_0\,\mathrm{e}^{\mathrm{i}\lambda z} - \mathrm{e}^{\mathrm{i}\alpha z} = \mathcal{A}\,\mathrm{e}^{\mathrm{i}\lambda z} + \mathcal{B}\,\mathrm{e}^{\mathrm{i}\alpha z} \quad \text{for } z > \ell + b,$$

where

$$\mathcal{A} = \frac{4\pi\mathrm{i}gn_0V_0}{k(k^2-K^2)}\,\mathcal{N}_0(Kb), \qquad \mathcal{B} = \frac{2\pi g n_0}{k\alpha}\int_0^{\ell} V(t)\,\mathrm{e}^{-\mathrm{i}\alpha t}\,\mathrm{d}t + \frac{2\pi\mathrm{i}gn_0V_0}{k\alpha(\lambda-\alpha)}\,\mathrm{e}^{\mathrm{i}(\lambda-\alpha)\ell}.$$

For a solution, we must have $\mathcal{A} = V_0$, whence

$$K^2 = k^2 - 4\pi\mathrm{i}g(n_0/k)\,\mathcal{N}_0(Kb), \tag{3.25}$$

which is a nonlinear equation for K. Notice that this equation does not depend on the angle of incidence, θ_{in}.

We have $\mathcal{N}_0(Kb) \to 1$ as $Kb \to 0$ so that, in this limit, we recover Foldy's formula for the effective wavenumber, (1.1).

Let us solve (3.25) for small n_0. (We could use the dimensionless volume fraction $\frac{4}{3}\pi a^3 n_0$, but it is customary to use n_0.) Begin by writing

$$K^2 = k^2 + \delta_1 n_0 + \delta_2 n_0^2 + \cdots, \tag{3.26}$$

where δ_1 and δ_2 are to be found; for δ_1, we expect to obtain the result given by (1.1). It follows that $K = k + \frac{1}{2}\delta_1 n_0/k + O(n_0^2)$ and then

$$\begin{aligned} \mathcal{N}_0(Kb) &= \mathcal{N}_0(kb) + (Kb-kb)\mathcal{N}_0'(kb) + \cdots \\ &= 1 - \tfrac{1}{2}\mathrm{i}b(n_0/k)\delta_1 d_0(kb) + O(n_0^2), \end{aligned}$$

where $d_0(x) = 1 - x^{-1}\mathrm{e}^{\mathrm{i}x}\sin x$. When this approximation for $\mathcal{N}_0(Kb)$ is used in (3.25), we obtain

$$K^2 = k^2 - 4\pi\mathrm{i}gn_0/k - 2\pi bg(n_0/k)^2\delta_1 d_0(kb).$$

Comparison of this formula with (3.26) gives $\delta_1 = -4\pi\mathrm{i}g/k$ (as expected) and $\delta_2 = 8\pi^2\mathrm{i}g^2bk^{-3}d_0(kb)$, so that we obtain the approximation

$$K^2 = k^2 - \frac{4\pi\mathrm{i}g}{k}n_0 + \frac{8\mathrm{i}b(\pi g n_0)^2}{k^3}\left(1 - \mathrm{e}^{\mathrm{i}kb}\frac{\sin kb}{kb}\right). \tag{3.27}$$

(Recall that a common choice for the hole radius is $b = 2a$.) As far as we know, the formula (3.27) is new. Note that the second-order term in (3.27) vanishes in the limit $kb \to 0$.

4. A finite array of identical spheres: Exact theory. Let O be the origin of three-dimensional Cartesian coordinates, so that a typical point has position vector $\boldsymbol{r} = (x, y, z)$ with respect to O. Define spherical polar coordinates (r, θ, ϕ) at O, so that $\boldsymbol{r} = r\hat{\boldsymbol{r}} = r(\sin\theta\cos\phi,\ \sin\theta\sin\phi,\ \cos\theta)$. We consider N identical spheres, S_j, $j = 1, 2, \ldots, N$. The sphere S_j has radius a and center O_j at $\boldsymbol{r} = \boldsymbol{r}_j$. We define spherical polar coordinates $(\rho_j, \theta_j, \phi_j)$ at O_j, so that $\boldsymbol{r} = \boldsymbol{\rho}_j + \boldsymbol{r}_j$ with

$$\boldsymbol{\rho}_j = \rho_j\hat{\boldsymbol{\rho}}_j = \rho_j(\sin\theta_j\cos\phi_j,\ \sin\theta_j\sin\phi_j,\ \cos\theta_j).$$

We assume that $\theta_j = 0$ is in the z-direction ($\theta = 0$).

Exterior to the spheres the pressure field is u, where

$$\nabla^2 u + k^2 u = 0. \tag{4.1}$$

Inside S_j, the field is u_j, where

$$\nabla^2 u_j + \kappa^2 u_j = 0, \tag{4.2}$$

$\kappa = \omega/\tilde{c}$, and $\tilde{c}$ is the sound speed inside the spheres. The transmission conditions on the spheres are

$$u = u_j, \qquad \frac{1}{\rho}\frac{\partial u}{\partial \rho_j} = \frac{1}{\tilde{\rho}}\frac{\partial u_j}{\partial \rho_j} \quad \text{on} \quad \rho_j = a, \quad j = 1, \ldots, N, \tag{4.3}$$

where $\tilde{\rho}$ is the fluid density inside the spheres.

A plane wave, given by (3.9), is incident on the spheres. The problem is to calculate the scattered field outside the spheres, defined as $u_{\mathrm{sc}} = u - u_{\mathrm{in}}$. We start with just one sphere, in order to fix our notation.

4.1. Scattering by one sphere. For the incident plane wave, we have

$$u_{\mathrm{in}}(\boldsymbol{r}) = \exp(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}) = 4\pi\sum_{n,m}\mathrm{i}^n\hat{\psi}_n^m(\boldsymbol{r})\,\overline{Y_n^m(\hat{\boldsymbol{k}})}, \tag{4.4}$$

where $\hat{\psi}_n^m(\boldsymbol{r}) = j_n(kr)Y_n^m(\hat{\boldsymbol{r}})$, $j_n(w)$ is a spherical Bessel function, $Y_n^m(\hat{\boldsymbol{r}}) = Y_n^m(\theta, \phi)$ is a spherical harmonic (see Appendix A), the overbar denotes complex conjugation, and we have used the shorthand notation

$$\sum_{n,m} \equiv \sum_{n=0}^{\infty}\sum_{m=-n}^{n}.$$

With our choice of normalization, the spherical harmonics are orthonormal; see (A.1).

For the scattered and interior fields, we can write

$$u_{\rm sc}(\boldsymbol{r}) = 4\pi \sum_{n,m} \mathrm{i}^n A_n^m Z_n \psi_n^m(\boldsymbol{r}) \quad \text{and} \quad u_{\rm int}(\boldsymbol{r}) = 4\pi \sum_{n,m} \mathrm{i}^n B_n^m j_n(\kappa r)\, Y_n^m(\hat{\boldsymbol{r}}),$$

respectively, where $\psi_n^m(\boldsymbol{r}) = h_n(kr)\, Y_n^m(\hat{\boldsymbol{r}})$, the coefficients A_n^m and B_n^m are to be found, and the factor

$$Z_n = \frac{q j_n'(ka) j_n(\kappa a) - j_n(ka) j_n'(\kappa a)}{q h_n'(ka) j_n(\kappa a) - h_n(ka) j_n'(\kappa a)}, \tag{4.5}$$

with $q = \tilde{\rho}\tilde{c}/(\rho c)$, has been introduced for later convenience. Then, the transmission conditions on $r = a$ yield A_n^m and B_n^m; in particular, we obtain $A_n^m = -\overline{Y_n^m(\hat{\boldsymbol{k}})}$. Also, the far-field pattern, defined by (1.3), is given by

$$f(\Theta) = 4\pi \sum_{n,m} Z_n A_n^m\, Y_n^m(\hat{\boldsymbol{r}}) = -\sum_{n=0}^{\infty} (2n+1) Z_n P_n(\cos\Theta), \tag{4.6}$$

where $\cos\Theta = \hat{\boldsymbol{r}} \cdot \hat{\boldsymbol{k}}$ and we have used (A.3) in order to evaluate the sum over m. Note that we recover the sound-soft results in the limit $q \to 0$, whereas the limit $q \to \infty$ gives the sound-hard results.

4.2. Scattering by N spheres. A phase factor for each sphere is defined by $I_j = \exp(\mathrm{i}\boldsymbol{k} \cdot \boldsymbol{r}_j)$, and then we can write

$$u_{\rm in} = I_j \exp(\mathrm{i}\boldsymbol{k} \cdot \boldsymbol{\rho}_j) = 4\pi I_j \sum_{n,m} \mathrm{i}^n \hat{\psi}_n^m(\boldsymbol{\rho}_j)\, \overline{Y_n^m(\hat{\boldsymbol{k}})}. \tag{4.7}$$

We seek a solution to (4.1) and (4.2) in the form

$$u = u_{\rm in} + 4\pi \sum_{j=1}^{N} \sum_{n,m} \mathrm{i}^n A_{nj}^m Z_n \psi_n^m(\boldsymbol{\rho}_j), \qquad u_j = 4\pi \sum_{n,m} \mathrm{i}^n B_{nj}^m\, j_n(\kappa\rho_j)\, Y_n^m(\hat{\boldsymbol{\rho}}_j),$$

for some set of unknown complex coefficients A_{nj}^m and B_{nj}^m.

Now, in order to apply the transmission conditions on each sphere, we shall need an *addition theorem*. Thus, given vectors $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c} = \boldsymbol{a} + \boldsymbol{b}$, we have

$$\psi_n^m(\boldsymbol{c}) = \sum_{\nu,\mu} S_{n\nu}^{m\mu}(\boldsymbol{b})\, \hat{\psi}_\nu^\mu(\boldsymbol{a}) \quad \text{for } |\boldsymbol{a}| < |\boldsymbol{b}|, \tag{4.8}$$

where the separation matrix $S_{\nu n}^{\mu m}$ is given by

$$S_{\nu n}^{\mu m}(\mathbf{R}) = 4\pi \mathrm{i}^{n-\nu} (-1)^m \sum_q \mathrm{i}^q \psi_q^{\mu - m}(\mathbf{R})\, \mathcal{G}(n, m; \nu, -\mu; q). \tag{4.9}$$

In this formula, $\mathcal{G}$ is a Gaunt coefficient (defined by (A.5)), and the sum has a finite number of terms; in fact, q runs from $|n - \nu|$ to $(n + \nu)$ in steps of 2, so that

$$(q + n + \nu) \text{ is even.} \tag{4.10}$$

For more information on the addition theorem, see [3, 10, 11] and references therein.

Let $\mathbf{R}_{sj} = \boldsymbol{r}_s - \boldsymbol{r}_j = \boldsymbol{\rho}_j - \boldsymbol{\rho}_s$ be the position vector of O_s with respect to O_j. Then, provided that $\rho_s < R_{sj} = |\mathbf{R}_{sj}|$ for all j, we can write the field exterior to the sphere S_s as

$$u = 4\pi \sum_{n,m} \mathrm{i}^n \left\{ I_s \hat{\psi}_n^m(\boldsymbol{\rho}_s)\, \overline{Y_n^m(\hat{\boldsymbol{k}})} + A_{ns}^m Z_n \psi_n^m(\boldsymbol{\rho}_s) \right\} + 4\pi \sum_{n,m} \hat{\psi}_n^m(\boldsymbol{\rho}_s) \sum_{\substack{j=1 \\ j\neq s}}^{N} \sum_{\nu,\mu} \mathrm{i}^\nu A_{\nu j}^\mu Z_\nu S_{\nu n}^{\mu m}(\mathbf{R}_{sj}). \tag{4.11}$$

The geometrical restriction implies that this expression is valid near the surface of S_s, and so (4.11) can be used to apply the transmission conditions on $\rho_s = a$. Thus, after using the orthogonality of the functions $Y_n^m(\hat{\boldsymbol{\rho}}_s)$, (A.1), and then eliminating the coefficients B_{nj}^m, we obtain

$$A_{ns}^m + \sum_{\substack{j=1 \\ j\neq s}}^{N} \sum_{\nu,\mu} \mathrm{i}^{\nu-n} A_{\nu j}^\mu Z_\nu S_{\nu n}^{\mu m}(\mathbf{R}_{sj}) = -I_s \overline{Y_n^m(\hat{\boldsymbol{k}})}, \qquad \begin{array}{l} s = 1, 2, \ldots, N, \\ n = 0, 1, 2, \ldots, \\ m = -n, \ldots, n, \end{array} \tag{4.12}$$

an infinite linear system of equations for A_{nj}^m. Note that the quantities q, κ, and a enter the equations only through the terms Z_ν.

4.3. Arrays of spheres: Averaged equations. The above analysis applies to a specific configuration of scatterers. Now we take ensemble averages. Specifically, setting $s = 1$ in (4.12) and then taking the conditional average, using (2.5), we obtain

$$\langle A_{n1}^m \rangle_1 + n_0 \frac{N-1}{N} \sum_{\nu,\mu} \mathrm{i}^{\nu-n} Z_\nu \int_{B_N:\, R_{12}>b} S_{\nu n}^{\mu m}(\mathbf{R}_{12})\, \langle A_{\nu 2}^\mu \rangle_{12}\, \mathrm{d}V_2 = -I_1 \overline{Y_n^m(\hat{\boldsymbol{k}})}, \tag{4.13}$$

for $n = 0, 1, 2, \ldots$ and $m = -n, \ldots, n$. Then, we let $N \to \infty$ so that B_N becomes the half-space $z > 0$, and invoke Lax's QCA, (3.7). This implies that $\langle A_{n2}^m \rangle_{12} = \langle A_{n2}^m \rangle_2$. Hence, (4.13) reduces to

$$\langle A_{n1}^m \rangle_1 + n_0 \sum_{\nu,\mu} \mathrm{i}^{\nu-n} Z_\nu \int_{z_2>0,\, R_{12}>b} S_{\nu n}^{\mu m}(\mathbf{R}_{12})\, \langle A_{\nu 2}^\mu \rangle_2\, \mathrm{d}V_2 = -I_1 \overline{Y_n^m(\hat{\boldsymbol{k}})}, \tag{4.14}$$

for $n = 0, 1, 2, \ldots$ and $m = -n, \ldots, n$. As $I_1 = \exp(\mathrm{i}\boldsymbol{k} \cdot \boldsymbol{r}_1) = \mathrm{e}^{\mathrm{i}\alpha z_1} \exp(\mathrm{i}\boldsymbol{k}_T \cdot \boldsymbol{q}_1)$ with $\alpha = k \cos\theta_{\mathrm{in}}$ and $\boldsymbol{q}_s = (x_s, y_s, 0)$, we seek a solution to (4.14) in the form

$$\langle A_{ns}^m \rangle_s = \Phi_n^m(z_s)\, \exp(\mathrm{i}\boldsymbol{k}_T \cdot \boldsymbol{q}_s) \tag{4.15}$$

so that

$$\Phi_n^m(z_1) + n_0 \sum_{\nu,\mu} \mathrm{i}^{\nu-n} Z_\nu \int_{z_2>0,\, R_{12}>b} S_{\nu n}^{\mu m}(\mathbf{R}_{12})\, \exp(\mathrm{i}\boldsymbol{k}_T \cdot \boldsymbol{q}_{21})\, \Phi_\nu^\mu(z_2)\, \mathrm{d}V_2 = -\mathrm{e}^{\mathrm{i}\alpha z_1} \overline{Y_n^m(\hat{\boldsymbol{k}})} \tag{4.16}$$

for $n = 0, 1, 2, \ldots$ and $m = -n, \ldots, n$, where $\boldsymbol{q}_{sj} = \boldsymbol{q}_s - \boldsymbol{q}_j$.

Proceeding as before, suppose that for sufficiently large z (say, $z > \ell$) we can write

$$\Phi_n^m(z) = F_n^m \mathrm{e}^{\mathrm{i}\lambda z}. \tag{4.17}$$

Then if $z_1 > \ell + b$, (4.16) becomes

(4.18)

$$F_n^m \mathrm{e}^{\mathrm{i}\lambda z_1} + n_0 \sum_{\nu,\mu} (-\mathrm{i})^{\nu-n} Z_\nu \left\{ \int_0^\ell \Phi_\nu^\mu(z_2) \mathcal{L}_{\nu n}^{\mu m}(z_{21})\, \mathrm{d}z_2 + F_\nu^\mu \mathrm{e}^{\mathrm{i}\lambda z_1} \mathcal{M}_{\nu n}^{\mu m} \right\} = -\mathrm{e}^{\mathrm{i}\alpha z_1} \overline{Y_n^m(\hat{\boldsymbol{k}})}$$

for $n = 0, 1, 2, \ldots$ and $m = -n, \ldots, n$, where $z_{21} = z_2 - z_1$,

$$\mathcal{L}_{\nu n}^{\mu m}(z_{21}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S_{\nu n}^{\mu m}(\mathbf{R}_{21}) \, \exp\left(\mathrm{i}\boldsymbol{k}_T \cdot \boldsymbol{q}_{21}\right) \mathrm{d}x_2 \, \mathrm{d}y_2,$$

$$\mathcal{M}_{\nu n}^{\mu m} = \int_{z_2 > \ell,\, R_{21} > b} S_{\nu n}^{\mu m}(\mathbf{R}_{21}) \, \exp\left(\mathrm{i}\boldsymbol{k}_T \cdot \boldsymbol{q}_{21}\right) \mathrm{e}^{\mathrm{i}\lambda z_{21}} \, \mathrm{d}V_2,$$

and we have used $S_{\nu n}^{\mu m}(-\boldsymbol{r}) = (-1)^{n+\nu} S_{\nu n}^{\mu m}(\boldsymbol{r})$, a relation that follows from (4.9). Indeed, because of (4.9), it is sufficient to consider

$$L_n^m(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \psi_n^m(\mathbf{R}) \, \exp\left(\mathrm{i}\boldsymbol{k}_T \cdot \mathbf{Q}\right) \mathrm{d}X \, \mathrm{d}Y$$

for $z < 0$ and

$$M_n^m = \int_{z_2 > \ell,\, R_{21} > b} \psi_n^m(\mathbf{R}_{21}) \, \Psi(\mathbf{R}_{21}) \, \mathrm{d}V_2,$$

where $\Psi(\mathbf{R}) = \mathrm{e}^{\mathrm{i}\lambda z} \exp\left(\mathrm{i}\boldsymbol{k}_T \cdot \mathbf{Q}\right) = \exp\left(\mathrm{i}\mathbf{K} \cdot \mathbf{R}\right)$ and $\mathbf{K} = K\hat{\mathbf{K}}$ is defined by (3.15).

From (B.5), we have

$$L_n^m(z_{21}) = \frac{2\pi \mathrm{i}^n}{k\alpha} Y_n^m(\hat{\boldsymbol{k}}) \, \mathrm{e}^{\mathrm{i}\alpha(z_1 - z_2)} \quad \text{for } z_1 > z_2.$$

Hence, $\mathcal{L}_{\nu n}^{\mu m}$ is proportional to $\mathrm{e}^{\mathrm{i}\alpha(z_1 - z_2)}$, and so the integral term in (4.18) is proportional to $\mathrm{e}^{\mathrm{i}\alpha z_1}$.

The volume integral M_n^m can be evaluated readily using Green's theorem. We have $\psi_n^m \nabla^2 \Psi - \Psi \nabla^2 \psi_n^m = (k^2 - K^2) \psi_n^m \Psi$. It follows that

$$M_n^m = \frac{1}{k^2 - K^2} \int_{\partial B} \left[\psi_n^m \frac{\partial \Psi}{\partial n} - \Psi \frac{\partial \psi_n^m}{\partial n} \right] \mathrm{d}S_2,$$

where ∂B consists of two parts, the plane $z_2 = \ell$ and the sphere $R_{12} = b$. Now, on $z_2 = \ell$, $\partial/\partial n = -\partial/\partial z_2$, and so we have

$$-\int_{z_2 = \ell} \left[\psi_n^m \frac{\partial \Psi}{\partial z_2} - \Psi \frac{\partial \psi_n^m}{\partial z_2} \right] \mathrm{d}x_2 \, \mathrm{d}y_2 = \frac{2\pi}{k\alpha} \mathrm{e}^{\mathrm{i}(\alpha - \lambda)(z_1 - \ell)} \, \mathrm{i}^{n-1} (\lambda + \alpha) \, Y_n^m(\hat{\boldsymbol{k}}),$$

using (B.8). Thus, the plane part of ∂B contributes a term to $\mathcal{M}_{\nu n}^{\mu m}$ proportional to $\mathrm{e}^{\mathrm{i}(\alpha - \lambda) z_1}$, which in turn gives a contribution to (4.18) proportional to $\mathrm{e}^{\mathrm{i}\alpha z_1}$.

Next, from (4.4), we have

$$\Psi = \exp\left(\mathrm{i}\mathbf{K} \cdot \mathbf{R}\right) = 4\pi \sum_{\nu,\mu} \mathrm{i}^\nu j_\nu(KR) \, \overline{Y_\nu^\mu(\hat{\mathbf{R}})} \, Y_\nu^\mu(\hat{\mathbf{K}}).$$

Then, the contribution from the sphere $R_{12} = b$ is

$$\begin{aligned} &-\int_\Omega \left[\psi_n^m \frac{\partial \Psi}{\partial R} - \Psi \frac{\partial \psi_n^m}{\partial R} \right]_{R=b} b^2 \, \mathrm{d}\Omega \\ &\quad = 4\pi b^2 \sum_{\nu,\mu} \mathrm{i}^\nu Y_\nu^\mu(\hat{\mathbf{K}}) \{ k j_\nu(Kb) \, h_n'(kb) - K j_\nu'(Kb) \, h_n(kb) \} \int_\Omega Y_n^m \overline{Y_\nu^\mu} \, \mathrm{d}\Omega \\ &\quad = 4\pi b^2 \mathrm{i}^n Y_n^m(\hat{\mathbf{K}}) \{ k j_n(Kb) \, h_n'(kb) - K j_n'(Kb) \, h_n(kb) \}, \end{aligned}$$

which is independent of z_1; here, Ω is the unit sphere and we have used (A.1).

Collecting up our results, we find that (4.18) can be written as

$$\mathcal{A}_n^m \,\mathrm{e}^{\mathrm{i}\lambda z_1} + \mathcal{B}_n^m \,\mathrm{e}^{\mathrm{i}\alpha z_1} = -\mathrm{e}^{\mathrm{i}\alpha z_1}\,\overline{Y_n^m(\hat{\boldsymbol{k}})}, \tag{4.19}$$

for $n = 0, 1, 2, \ldots, m = -n, \ldots, n$, and $z_1 > \ell + b$, where

$$\mathcal{A}_n^m = F_n^m + \frac{(4\pi)^2 \mathrm{i} n_0 (-1)^m}{k(k^2 - K^2)} \sum_{\nu,\mu} Z_\nu F_\nu^\mu \sum_q Y_q^{\mu-m}(\hat{\mathbf{K}})\, \mathcal{N}_q(Kb)\, \mathcal{G}(n, m; \nu, -\mu; q), \tag{4.20}$$

$$\mathcal{N}_n(x) = \mathrm{i}kb\{x\, j_n'(x)\, h_n(kb) - kb\, j_n(x)\, h_n'(kb)\}, \tag{4.21}$$

and we have used (4.10) to remove a factor of $(-1)^{q+n+\nu}$. In particular, we note that $\mathcal{N}_0$ appeared in section 3.3 during our analysis of Lax's integral equation.

From (4.19), we immediately obtain $\mathcal{A}_n^m = 0$ for $n = 0, 1, 2, \ldots$, $m = -n, \ldots, n$. These equations yield an infinite homogeneous system of linear algebraic equations for F_n^m. The existence of a nontrivial solution to this system determines K.

It is worth noting that even though the solution of the system $\mathcal{A}_n^m = 0$ can depend on θ_{in} via $\hat{\mathbf{K}}$ (see (3.15)), the effective wavenumber itself, K, should not depend on θ_{in}.

4.4. Approximate determination of K for small n_0. The only approximation made in the derivation of the system $\mathcal{A}_n^m = 0$ is the QCA, which is expected to be valid for small values of the scatterer concentration ($n_0 a^3 \ll 1$). We now assume (as in section 3.3) that $n_0 b/k^2$ is also small and write $K^2 = k^2 + \delta_1 n_0 + \delta_2 n_0^2 + \cdots$. Then

$$\mathcal{N}_n(Kb) = 1 - \frac{\mathrm{i}bn_0}{2k}\delta_1 d_n(kb) + O(n_0^2), \tag{4.22}$$

where

$$d_n(x) = x j_n'(x)[x h_n'(x) + h_n(x)] + [x^2 - n(n+1)] j_n(x)\, h_n(x), \tag{4.23}$$

and so

$$\frac{\mathcal{N}_n(Kb)}{k^2 - K^2} = -\frac{1}{\delta_1 n_0} + \frac{\mathrm{i}b d_n(kb)}{2k} + \frac{\delta_2}{\delta_1^2} + O(n_0). \tag{4.24}$$

If (4.24) is substituted into $\mathcal{A}_n^m = 0$, with $\mathcal{A}_n^m$ defined by (4.20) and $O(n_0^2)$ terms neglected, we obtain

$$F_n^m = \frac{(4\pi)^2 \mathrm{i}}{k\delta_1}(-1)^m \left(1 - \frac{n_0\delta_2}{\delta_1}\right) \sum_{\nu,\mu} Z_\nu F_\nu^\mu W_{n\nu}^{m\mu} + \frac{(4\pi)^2 b n_0}{2k^2}(-1)^m \sum_{\nu,\mu} Z_\nu F_\nu^\mu X_{n\nu}^{m\mu}, \tag{4.25}$$

where

$$W_{n\nu}^{m\mu} = \sum_q Y_q^{\mu-m}(\hat{\mathbf{K}})\, \mathcal{G}(n, m; \nu, -\mu; q), \tag{4.26}$$

$$X_{n\nu}^{m\mu} = \sum_q Y_q^{\mu-m}(\hat{\mathbf{K}})\, \mathcal{G}(n, m; \nu, -\mu; q)\, d_q(kb). \tag{4.27}$$

The Gaunt coefficients appear in the linearization formula for spherical harmonics, (A.4). Replacing μ by $-\mu$ and $\hat{\boldsymbol{r}}$ by $\hat{\mathbf{K}}$ in the complex conjugate of (A.4), we obtain

$$W_{n\nu}^{m\mu} = Y_n^{-m}(\hat{\mathbf{K}})\, Y_\nu^\mu(\hat{\mathbf{K}}).$$

Thus, at leading order, (4.25) gives

$$F_n^m = \frac{(4\pi)^2\mathrm{i}}{k\delta_1}\overline{Y_n^m(\hat{\mathbf{K}})}\sum_{\nu,\mu} Z_\nu F_\nu^\mu Y_\nu^\mu(\hat{\mathbf{K}})$$

for $n = 0, 1, 2, \ldots$ and $m = -n, \ldots, n$. Set $F_n^m = \overline{Y_n^m(\hat{\mathbf{K}})}\,\widetilde{F}_n^m$, whence

$$\widetilde{F}_n^m = \frac{(4\pi)^2\mathrm{i}}{k\delta_1}\sum_{\nu,\mu} Z_\nu \widetilde{F}_\nu^\mu Y_\nu^\mu(\hat{\mathbf{K}})\,\overline{Y_\nu^\mu(\hat{\mathbf{K}})}$$

for $n = 0, 1, 2, \ldots$ and $m = -n, \ldots, n$. However, the right-hand side of this equation does not depend on n or m, so that $\widetilde{F}_n^m = \widetilde{F}$, say. Hence

$$\delta_1 = \frac{(4\pi)^2\mathrm{i}}{k}\sum_{\nu=0}^{\infty} Z_\nu \sum_{\mu=-\nu}^{\nu} Y_\nu^\mu(\hat{\mathbf{K}})\,\overline{Y_\nu^\mu(\hat{\mathbf{K}})}. \tag{4.28}$$

The sum over μ can be evaluated using Legendre's addition theorem, (A.3). Setting $\hat{\boldsymbol{r}}_1 = \hat{\boldsymbol{r}}_2 = \hat{\mathbf{K}}$ in (A.3) and noting that $P_n(1) = 1$, we obtain

$$\delta_1 = \frac{4\pi\mathrm{i}}{k}\sum_{\nu=0}^{\infty}(2\nu+1)Z_\nu = -\frac{4\pi\mathrm{i}}{k}f(0), \tag{4.29}$$

where f is the far-field pattern given by (4.6).

Returning to (4.25), we now set

$$F_n^m = \overline{Y_n^m(\hat{\mathbf{K}})}\,\widetilde{F} + n_0 G_n^m,$$

and then the $O(n_0)$ terms give

$$G_n^m = \overline{Y_n^m(\hat{\mathbf{K}})}\,V + \frac{(4\pi)^2 b}{2k^2}(-1)^m\widetilde{F}\sum_{\nu,\mu} Z_\nu\,\overline{Y_\nu^\mu(\hat{\mathbf{K}})}\,X_{n\nu}^{m\mu}, \tag{4.30}$$

where

$$V = \frac{(4\pi)^2\mathrm{i}}{k\delta_1}\sum_{\nu,\mu} Z_\nu G_\nu^\mu Y_\nu^\mu(\hat{\mathbf{K}}) - \frac{\delta_2}{\delta_1}\widetilde{F}. \tag{4.31}$$

Note that V does not depend on n or m. Substituting for G_ν^μ from (4.30) in (4.31), making use of (4.28), gives a formula for δ_2:

$$\delta_2 = \frac{(4\pi)^4\mathrm{i}b}{2k^3}\sum_{n,m}\sum_{\nu,\mu}(-1)^m Z_n Z_\nu Y_n^m(\hat{\mathbf{K}})\,\overline{Y_\nu^\mu(\hat{\mathbf{K}})}\,X_{n\nu}^{m\mu}. \tag{4.32}$$

So far we have not made any assumptions about the size of ka or kb (though clearly $kb \geq 2ka$). Now we will assume that kb is small. In the limit $x \to 0$, we have $d_n(x) \sim \mathrm{i}n/x$. Using this approximation simplifies $X_{n\nu}^{m\mu}$, defined by (4.27). Hence,

$$\delta_2 \sim -\frac{1}{2}(4\pi/k)^4\sum_{n=0}^{\infty}\sum_{\nu=0}^{\infty} Z_n Z_\nu\, K_{n\nu}(\hat{\mathbf{K}}) \quad \text{as } kb \to 0, \tag{4.33}$$

where

$$K_{n\nu}(\hat{\mathbf{K}}) = \sum_{m=-n}^{n} \sum_{\mu=-\nu}^{\nu} (-1)^m Y_n^m(\hat{\mathbf{K}})\, \overline{Y_\nu^\mu(\hat{\mathbf{K}})} \sum_q q\, Y_q^{\mu-m}(\hat{\mathbf{K}})\, \mathcal{G}(n,m;\nu,-\mu;q). \tag{4.34}$$

From (A.5) and $\int_0^{2\pi} \mathrm{e}^{\mathrm{i}m\phi}\,\mathrm{d}\phi = 2\pi\delta_{0m}$, we have

$$\begin{aligned} Y_q^{\mu-m}(\hat{\mathbf{K}})\, \mathcal{G}(n,m;\nu,-\mu;q) &= (-1)^m \sum_M Y_q^M(\hat{\mathbf{K}}) \int_\Omega \overline{Y_n^m} Y_\nu^\mu \overline{Y_q^M}\,\mathrm{d}\Omega \\ &= (-1)^m \frac{2q+1}{4\pi} \int_\Omega \overline{Y_n^m}(\hat{\boldsymbol{r}})\, Y_\nu^\mu(\hat{\boldsymbol{r}})\, P_q(\hat{\boldsymbol{r}}\cdot\hat{\mathbf{K}})\,\mathrm{d}\Omega(\hat{\boldsymbol{r}}), \end{aligned}$$

using (A.3). Hence, using (A.3) two more times, we obtain

$$\begin{aligned} K_{n\nu}(\hat{\mathbf{K}}) &= \frac{(2n+1)(2\nu+1)}{(4\pi)^3} \sum_q q(2q+1) \int_\Omega P_n(\hat{\boldsymbol{r}}\cdot\hat{\mathbf{K}})\, P_\nu(\hat{\boldsymbol{r}}\cdot\hat{\mathbf{K}})\, P_q(\hat{\boldsymbol{r}}\cdot\hat{\mathbf{K}})\,\mathrm{d}\Omega(\hat{\boldsymbol{r}}) \\ &= \frac{\sqrt{(2n+1)(2\nu+1)}}{(4\pi)^{3/2}} \sum_q q\sqrt{2q+1}\, \mathcal{G}(n,0;\nu,0;q), \end{aligned}$$

where we have used $Y_n^0 = \sqrt{(2n+1)/(4\pi)}\, P_n$ and (A.5). When this formula for $K_{n\nu}$ is substituted into (4.33), we obtain complete agreement with the formula of Lloyd and Berry [21]; see Appendix C.

In conclusion, we note that if we were to replace (2.5) with the (clearly unreasonable)

$$p(\boldsymbol{r}_2|\boldsymbol{r}_1) = (n_0/N)H(|z_2 - z_1| - a),$$

an analysis similar to that given above yields Twersky's erroneous expression for δ_2, as given in (1.4). We omit the details of this calculation, but see [21] for a related discussion and [20] for analogous calculations in two dimensions.

It is perhaps worth summarizing the various approximations that are needed to arrive at the Lloyd–Berry formula. The system $\mathcal{A}_n^m = 0$ (with $\mathcal{A}_n^m$ defined by (4.20)) serves to determine the effective wavenumber, subject only to the QCA. The QCA is certainly appropriate only for low volume fractions, but it is very difficult to make a precise quantitative assessment of its range of validity. Some numerical estimates of the accuracy of the QCA can be found in, for example, [16] and [9]. If we assume that the concentration of scatterers is small, in the sense that $n_0 a^3 \ll 1$ (which is consistent with the QCA) and also that $n_0 b/k^2 \ll 1$, then the effective wavenumber, up to second order in concentration, follows from (4.29) and (4.32). If we finally let $kb \to 0$, we obtain the Lloyd–Berry formula in the form of (4.29) and (4.33).

Appendix A. Spherical harmonics. We define spherical harmonics Y_n^m by

$$Y_n^m(\hat{\boldsymbol{r}}) = Y_n^m(\theta,\phi) = (-1)^m \sqrt{\frac{2n+1}{4\pi}} \sqrt{\frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta)\, \mathrm{e}^{\mathrm{i}m\phi},$$

where P_n^m is an associated Legendre function. We have orthonormality,

$$\int_\Omega Y_n^m\, \overline{Y_\nu^\mu}\,\mathrm{d}\Omega = \delta_{n\nu}\delta_{m\mu}, \tag{A.1}$$

where Ω is the unit sphere. Also, $Y_n^{-m} = (-1)^m \overline{Y_n^m}$.

For $0 \le m \le n$, we have the expansion

$$\frac{P_n^m(t)}{(1-t^2)^{m/2}} = \frac{1}{2^n\, n!}\frac{\mathrm{d}^{m+n}}{\mathrm{d}t^{m+n}}(t^2-1)^n = \sum_{l=0}^{[(n-m)/2]} B_l^{n,m} t^{n-m-2l}, \tag{A.2}$$

where $[n]$ denotes the integer part of n. The coefficients $B_l^{n,m}$ are known explicitly, but we shall not need them.

We shall make use of Legendre's addition theorem, namely,

$$P_n(\hat{\boldsymbol{r}}_1 \cdot \hat{\boldsymbol{r}}_2) = \frac{4\pi}{2n+1} \sum_{m=-n}^{n} Y_n^m(\hat{\boldsymbol{r}}_1)\, \overline{Y_n^m(\hat{\boldsymbol{r}}_2)}, \tag{A.3}$$

where $P_n(t)$ is a Legendre polynomial.

The linearization formula for spherical harmonics is

$$Y_n^m(\hat{\boldsymbol{r}})\, Y_\nu^\mu(\hat{\boldsymbol{r}}) = \sum_q Y_q^{m+\mu}(\hat{\boldsymbol{r}})\, \mathcal{G}(n,m;\nu,\mu;q), \tag{A.4}$$

where $\mathcal{G}$ is a Gaunt coefficient. Note that $\mathcal{G}$ is real. Making use of (A.1), we obtain

$$\mathcal{G}(n,m;\nu,-\mu;q) = (-1)^m \int_\Omega \overline{Y_n^m} Y_\nu^\mu \overline{Y_q^{\mu-m}}\, \mathrm{d}\Omega. \tag{A.5}$$

Appendix B. Some integrals. Consider the integral

$$L(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_0(kR)\, \exp\left(\mathrm{i}\boldsymbol{k}_T \cdot \mathbf{Q}\right) \mathrm{d}X\, \mathrm{d}Y,$$

where $R = |\mathbf{R}|$, $\mathbf{R} = (X, Y, z)$, and $\mathbf{Q} = (X, Y, 0)$. Set $\mathbf{Q} = Q(\cos\Phi,\ \sin\Phi,\ 0)$ so that $\boldsymbol{k}_T \cdot \mathbf{Q} = kQ\sin\theta_{\mathrm{in}}\, \cos(\Phi - \phi_{\mathrm{in}})$. Hence, as $\mathrm{d}X\, \mathrm{d}Y = Q\, \mathrm{d}Q\, \mathrm{d}\Phi$, we can integrate over Φ, giving

$$L(z) = 2\pi \int_0^\infty h_0(k\sqrt{Q^2+z^2})\, J_0(kQ\sin\theta_{\mathrm{in}})\, Q\, \mathrm{d}Q.$$

We have

$$Qh_0(k\sqrt{Q^2+z^2}) = \frac{Q\, \mathrm{e}^{\mathrm{i}k\sqrt{Q^2+z^2}}}{\mathrm{i}k\sqrt{Q^2+z^2}} = \frac{1}{(\mathrm{i}k)^2}\frac{\mathrm{d}}{\mathrm{d}Q}\mathrm{e}^{\mathrm{i}k\sqrt{Q^2+z^2}},$$

so that an integration by parts (using $J_0' = -J_1$) gives

$$L(z) = 2\pi k^{-2}\{\mathrm{e}^{\mathrm{i}k|z|} - \hat{L}(z)\}, \tag{B.1}$$

where

$$\hat{L}(z) = k\sin\theta_{\mathrm{in}} \int_0^\infty J_1(kQ\sin\theta_{\mathrm{in}})\, \mathrm{e}^{\mathrm{i}k\sqrt{Q^2+z^2}}\, \mathrm{d}Q.$$

Now, from [8, equation 6.637(1)] (with $\nu = 1$ therein), we have

$$\int_0^\infty \frac{\mathrm{e}^{-a\sqrt{x^2+\beta^2}}}{\sqrt{x^2+\beta^2}}\, J_1(\gamma x)\, \mathrm{d}x = I_{1/2}(X_-)\, K_{1/2}(X_+), \tag{B.2}$$

where $X_\pm = \frac{1}{2}\beta\{\sqrt{a^2+\gamma^2} \pm a\}$, $\operatorname{Re} a > 0$, $\operatorname{Re}\beta > 0$, and $\operatorname{Re}\gamma > 0$. From [1, 10.2.13 and 10.2.17], the modified Bessel functions are given by

$$I_{1/2}(w) = \{2/(\pi w)\}^{1/2} \sinh w \quad \text{and} \quad K_{1/2}(w) = \{\pi/(2w)\}^{1/2} \mathrm{e}^{-w},$$

so that

$$I_{1/2}(X_-)\, K_{1/2}(X_+) = (\beta\gamma)^{-1} \left\{ \mathrm{e}^{-a\beta} - \mathrm{e}^{-\beta\sqrt{a^2+\gamma^2}} \right\}.$$

Then, differentiating (B.2) with respect to a gives

$$\gamma \int_0^\infty J_1(\gamma x)\, \mathrm{e}^{-a\sqrt{x^2+\beta^2}}\, \mathrm{d}x = \mathrm{e}^{-a\beta} - \frac{a}{\sqrt{a^2+\gamma^2}} \mathrm{e}^{-\beta\sqrt{a^2+\gamma^2}}. \tag{B.3}$$

The calculations leading to (B.3) are certainly valid for $\operatorname{Re} a > 0$, $\operatorname{Re}\beta > 0$, and $\operatorname{Re}\gamma > 0$. We want to use (B.3) for $\beta = z$; as the left-hand side of (B.3) is an even function of β, we can replace β by $|\beta|$ on the right-hand side. We also want to substitute $a = -\mathrm{i}k$ and $\gamma = k\sin\theta_{\mathrm{in}}$, so that $\sqrt{a^2+\gamma^2} = \pm \mathrm{i}k\cos\theta_{\mathrm{in}}$. To determine the sign, we note that (from [8, equations 6.671(1) and 6.671(2)])

$$\gamma \int_0^\infty J_1(\gamma x)\, \mathrm{e}^{\mathrm{i}kx}\, \mathrm{d}x = 1 - \frac{k}{\sqrt{k^2-\gamma^2}} \quad \text{for } k > \gamma,$$

implying that we should take $\sqrt{a^2+\gamma^2} = -\mathrm{i}k\cos\theta_{\mathrm{in}}$. (Alternatively, we note that the right-hand side of (B.3) is an analytic function of a in a cut plane; we can take the cut between $a = \mathrm{i}\gamma$ and $a = -\mathrm{i}\gamma$ (γ real and positive), and we choose the branch so that the right-hand side of (B.3) is real when a is real and positive. This leads to $\sqrt{a^2+\gamma^2} = -\mathrm{i}\sqrt{k^2-\gamma^2}$ when $a = -\mathrm{i}k$ with $k > \gamma > 0$.) Hence,

$$\hat{L}(z) = \mathrm{e}^{\mathrm{i}k|z|} - \mathrm{e}^{\mathrm{i}k|z|\cos\theta_{\mathrm{in}}} \sec\theta_{\mathrm{in}},$$

and so (B.1) gives

$$L(z) = \int_{-\infty}^\infty \int_{-\infty}^\infty h_0(kR)\, \exp\left(\mathrm{i}\boldsymbol{k}_T \cdot \mathbf{Q}\right) \mathrm{d}X\, \mathrm{d}Y = \frac{2\pi}{k^2\cos\theta_{\mathrm{in}}}\, \mathrm{e}^{\mathrm{i}k|z|\cos\theta_{\mathrm{in}}}. \tag{B.4}$$

This formula generalizes. Thus, let

$$L_n^m(z) = \int_{-\infty}^\infty \int_{-\infty}^\infty \psi_n^m(\mathbf{R})\, \exp\left(\mathrm{i}\boldsymbol{k}_T \cdot \mathbf{Q}\right) \mathrm{d}X\, \mathrm{d}Y,$$

with $\psi_n^m(\boldsymbol{r}) = h_n(kr) Y_n^m(\hat{\boldsymbol{r}})$. Then,

$$L_n^m(z) = \frac{2\pi\,\mathrm{i}^n}{k^2\cos\theta_{\mathrm{in}}}\, Y_n^m(\hat{\boldsymbol{k}})\, \mathrm{e}^{-\mathrm{i}kz\cos\theta_{\mathrm{in}}} \quad \text{for } z < 0, \tag{B.5}$$

with a similar formula for $z > 0$ (which we shall not need). When both $m = 0$ and $\theta_{\mathrm{in}} = 0$, (B.5) reduces to a result obtained in [27].

To prove (B.5), begin by assuming that $0 \le m \le n$. Let

$$\Omega_n^m(\boldsymbol{r}) = h_n(kr)\, P_n^m(\cos\theta)\, \mathrm{e}^{\mathrm{i}m\phi}. \tag{B.6}$$

($\psi_n^m(\boldsymbol{r})$ is a normalized form of $\Omega_n^m(\boldsymbol{r})$.) Then, we have [5, 29, 4]

$$\Omega_n^m(\boldsymbol{r}) = \mathcal{Y}_n^m h_0(kr),$$

where the Erdélyi operator $\mathcal{Y}_n^m$ is defined by

$$\mathcal{Y}_n^m = (\mathcal{D}_{xy})^m \sum_{l=0}^{[(n-m)/2]} (-1)^l B_l^{n,m} (\mathcal{D}_z)^{n-m-2l}, \qquad \mathcal{D}_{xy} = -\frac{1}{k}\left(\frac{\partial}{\partial x} + \mathrm{i}\frac{\partial}{\partial y}\right),$$

and $\mathcal{D}_z = -k^{-1}\partial/\partial z$; the coefficients $B_l^{n,m}$ appear in the expansion (A.2). Hence,

$$O_n^m(z) \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Omega_n^m(\mathbf{R}) \exp(\mathrm{i}\boldsymbol{k}_T \cdot \mathbf{Q})\, \mathrm{d}X\, \mathrm{d}Y = \sum_l (-1)^l B_l^{n,m} (\mathcal{D}_z)^{n-m-2l} \mathcal{I}_m(z),$$

where the sum is from $l = 0$ to the integer part of $(n-m)/2$, and

$$\begin{aligned}\mathcal{I}_m(z) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(\mathrm{i}\boldsymbol{k}_T \cdot \mathbf{Q})\, (\mathcal{D}_{XY})^m h_0(kR)\, \mathrm{d}X\, \mathrm{d}Y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_0(kR)\, (-\mathcal{D}_{XY})^m \exp(\mathrm{i}\boldsymbol{k}_T \cdot \mathbf{Q})\, \mathrm{d}X\, \mathrm{d}Y = \mathrm{i}^m \sin^m\theta_{\mathrm{in}}\, \mathrm{e}^{\mathrm{i}m\phi_{\mathrm{in}}} L(z).\end{aligned}$$

Hence, substituting for $L(z)$ from (B.4) and carrying out the differentiations with respect to z, we obtain

$$O_n^m(z) = \frac{2\pi\,\mathrm{i}^n}{k^2\cos\theta_{\mathrm{in}}}\, P_n^m(\cos\theta_{\mathrm{in}})\, \mathrm{e}^{\mathrm{i}m\phi_{\mathrm{in}}}\, \mathrm{e}^{-\mathrm{i}kz\cos\theta_{\mathrm{in}}} \tag{B.7}$$

for $z < 0$. The result (B.5) follows after multiplication by the appropriate normalization constant. It can be shown that the same result is also true for $-n \le m \le 0$.

Next, we consider an integral required in section 4.3. We have

$$\begin{aligned}-\int_{z_2=\ell} &\left[\Omega_n^m \frac{\partial\Psi}{\partial z_2} - \Psi\frac{\partial\Omega_n^m}{\partial z_2}\right] \mathrm{d}x_2\, \mathrm{d}y_2 \\ &= \mathrm{e}^{\mathrm{i}\lambda(\ell - z_1)} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(\mathrm{i}\boldsymbol{k}_T \cdot \boldsymbol{q}_{21}) \left[-\mathrm{i}\lambda\Omega_n^m + \frac{\partial\Omega_n^m}{\partial z_2}\right]_{z_2=\ell} \mathrm{d}x_2\, \mathrm{d}y_2,\end{aligned}$$

where $\Psi = \exp(\mathrm{i}\mathbf{K}\cdot\mathbf{R}_{21})$. Using

$$\mathcal{D}_z\Omega_n^m = (2n+1)^{-1}\{(n-m+1)\Omega_{n+1}^m - (n+m)\Omega_{n-1}^m\}$$

and (B.7) thrice gives the integral's value as

$$\frac{2\pi}{k\alpha}\mathrm{e}^{\mathrm{i}(\alpha-\lambda)(z_1-\ell)}\, \mathrm{i}^{n-1}(\lambda+\alpha)\, P_n^m(\cos\theta_{\mathrm{in}})\, \mathrm{e}^{\mathrm{i}m\phi_{\mathrm{in}}}, \tag{B.8}$$

where we have also used $(2n+1)tP_n^m(t) = (n-m+1)P_{n+1}^m(t) + (n+m)P_{n-1}^m(t)$.

Appendix C. The Lloyd–Berry formula. Recall the formula (1.6). From (4.6) and $Y_n^0(\hat{\boldsymbol{r}}) = \sqrt{(2n+1)/(4\pi)}P_n(\cos\theta)$, we obtain

$$f(\theta) = -\sqrt{4\pi}\sum_{n=0}^{\infty}\sqrt{2n+1}\, Z_n Y_n^0.$$

Then, the linearization formula (A.4) gives

$$[f(\theta)]^2 = \sum_{n=0}^{\infty}\sum_{\nu=0}^{\infty}\sum_q T(n,\nu;q)\, P_q(\cos\theta), \tag{C.1}$$

where

$$T(n,\nu;q) = \sqrt{4\pi(2n+1)(2\nu+1)(2q+1)}\, Z_n Z_\nu\, \mathcal{G}(n,0;\nu,0;q).$$

Hence,

$$-[f(\pi)]^2 + [f(0)]^2 = \sum_{n=0}^{\infty}\sum_{\nu=0}^{\infty}\sum_{q} T(n,\nu;q)\,\{1-(-1)^q\}. \tag{C.2}$$

For the integral term in (1.6), we use (C.1) and

$$\int_0^\pi \frac{1}{\sin(\theta/2)}\frac{\mathrm{d}}{\mathrm{d}\theta}P_q(\cos\theta)\,\mathrm{d}\theta = -\int_{-1}^{1}\sqrt{\frac{2}{1-x}}\,P_q'(x)\,\mathrm{d}x = (-1)^q - 1 - 2q. \tag{C.3}$$

(The last equality was obtained as follows. From [8, equation 7.225(1)], we have

$$\begin{aligned}\frac{2}{2n+1}\frac{1}{\sqrt{1+x}}\{T_n(x)+T_{n+1}(x)\} &= \int_{-1}^{x}\frac{1}{\sqrt{x-t}}\,P_n(t)\,\mathrm{d}t \\ &= 2(-1)^n\sqrt{1+x} + 2\int_{-1}^{x}\sqrt{x-t}\,P_n'(t)\,\mathrm{d}t,\end{aligned}$$

after an integration by parts, where $T_n(\cos\theta) = \cos n\theta$ is a Chebyshev polynomial, and we have used $P_n(-1) = (-1)^n$. Now, differentiate this formula with respect to x and then let $x \to 1$, using $T_n(1) = 1$ and $T_n'(1) = n^2$.) Substituting (C.1), (C.2), and (C.3) into (1.6) gives

$$\delta_2 = -\frac{8\pi^2}{k^4}\sum_{n=0}^{\infty}\sum_{\nu=0}^{\infty}\sum_{q} q\,T(n,\nu;q), \tag{C.4}$$

which is the same as (4.33).

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] D. G. Aggelis, D. Polyzos, and T. P. Philippidis, *Wave dispersion and attenuation in fresh mortar: Theoretical predictions vs. experimental results*, J. Mech. Phys. Solids, 53 (2005), pp. 857–883.

[3] M. A. Epton and B. Dembart, *Multipole translation theory for the three-dimensional Laplace and Helmholtz equations*, SIAM J. Sci. Comput., 16 (1995), pp. 865–897.

[4] H. J. H. Clercx and P. P. J. M. Schram, *An alternative expression for the addition theorems of spherical wave solutions of the Helmholtz equation*, J. Math. Phys., 34 (1993), pp. 5292–5302.

[5] A. Erdélyi, *Zur Theorie der Kugelwellen*, Physica, 4 (1937), pp. 107–120.

[6] J. G. Fikioris and P. C. Waterman, *Multiple scattering of waves.* II. *"Hole corrections" in the scalar case*, J. Math. Phys., 5 (1964), pp. 1413–1420.

[7] L. L. Foldy, *The multiple scattering of waves.* I. *General theory of isotropic scattering by randomly distributed scatterers*, Phys. Rev., 67 (1945), pp. 107–119.

[8] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series, and Products*, 4th ed., Academic Press, New York, 1980.

[9] C.-A. Guérin, P. Mallet, and A. Sentenac, *Effective-medium theory for finite-size aggregates*, J. Opt. Soc. Amer. A, 23 (2006), pp. 349–358.

[10] N. A. Gumerov and R. Duraiswami, *Computation of scattering from N spheres using multipole reexpansion*, J. Acoust. Soc. Amer., 112 (2002), pp. 2688–2701.

[11] N. A. Gumerov and R. Duraiswami, *Recursions for the computation of multipole translation and rotation coefficients for the* 3-D *Helmholtz equation*, SIAM J. Sci. Comput., 25 (2003), pp. 1344–1381.
[12] N. A. Gumerov and R. Duraiswami, *Computation of scattering from clusters of spheres using the fast multipole method*, J. Acoust. Soc. Amer., 117 (2005), pp. 1744–1761.
[13] A. Ishimaru, *Wave Propagation and Scattering in Random Media*, Vol. 2, Academic Press, New York, 1978.
[14] C. Javanaud and A. Thomas, *Multiple scattering using the Foldy–Twersky integral equation*, Ultrasonics, 26 (1988), pp. 341–343.
[15] F. H. Kerr, *The scattering of a plane elastic wave by spherical elastic inclusions*, Int. J. Engrg. Sci., 30 (1992), pp. 169–186.
[16] J.-Y. Kim, *Dynamic self-consistent analysis for elastic wave propagation in fiber reinforced composites*, J. Acoust. Soc. Amer., 100 (1996), pp. 2002–2010.
[17] S. Koc and W. C. Chew, *Calculation of acoustical scattering from a cluster of scatterers*, J. Acoust. Soc. Amer., 103 (1998), pp. 721–734.
[18] M. Lax, *Multiple scattering of waves*, Rev. Modern Phys., 23 (1951), pp. 287–310.
[19] M. Lax, *Multiple scattering of waves.* II. *The effective field in dense systems*, Phys. Rev., 85 (1952), pp. 621–629.
[20] C. M. Linton and P. A. Martin, *Multiple scattering by random configurations of circular cylinders: Second-order corrections for the effective wavenumber*, J. Acoust. Soc. Amer., 117 (2005), pp. 3413–3423.
[21] P. Lloyd and M. V. Berry, *Wave propagation through an assembly of spheres* IV. *Relations between different multiple scattering theories*, Proc. Phys. Soc., 91 (1967), pp. 678–688.
[22] P. A. Martin, *Multiple Scattering*, Cambridge University Press, Cambridge, UK, 2006.
[23] V. J. Pinfield, O. G. Harlen, M. J. W. Povey, and B. D. Sleeman, *Acoustic propagation in dispersions in the long wavelength limit*, SIAM J. Appl. Math., 66 (2005), 489–509.
[24] M. J. W. Povey, *Ultrasonic Techniques for Fluids Characterization*, Academic Press, San Diego, 1997.
[25] L. Tsang, J. A. Kong, K.-H. Ding, and C. O. Ao, *Scattering of Electromagnetic Waves: Numerical Simulations*, Wiley, New York, 2001.
[26] V. Twersky, *On scattering of waves by random distributions.* I. *Free-space scatterer formalism*, J. Math. Phys., 3 (1962), pp. 700–715.
[27] R. J. Urick and W. S. Ament, *The propagation of sound in composite media*, J. Acoust. Soc. Amer., 21 (1949), pp. 115–119.
[28] P. C. Waterman and R. Truell, *Multiple scattering of waves*, J. Math. Phys., 2 (1961), pp. 512–537.
[29] R. C. Wittmann, *Spherical wave operators and the translation formulas*, IEEE Trans. Antennas & Propag., 36 (1988), pp. 1078–1087.
[30] Z. Ye and L. Ding, *Acoustic dispersion and attenuation relations in bubbly mixture*, J. Acoust. Soc. Amer., 98 (1995), pp. 1629–1636.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1669–1688

NON-LORENTZIAN SPECTRAL LINESHAPES NEAR A HOPF BIFURCATION*

J. P. GLEESON† AND F. O'DOHERTY†

Abstract. The effects of additive white noise upon the dynamics of a system described by its Hopf normal form are investigated, with particular reference to the well-known model of a detuned single-mode laser. The power spectrum corresponding to the laser amplitude is determined by finite-difference solution of a partial differential equation, and analytical formulas are determined in the asymptotic limits of large parameters. The effect of the amplitude-phase coupling parameter in generating non-Lorentzian lineshapes is highlighted, and the regions of parameter space where accurate first-eigenvalue approximations of the Fokker–Planck equation exist are indicated.

Key words. white noise, stochastic differential equations, Hopf normal form, Fokker–Planck equation, semiconductor lasers, oscillators

AMS subject classifications. 70K45, 82C31, 37N20, 60H40

DOI. 10.1137/040615146

1. Introduction.

In this paper we consider the effects of white noise upon the dynamics of a system near a Hopf bifurcation point. The well-known deterministic (noise-free) Hopf normal form is [1]

$$\frac{dE}{dt} = i\Omega E + (1 + i\delta)\left(a - |E|^2\right) E, \tag{1}$$

where $E(t)$ is a complex-valued function of time representing, for example, the electric field of a single-mode laser in a semiclassical approximation [2] with fundamental frequency Ω. The parameters a and δ are real-valued and dimensionless; see the Appendix for details of their derivation from the standard normal form parameters. This equation can be written in amplitude-phase coordinates using $E = r\exp(i\theta)$, with the dynamics of the polar coordinates $r(t)$ and $\theta(t)$ being governed by the pair of equations

$$\begin{aligned} \frac{dr}{dt} &= \left(a - r^2\right) r, \\ \frac{d\theta}{dt} &= \Omega + \delta\left(a - r^2\right). \end{aligned} \tag{2}$$

As discussed in Chapter 12 of [2], noise effects may be incorporated into the deterministic system by the addition of a complex Langevin fluctuation term

$$\Gamma(t) = \xi_x(t) + i\,\xi_y(t) \tag{3}$$

to the right-hand side of (1). The zero-mean, real-valued white noise processes $\xi_x(t)$ and $\xi_y(t)$ are described by their delta-function autocorrelation functions [2]:

$$\langle \xi_x(t)\xi_x(t')\rangle = \langle \xi_y(t)\xi_y(t')\rangle = 2\delta(t - t'), \qquad \langle \xi_x(t)\xi_y(t')\rangle = 0. \tag{4}$$

*Received by the editors September 15, 2004; accepted for publication (in revised form) February 7, 2006; published electronically July 17, 2006. This work was supported by a Science Foundation Ireland Investigator Award to the first author, program 02/IN.1/IM062.
http://www.siam.org/journals/siap/66-5/61514.html
†Applied Mathematics, University College Cork, Cork, Ireland (j.gleeson@ucc.ie, fergal1979@hotmail.com).

Note that the noise intensity is set to unity by our choice of length and time units (see the Appendix for details). Angle brackets are used throughout this paper to denote averaging over an ensemble, i.e., taking an expectation value. Our goal is to describe the effects of the noise terms upon the dynamics of the system (2).

Any system undergoing a Hopf bifurcation may be written in the Hopf normal form (1). The deterministic dynamics of the Hopf bifurcation are well understood. When the bifurcation parameter a is negative, the origin $r = 0$ is an attracting point for all trajectories of the system (2) in the absence of noise. As the bifurcation parameter passes through zero and becomes positive, the origin becomes unstable and trajectories are attracted to the stable limit cycle at $r = \sqrt{a}$. The flow on the limit cycle is purely circular, with angular speed Ω. The parameter δ does not affect these steady-state results, but it has important consequences when the system is subject to fluctuations. For example, in the stable limit cycle case $a > 0$, trajectories which are kicked off the limit cycle eventually flow back onto it, moving at an angular speed which varies with their amplitude when $\delta \neq 0$ and being equal to the steady-state speed Ω only when on the limit cycle. Because δ quantifies this effect of amplitude fluctuations upon the phase angle θ, it is sometimes referred to as the *amplitude-phase coupling parameter*. When $a < 0$ the flow into the attracting origin has a spiral structure. Nonzero δ alters the angular speed of the spiral motion and induces a differential rotation effect; i.e., trajectories further from the origin spiral inwards at rates different from those closer to $r = 0$.

In this paper we study the effect of noise on the dynamics of (2), with particular reference to the *correlation function* of E,[1]

$$R(\tau) = \langle E(t)E^*(t+\tau)\rangle\,, \tag{5}$$

or to its *power spectrum*, which is found by Fourier transforming $R(\tau)$. The deterministic limit cycle solution described above in the case $a > 0$ is periodic, and so the power spectrum of E exhibits a delta-function spike at the frequency Ω. The effect of small random fluctuations modeled by the noise terms ξ_x and ξ_y in (3) is to broaden the peak in the power spectrum, and so generate a finite linewidth. For $a < 0$ the stable solution is $r = 0$, but "precursor" peaks in the power spectrum of E can be seen: these grow in intensity as the bifurcation parameter a is increased towards the bifurcation point [3, 4, 5]. The stable oscillation generated when $a > 0$ is generic and has been studied in many fields, including electric and electronic engineering, chemical physics, and laser physics. As the system (2) has been extensively studied in the laser literature (at least for the case $\delta = 0$), for convenience we will adopt the nomenclature and notation specific to that field. However, our results are applicable to any system of the form (2) in the presence of additive white noise. Recent examples where nonzero values of the parameter δ are important arise in applications to electronic oscillator circuits [5, 6, 7] and to chemical reaction systems [8]. For instance, Coram [7], as a test case for various recently proposed methods [5, 9] for predicting the spectral lineshape of noisy electronic oscillator circuits, proposes a model which (in the limit of small noise) is essentially equivalent to (2). The important new effects due to the amplitude-phase coupling parameter δ demonstrated in this paper are thus also immediately applicable to electronic oscillator models such as Coram's.

In the laser literature the parameter a is commonly referred to as the *pump parameter*, with the regime $a < 0$ termed *below (lasing) threshold*, and $a > 0$ *above*

[1] In the laser physics literature the function $R(\tau)$ is known as the amplitude correlation function; here star denotes complex conjugation.

threshold. The parameter δ has received comparatively little attention in the literature, owing to its relative insignificance in the most commonly studied laser systems. Depending on the physical causes of the effect, the parameter has variously been named the *detuning parameter* [2], the *linewidth enhancement factor* [10], or as noted above, the *amplitude-phase coupling parameter*. The effect of δ is known to be non-negligible in semiconductor lasers, and so we believe this study of its effects is timely. As mentioned above, the Hopf normal form (2) is frequently used as the simplest model of a noisy, self-sustained oscillator in many other fields, e.g., electronic circuit design [7] and chemical reaction dynamics [8], and so a complete description of the power spectrum beyond the well-studied $\delta = 0$ case is of some importance.

The dynamics of Hopf bifurcations with additive noise, but with $\delta = 0$, have been studied by several authors. For instance, Baras, Mansour, and Van den Broek [11] discuss in detail the stationary probability density near a Hopf bifurcation point in the weak noise limit. In the review paper [12] the more general stationary distribution problem for a noisy Hopf bifurcation with a noncircular limit cycle is discussed. The effects of multiplicative [13, 14] or colored noises [15, 16] near a Hopf bifurcation have also been considered, as well as noisy bifurcations in delay-differential equations [17]. Our work concentrates on the simplest additive white noise case, and is new in two respects: first, it describes the spectral lineshape (not just the stationary probability density) of the dynamical system near the bifurcation point, and, second, to our knowledge, this is the first comprehensive study of the important effects of the amplitude-phase coupling parameter δ—most previous studies assume δ to be zero or negligibly small. While we use standard results for the stationary probability density near the bifurcation point (e.g., see (8) below), the focus of our work is on the dynamical effect of δ, as evidenced by the shape of the power spectrum. The distinction is crucial—the stationary probability density is independent of δ, while the spectral lineshape is very sensitive to nonzero δ values.

The most complete study of the δ-effect to date appears to be that of Seybold and Risken [18], referred to hereafter as SR. They approach the Fokker–Planck equation corresponding to (2) using the same methods that were successful in the $\delta = 0$ case [2], i.e., an eigenfunction expansion of the transition probability density. The resulting correlation function is then an infinite sum of exponentials, yielding a power spectrum composed of a sum of Lorentzians. Although the calculation of eigenvalues must be done numerically, a crucial result of SR is that for $\delta < 1.2$ the first term in the infinite series is dominant, and so only this first eigenvalue term need be used to find the spectrum to a high order of accuracy. As shown in Figure 1, this single-Lorentzian expression for the spectrum is accurate to power levels many orders of magnitude below the peak. The divergence between the exact spectrum and the SR approximation seen at high frequencies in Figure 2 can be traced to the inaccuracies in the small-time expansion of the correlation function, as noted in SR. However, near the peak of the spectrum (within, say, three orders of magnitude in power level) the SR Lorentzian form of the spectrum is very accurate for $\delta \lesssim 1$. Thus the description of the spectral lineshape (and linewidth) is reduced to finding the appropriate eigenvalue of the Fokker–Planck equation. We note in passing that this first-eigenvalue approach is also used in recent papers attempting to predict the spectral lineshape arising from noise in electronic oscillator circuits [5, 6] and limit cycles modeling chemical clocks [8].

The SR paper covers only the case $\delta < 1.2$, which encompassed the values of the parameter δ of interest to the laser community at the time. However, more recent research in the field of semiconductor lasers indicates that the values of δ

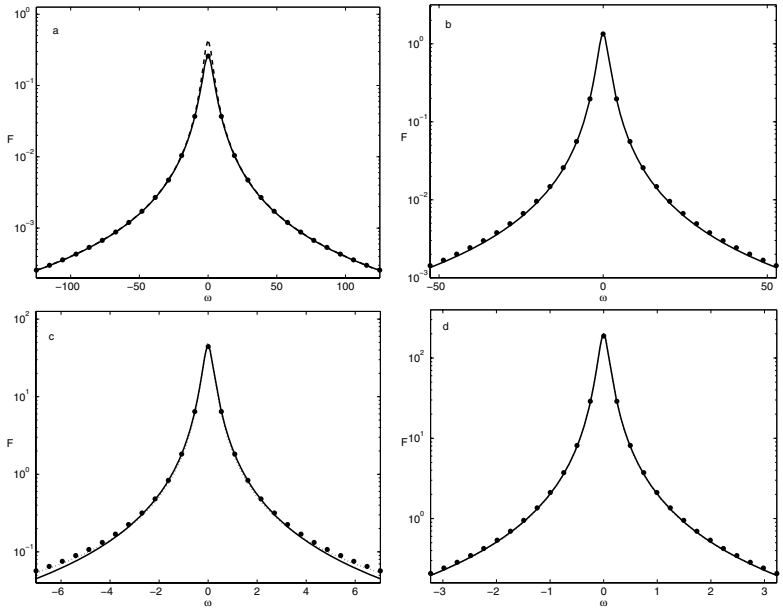


FIG. 1. *Spectrum $F(\omega)$ for $\delta = 0$ and various values of a:* (a) $a = -3$, (b) $a = 0$, (c) $a = 5$, (d) $a = 10$. *Results from the numerical solution* (25) *are shown as symbols; the solid line is the Lorentzian lineshape* (31)*; the dashed (for $a < 0$) and dotted (for $a > 0$) lines show the asymptotic results of section* 4.

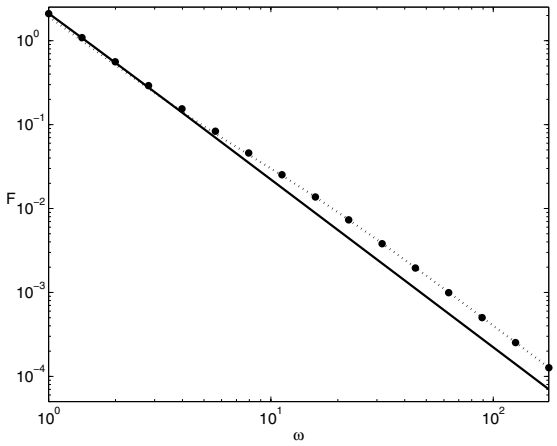


FIG. 2. *Log-log plot of the spectrum $F(\omega)$ for large frequencies, with parameters $a = 5$ and $\delta = 0$ (cf. Figure* 1(c)*). Line types are as in Figure* 1.

appropriate to the model (2) are on the order of 5 to 7. The most obvious effect of this larger δ is a significant widening of the spectral line—hence the title of *linewidth enhancement factor* (or simply *α-factor*) which is sometimes applied to δ. Implicit in most discussions of the linewidth of semiconductor lasers is the belief that the lineshape is intrinsically Lorentzian. Indeed some workers have followed the methods of SR to calculate the first eigenvalue in the case $\delta \geq 5$, and to compare the resulting linewidth to experimental calculations [19]. However, our results, using numerical

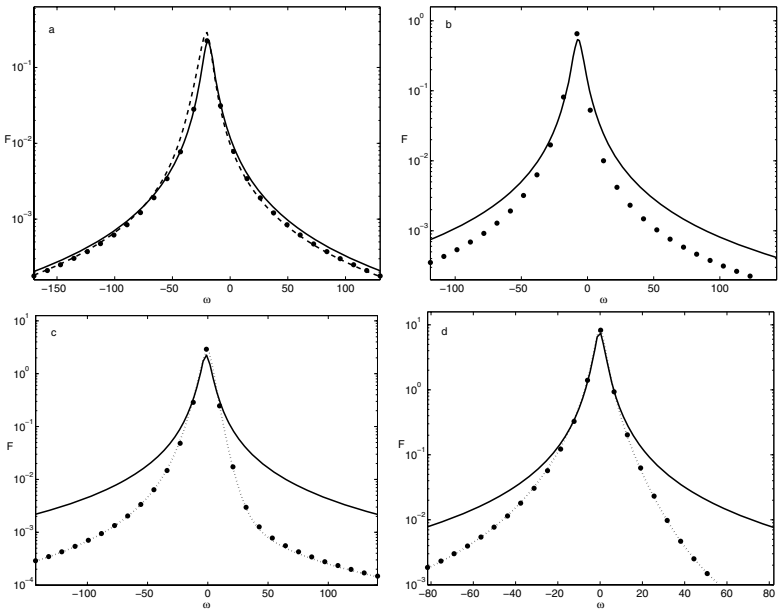


FIG. 3. *Spectrum $F(\omega)$ for $\delta = 5$ and various values of a:* (a) $a = -3$, (b) $a = 0$, (c) $a = 5$, (d) $a = 10$. *See Figure* 1 *for description of line types.*

solutions and asymptotic analysis of the Fokker–Planck equation (see Figures 3 and 4), indicate that the lineshape is markedly non-Lorentzian at $\delta = 5$ (at least for certain values of the pump parameter a). It is therefore useful to experimentalists working with δ levels higher than those examined in SR to have numerical methods and closed form asymptotic solutions which are valid for arbitrarily large values of the parameter δ.

The main results of this paper are presented in caricature form in Figure 5, showing the (a, δ) parameter plane. Since the spectrum for negative δ can be found from the corresponding positive δ result by changing the sign of frequencies ($\omega \to -\omega$, $\Omega \to -\Omega$), we consider only $\delta \geq 0$. The hatched region is where the SR first-eigenvalue Lorentzian lineshape fits reasonably well[2] to the exact spectrum, which we determine by a finite-difference numerical solution of (23) below. As originally shown in SR, the Lorentzian lineshape fits well for all values of the pumping parameter a, provided that δ is less than approximately unity. When δ is significantly greater than one, SR remains accurate very far from the threshold at $a = 0$, but the required magnitude of the pumping increases as δ increases. For instance, we show that for $a < 0$ the SR solution is accurate when the ratio $\delta/a^2 \ll 1$; the corresponding requirement for positive pumping parameter is that $\delta/a \ll 1$. These constraints on the SR solution give the borders of the hatched region above $\delta = 1$. These limits on the single-Lorentzian form of the spectrum are derived using new asymptotic solutions, valid for $|a| \gg 1$ and for all δ values. Our new asymptotic solutions are found to be accurate outside of the region of the (a, δ) plane enclosed by the dashed lines at $a = -5$ and $a = 3$, and show that non-Lorentzian lineshapes are increasingly

[2]The precise meaning of "reasonably well" of course depends on the level of accuracy desired; Figure 5 is intended to indicate only the estimated (order of magnitude) borders where various approximations hold.

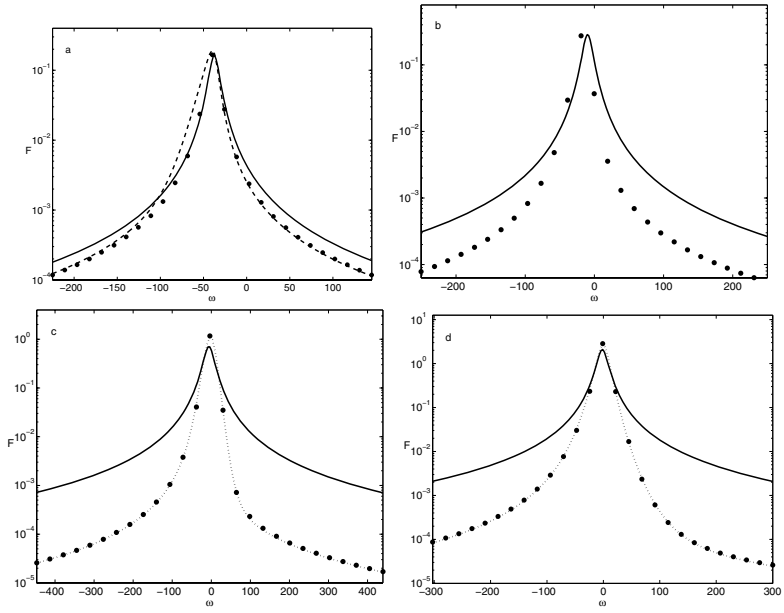


FIG. 4. *Spectrum $F(\omega)$ for $\delta = 10$ and various values of a:* (a) $a = -3$, (b) $a = 0$, (c) $a = 5$, (d) $a = 10$. *See Figure* 1 *for description of line types.*

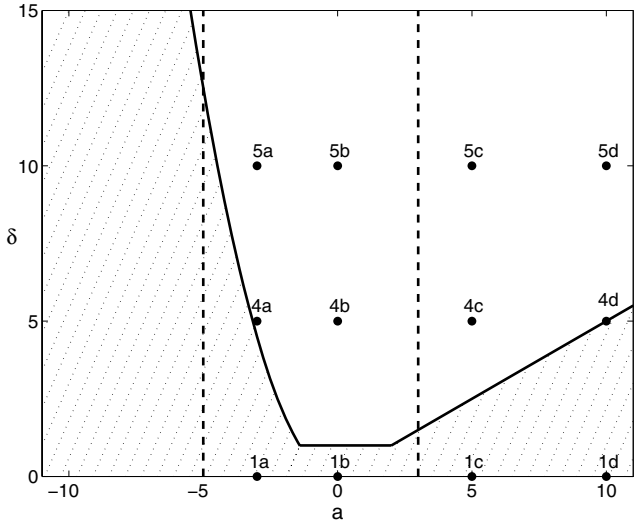


FIG. 5. *The (a, δ) parameter plane. The shaded region shows where the single-Lorentzian spectrum of SR matches the exact result reasonably well; the boundaries are taken to be $\delta/a^2 = 1/2$ for $a < 0$, and $\delta/a = 1/2$ for $a > 0$. The new asymptotic results derived in section* 4 *match the exact results well when $a \ll -5$ or $a \gg 3$. These limits are shown as dashed lines in the figure—note the considerable regions of the parameter plane (e.g., for $a \gg 3$ and $\delta \gg a/2$) where our asymptotic results match the exact values, but those of SR do not. The labeled dots mark the parameter values where spectra are plotted in the corresponding figures.*

important as δ increases. The analytical expressions are likely to be useful not only to experimentalists characterizing lineshapes with $\delta > 1$, but also in providing a novel test case for schemes attempting to characterize spectral lineshapes in self-sustained oscillators of arbitrary dimension in various applications [5, 20].

2. Fokker–Planck equation. The dynamics of the system (1) in the presence of the noise term (3) is fully described by the solution of the Fokker–Planck equation [2], which gives the transition probability density $P(\mathbf{x}, \tau|\mathbf{x}', 0)$ of the two-dimensional stochastic process $\mathbf{x}(t)$, written in polar coordinates as $\mathbf{x}(t) = (r(t)\cos\theta(t), r(t)\sin\theta(t))$ using (2). The transition probability density is the probability that a trajectory of the system passes through the point $\mathbf{x}$ at time $t = \tau$, given that it passed through $\mathbf{x}'$ at time $t = 0$. The Fokker–Planck equation is a partial differential equation for P:

(6)
$$\frac{\partial P}{\partial \tau} + \frac{1}{r}\frac{\partial}{\partial r}\left[\left(a - r^2\right) r^2 P\right] + \frac{\partial}{\partial \theta}\left[\left(\Omega + \delta\left(a - r^2\right)\right) P\right] - \frac{1}{r}\frac{\partial}{\partial r}\left[r\frac{\partial P}{\partial r}\right] - \frac{1}{r^2}\frac{\partial^2 P}{\partial \theta^2} = 0,$$

with initial condition

$$P(\mathbf{x}, 0|\mathbf{x}', 0) = \delta(\mathbf{x} - \mathbf{x}'), \tag{7}$$

where here $\delta(\mathbf{x})$ is the Dirac delta-function (not to be confused with the parameter δ, as should be clear from the context). Equation (6) has a stationary ($\tau \to \infty$) solution which depends only on r (and is independent of δ), as shown in [11, 2]:

$$P_\infty(r) = C\exp\left(\frac{ar^2}{2} - \frac{r^4}{4}\right). \tag{8}$$

The normalization constant C is given in terms of the error function as

$$C = \frac{\exp\left(-\frac{a^2}{4}\right)}{\pi^{\frac{3}{2}}\left[1 + \operatorname{erf}\left(\frac{a}{2}\right)\right]}, \tag{9}$$

in order to ensure unit probability over the two-dimensional space

$$\int_0^{2\pi}\int_0^{\infty} P_\infty(r) r\, dr\, d\theta = 1. \tag{10}$$

Note that stationary moments may be calculated from $P_\infty(r)$, for example [2]:

$$\left\langle r^2 \right\rangle = a + 2\pi C. \tag{11}$$

The transition probability density may be used to calculate the correlation between arbitrary functionals $f(\mathbf{x})$ and $g(\mathbf{x})$ of the dynamical system trajectories at times separated by τ [2, 21]:

$$\langle f\left(\mathbf{x}(t)\right) g\left(\mathbf{x}(t+\tau)\right)\rangle = \int d\mathbf{x}'\, P_\infty(\mathbf{x}') f\left(\mathbf{x}'\right) \int d\mathbf{x}\, P(\mathbf{x}, \tau|\mathbf{x}', 0) g\left(\mathbf{x}\right), \tag{12}$$

where each of the integrals above is over the two-dimensional trajectory space. The correlation function may be directly related to the Fokker–Planck equation by defining an auxiliary function $Q(\mathbf{x}, \tau)$ as

$$Q(\mathbf{x}, \tau) = \int P_\infty\left(\mathbf{x}'\right) f\left(\mathbf{x}'\right) P(\mathbf{x}, \tau|\mathbf{x}', 0)\, d\mathbf{x}'. \tag{13}$$

Note that this auxiliary function is a solution of the Fokker–Planck differential equation (6):

(14)
$$\frac{\partial Q}{\partial \tau} + \frac{1}{r}\frac{\partial}{\partial r}\left[\left(a - r^2\right) r^2 Q\right] + \frac{\partial}{\partial \theta}\left[\left(\Omega + \delta\left(a - r^2\right)\right) Q\right] - \frac{1}{r}\frac{\partial}{\partial r}\left[r\frac{\partial Q}{\partial r}\right] - \frac{1}{r^2}\frac{\partial^2 Q}{\partial \theta^2} = 0,$$

but with an initial condition that differs from (7):

$$Q(\mathbf{x}, 0) = f(\mathbf{x})\, P_\infty(\mathbf{x}). \tag{15}$$

Assuming that the auxiliary function has been determined, the correlation function is then obtained by integration:

$$\langle f(\mathbf{x}(t))\, g(\mathbf{x}(t+\tau))\rangle = \int g(\mathbf{x}) Q(\mathbf{x}, \tau)\, d\mathbf{x}. \tag{16}$$

The correlation function of the complex quantity E defined in (5) may be calculated by applying this general method, using the functionals

$$f(\mathbf{x}) = r\exp(i\theta) \quad \text{and} \quad g(\mathbf{x}) = r\exp(-i\theta) \tag{17}$$

and the stationary probability density given by (8) and (9). In this case, the auxiliary function is defined as the solution of (14) with initial condition

$$Q(r, \theta, 0) = re^{i\theta} P_\infty(r), \tag{18}$$

and the desired correlation function is found from Q by integration:

$$R(\tau) = \langle E(t) E^*(t+\tau)\rangle = \int_0^{2\pi}\int_0^\infty r^2 e^{-i\theta} Q(r, \theta, \tau)\, dr\, d\theta. \tag{19}$$

The power spectrum of E is defined using the Fourier transform of the correlation function, following the convention of SR:

$$F(\omega) \equiv 2\,\mathrm{Re}\left[\int_0^\infty e^{i\omega\tau} R(\tau)\, d\tau\right]. \tag{20}$$

This formulation of the problem in terms of the auxiliary function Q proves extremely useful for both numerical calculations and asymptotic analysis, and thus forms the theoretical basis for the remainder of the paper. While the formulation (12) for correlation functions is common [2, 21], the reduction to a partial differential equation (pde) for an auxiliary function Q is a nonstandard approach. This idea is potentially very useful for the calculation of correlation functions and spectra in low-dimensional systems, where standard numerical methods for pdes may be applied, as in (23) below.

3. Numerical calculations. In this section we compare two methods for numerical calculation of the spectrum. The auxiliary function approach presented in section 3.1 provides the high-accuracy solutions used throughout the paper; the single-eigenvalue spectrum used by Seybold and Risken is also reviewed in section 3.2.

3.1. Auxiliary function approach. The solution of the pde (14) for the auxiliary function is facilitated by noting that the initial condition (18) permits a separable form:

$$Q(r,\theta,\tau) = e^{i\theta} q(r,\tau). \tag{21}$$

Moreover, the power spectrum $F(\omega)$ may be calculated without first finding the correlation function. To see this, multiply (14) by $\exp(i\omega\tau)$ and integrate over τ. The resulting equation for the quantity $\hat{q}(r,\omega)$, defined as

$$\begin{aligned} \hat{q}(r,\omega) &\equiv \int_0^\infty e^{i\omega\tau} q(r,\tau)\, d\tau \\ &= \int_0^\infty e^{i\omega\tau} e^{-i\theta} Q(r,\theta,\tau)\, d\tau, \end{aligned} \tag{22}$$

is an ordinary differential equation, parameterized by the frequency ω:

$$-i\omega\hat{q} - rP_\infty(r) + \frac{1}{r}\frac{d}{dr}\left[\left(a - r^2\right) r^2 \hat{q}\right] + i\Omega\hat{q} + i\delta\left(a - r^2\right)\hat{q} - \frac{1}{r}\frac{d}{dr}\left[r\frac{d\hat{q}}{dr}\right] + \frac{1}{r^2}\hat{q} = 0. \tag{23}$$

The boundary conditions on $\hat{q}$ are

$$\begin{aligned} \hat{q} &\to 0 \quad \text{as } r \to 0, \\ \hat{q} &\to 0 \quad \text{as } r \to \infty. \end{aligned} \tag{24}$$

Equation (23) is solved numerically for each frequency ω using a standard finite-difference algorithm. The boundary condition for $r \to \infty$ is implemented by truncating the finite-difference grid at a large value of r and setting a Dirichlet condition at the truncation point. We check that this truncation point is sufficiently large to have negligible effect on all spectra shown. Finally, the spectrum $F(\omega)$ is calculated directly from $\hat{q}$ by numerical integration:

$$F(\omega) = 4\pi\,\mathrm{Re}\left[\int_0^\infty r^2 \hat{q}(r,\omega)\, dr\right]. \tag{25}$$

The accuracy of the finite-difference solution depends only on the number of grid points and on the truncation point, and so effectively gives an exact solution for the spectrum. The values found from this numerical method are plotted as symbols in the figures and enable us to compare various analytical approximations for the spectrum.

3.2. Eigenfunction expansion. The formulation given in SR (see also [2]) uses (12) to calculate the correlation function, but instead of defining the auxiliary function and solving (14), the Fokker–Planck equation (6) for the transition probability density $P(\mathbf{x},\tau|\mathbf{x}',0)$ is solved directly. The solution may be found by an eigenfunction expansion [18] or by expansion into a complete set as in [2], i.e., writing P as

$$P(r,\theta,\tau|r',\theta',0) = \frac{1}{2\pi} e^{-r^2/\alpha} \sum_{n=0}^{\infty} \sum_{\nu=-\infty}^{\infty} c_n^{(\nu)} \left(\frac{r'^2}{\alpha}, \tau\right) r^{|\nu|} \alpha^{-|\nu|/2} L_n^{(|\nu|)}\left(\frac{r^2}{\alpha}\right) e^{i\nu(\theta-\theta')}. \tag{26}$$

Here $L_n^{(|\nu|)}$ denotes the generalized Laguerre function, and α is a scaling parameter which can be varied to improve the accuracy of a truncated series approximation

to (26). The coefficients $c_n^{(\nu)}$ satisfy ordinary differential equations, which may be written as

$$\frac{dc_n^{(\nu)}}{d\tau} = \sum_{m=0}^{\infty} A_{nm}^{(\nu)} c_m^{(\nu)}, \tag{27}$$

with initial conditions derived from (7); see [2] for details. To solve numerically, we truncate the sums over n and m in (26) and (27) at a large integer N, and for each ν solve for $c_n^{(\nu)}$ using the matrix exponential of the $(N+1) \times (N+1)$ complex-valued matrix $\mathbf{A}^{(\nu)}\tau$. The final result for the correlation function (19) is then of the form

$$R_{SR}(\tau) = \langle r^2 \rangle \sum_{n=0}^{N} V_n \, e^{\Lambda_n \tau}, \tag{28}$$

where Λ_n are the eigenvalues of the matrix $\mathbf{A}^{(1)}$, and the V_n depend upon the eigenvectors and the initial conditions. The value of $\langle r^2 \rangle$ is given in (11). All Λ_n have negative real parts, so the dominant term when the separation time τ is large gives

$$R_{SR}(\tau) \sim \langle r^2 \rangle \, V_0 \, e^{\Lambda_0 \tau} \quad \text{as } \tau \to \infty, \tag{29}$$

where Λ_0 is the eigenvalue of $\mathbf{A}^{(1)}$ with largest real part. The corresponding spectrum is

$$F_{SR}(\omega) = 2 \langle r^2 \rangle \, \mathrm{Re}\left[\frac{V_0}{-\Lambda_0 - i\omega} \right]. \tag{30}$$

Seybold and Risken [18] take $V_0 \equiv 1$ (with error of less than 3% when $\delta \leq 1.2$), and we therefore compare our numerical results to the SR first-eigenvalue spectrum

$$F_{SR} = 2 \langle r^2 \rangle \, \frac{|\Lambda_{0r}|}{{\Lambda_{0r}}^2 + (\Lambda_{0i} + \omega)^2}, \tag{31}$$

with Λ_{0r} and Λ_{0i} being the real and imaginary parts of Λ_0. This Lorentzian spectrum is peaked at frequency $\omega = \omega_p = -\Lambda_{0i}$ and reaches half of its peak power at frequencies $\omega = \omega_p \pm \Delta\omega$, where $\Delta\omega = |\Lambda_{0r}|$. The frequency difference $\Delta\omega$ gives the half-width of the Lorentzian spectrum at half of the peak value, and when normalized by $\langle r^2 \rangle$ it is referred to as the half-width-at-half-maximum (HWHM), or simply *linewidth*, of the laser spectral line. The spectrum (31) is plotted as a solid line in the figures for comparison with numerical values calculated using the auxiliary function method (symbols), and with asymptotic approximations. The linewidth is examined further in Figures 6 and 7, see section 4.1.

4. Asymptotic solutions. To simplify our notation and conform with the usual conventions in the laser physics literature [18], we set the fundamental frequency Ω to zero in (2). Note that this can formally be accomplished by moving to a rotating reference frame, i.e., changing the angular variable from θ to $\theta - \Omega t$. This change of frame shifts the spectrum so that the peak at $\omega = \Omega$ in the original frame is moved to $\omega = 0$. All results reported here are thus at frequencies ω relative to the fundamental frequency Ω.

4.1. Above threshold: $a \gg 1$. The deterministic dynamics of (2) for positive values of the pump parameter a lead to limit cycles of radius $\sqrt{a}$, as discussed in section 1. When a is sufficiently large, the effects of the (order one) noise terms on the amplitude are small compared to $\sqrt{a}$. The behavior of the dynamics close to the deterministic limit cycle may be examined by making the change of variables

$$r = \sqrt{a} + \frac{\rho}{\sqrt{2a}} \tag{32}$$

in (14). This scaling focuses close to the limit cycle at $r = \sqrt{a}$, with (as shown below) the new variable ρ being of order one in the region of interest. The variable ρ is interpreted as a linear variable, with range $(-\infty, \infty)$ in the $a \to \infty$ limit. The two-dimensional area integral is modified as follows:

$$\int_0^{2\pi} \int_0^{\infty} r\, dr\, d\theta \longrightarrow \frac{1}{\sqrt{2}} \int_0^{2\pi} \int_{-\infty}^{\infty} d\rho\, d\theta, \tag{33}$$

and the stationary probability density becomes

$$P_\infty(\rho) = \frac{1}{2\pi^{3/2}} \exp\left(-\frac{\rho^2}{2}\right). \tag{34}$$

Using (32) in (14) leads to the linearized equation

$$\frac{\partial Q}{\partial \tau} - 2a \frac{\partial}{\partial \rho}(\rho Q) - \sqrt{2}\delta\rho \frac{\partial Q}{\partial \theta} - 2a \frac{\partial^2 Q}{\partial \rho^2} - \frac{1}{a} \frac{\partial^2 Q}{\partial \theta^2} = 0, \tag{35}$$

where the first term of an asymptotic expansion for $a \to \infty$ replaces each term in (14). The initial condition (18) is expanded to

$$Q(\tau = 0) = \sqrt{a} \left(1 + \frac{\rho}{\sqrt{2a}}\right) e^{i\theta} P_\infty(\rho). \tag{36}$$

Equation (35) with initial condition (36) may be solved exactly by Fourier transforming in ρ and using the method of characteristics [2]. Separating variables as in (21) and defining the spatial Fourier transform as

$$\tilde{q}(k, \tau) = \int_{-\infty}^{\infty} e^{i\rho k} q(\rho, \tau)\, d\rho, \tag{37}$$

we obtain the hyperbolic equation

$$\frac{\partial \tilde{q}}{\partial \tau} + \left(2ak - \sqrt{2}\delta\right) \frac{\partial \tilde{q}}{\partial k} = -\left(2ak^2 + \frac{1}{a}\right) \tilde{q} \tag{38}$$

with initial condition

$$\tilde{q}(k, 0) = \frac{1}{\pi} \sqrt{\frac{a}{2}} \left(1 + i \frac{k}{\sqrt{2a}}\right) e^{-\frac{1}{2}k^2}. \tag{39}$$

Solving (38) by the method of characteristics leads to the solution

$$\tilde{q}(k, \tau) = \frac{1}{\pi} \sqrt{\frac{a}{2}} \left[f_1(\tau) + f_2(\tau) k\right] \exp\left[-\frac{1}{2}k^2 + f_3(\tau) k - D(\tau)\right], \tag{40}$$

where

$$f_1(\tau) = 1 + i\frac{\delta}{2a^2}\left(1 - e^{-2a\tau}\right),$$
$$f_2(\tau) = i\frac{1}{\sqrt{2}a}e^{-2a\tau},$$
$$f_3(\tau) = \frac{\delta}{\sqrt{2}a}\left(e^{-2a\tau} - 1\right),$$
$$D(\tau) = \frac{1}{a}\tau + \frac{\delta^2}{2a^2}\left(2a\tau + e^{-2a\tau} - 1\right). \tag{41}$$

The correlation function (19) can be related directly to the Fourier transform of Q via

$$R(\tau) = \sqrt{2a}\pi\left[\tilde{q} - i\frac{1}{\sqrt{2}a}\frac{\partial\tilde{q}}{\partial k}\right]\Bigg|_{k=0}, \tag{42}$$

and the final result may be written as

$$R(\tau) = a\left(a_0 + a_1 e^{-2a\tau} + a_2 e^{-4a\tau}\right)e^{-D(\tau)}, \tag{43}$$

where the coefficients a_n depend upon the parameters as follows:

$$a_0 = \left(1 + \frac{\delta}{2a^2}i\right)^2,$$
$$a_1 = \frac{1}{2a^2}\left(1 + \frac{\delta^2}{a^2} - 2\delta i\right),$$
$$a_2 = -\frac{\delta^2}{4a^4}. \tag{44}$$

Note that for convenience we have assumed $\tau > 0$ in the above; the symmetry of R implies that for negative arguments τ should be replaced by $|\tau|$.

Analytical expressions may be found for the Fourier transform of (43). For instance, a series expansion of the $\exp(D(\tau))$ term allows the autocorrelation to be written as

$$R(\tau) = a\exp\left(\frac{\delta^2}{2a^2}\right)\sum_{n=0}^{\infty} b_n e^{-\lambda_n\tau}, \tag{45}$$

with $\lambda_n = 2an + \frac{1}{a}\left(\delta^2 + 1\right)$, and

$$b_0 = \left(1 + \frac{\delta^2}{2a^2}i\right)^2, \tag{46}$$

$$b_1 = \frac{1}{2a^2}\left(1 + \frac{\delta^2}{2a^2} - i\delta\right)^2, \tag{47}$$

$$b_n = \frac{(-1)^{n+1}}{n!}\frac{\delta^{2n-2}}{2^n a^{2n}}\left[n + \frac{\delta^2}{2a^2} - i\delta\right]^2 \quad \text{for } n \geq 2. \tag{48}$$

The corresponding spectrum is then an infinite sum of Lorentzians,

$$F(\omega) = 2a\exp\left(\frac{\delta^2}{2a^2}\right)\mathrm{Re}\left[\sum_{n=0}^{\infty}\frac{b_n}{\lambda_n - iw}\right]. \tag{49}$$

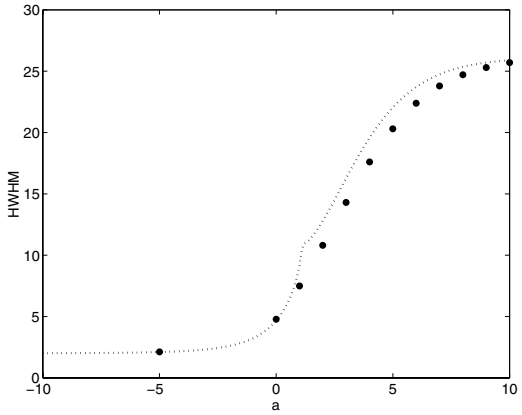


FIG. 6. *The normalized linewidth (HWHM) $\langle r^2 \rangle \Delta\omega$ of the spectra as a function of a for $\delta = 5$. Numerical results are shown by symbols; the dotted line is the SR first-eigenvalue result.*

Note that this sum converges very slowly when $|\delta| \gg 1$, so that many terms are required to give the spectrum accurately. Also, the leading-order dependence on δ is of the form δ/a; this implies that non-Lorentzian lineshapes arise when δ is a nonnegligible fraction of a, and leads us to sketch the estimated border $\delta = a/2$ for positive a in Figure 5.

An alternative expression for the spectrum may be given in terms of a finite sum, but involving the generalized incomplete gamma function [22, 23], defined as

$$\gamma(y, z) = \int_0^z t^{y-1} e^{-t} dt. \tag{50}$$

In terms of this function, the spectrum is

$$F(\omega) = \exp\left(\frac{\delta^2}{2a^2}\right) \operatorname{Re}\left[\sum_{n=0}^{2} a_n \left(\frac{2a^2}{\delta^2}\right)^{\frac{\lambda_n - i\omega}{2a}} \gamma\left(\frac{\lambda_n - i\omega}{2a}, \frac{\delta^2}{2a^2}\right)\right], \tag{51}$$

with the coefficients a_n as defined in (44). The spectrum (51) is plotted with a dotted line in parts (c) and (d) of Figures 1, 3, and 4, and matches the numerical spectrum (symbols) very well, even for values of a as low as 5. The SR Lorentzian spectrum (solid line) fits poorly when $\delta > 0$. Close to the peak of the spectrum the fit of the SR approximation can be quantified by the HWHM of the lineshape, as defined in section 3.2. The fit of the numerical linewidth to the SR prediction is reasonably good at $\delta = 5$ (Figure 6), indicating that the non-Lorentzian effects have most impact away from the peak; however, the SR linewidth estimation decays in accuracy as δ increases; see the $\delta = 10$ case in Figure 7.

4.1.1. Large-frequency spectrum. A small-τ expansion of $R(\tau)$ from (43) yields

$$R(\tau) \sim a - 2(1 - i\delta)|\tau| + O\left(\tau^2, a^{-1}\right) \quad \text{as } \tau \to 0,\ a \to \infty, \tag{52}$$

and the corresponding large-frequency asymptotic form of the spectrum $F(\omega)$ from (20) is

$$F(\omega) \sim \frac{4}{\omega^2} + O\left(\omega^{-3}\right) \quad \text{as } \omega \to \infty. \tag{53}$$

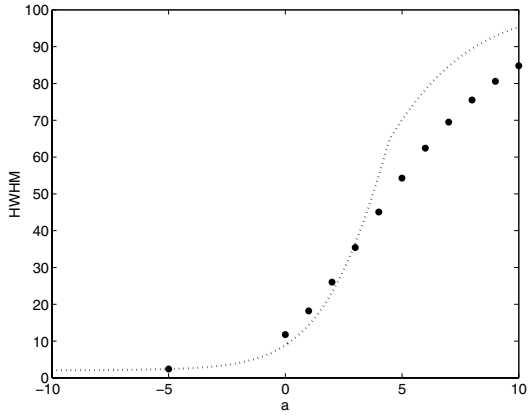


FIG. 7. *The normalized linewidth (HWHM)* $\langle r^2 \rangle \Delta\omega$ *of the spectra as a function of* a *for* $\delta = 10$. *Line types are as in Figure* 6.

Note that this leading order behavior of the spectrum can be derived directly from a small-τ expansion in (14) and (19), and is independent of the values of the parameters a and δ.

In the $a \gg 1$ asymptotic limit, the single-Lorentzian form of the spectrum used in SR corresponds to the correlation function

$$R_{\mathrm{SR}}(\tau) = a \exp\left[-\frac{1}{a}\left(1+\delta^2\right)|\tau|\right]. \tag{54}$$

The large-frequency asymptote of the SR spectrum is then [18]

$$F_{\mathrm{SR}} \sim \frac{2\left(1+\delta^2\right)}{\omega^2} + O\left(\omega^{-3}\right) \qquad \text{as } \omega \to \infty \tag{55}$$

and clearly does not match the exact result (53) for general δ. The mismatch of the exact spectrum and the SR approximation far from the peak can be clearly seen in Figure 2. As noted by SR, this effect occurs even when $\delta = 0$. We show in the following sections that the correct large-frequency asymptote depends upon the consistent inclusion of random fluctuations in the amplitude r of the oscillations.

4.1.2. Comparison with other approximations. The above-threshold laser regime has been the subject of many theoretical and experimental investigations. Here we compare our results for $|a| \gg 1$ with other approximation schemes yielding non-Lorentzian lineshapes.

The Langevin equations arising from (2) and (3) may be linearized near the deterministic orbit by the change of variables (32) to give (with $\Omega = 0$)

$$\frac{d\rho}{dt} = -2a\rho + \sqrt{2a}\xi_r, \tag{56}$$

$$\frac{d\theta}{dt} = -\sqrt{2}\delta\rho + \frac{1}{\sqrt{a}}\xi_\theta. \tag{57}$$

Here $\xi_r(t)$ and $\xi_\theta(t)$ are independent, unit intensity white noise terms; cf. the Fokker–Planck equation (35). Equation (56) is linear in ρ with a white-noise forcing, and so

generates an Ornstein–Uhlenbeck colored-noise process $\rho(t)$. Equation (57) describes the stochastic process of the phase angle $\theta(t)$, which is forced by the colored-noise process $\rho(t)$ and the white noise ξ_θ.

In the full description of the correlation function (5), fluctuations in both the phase $\theta(t)$ and the amplitude $r(t)$ of the oscillation have important effects. The linearized Langevin equations, however, suggest the use of a *phase oscillator* approximation [24], wherein the effects of the colored noise ρ are incorporated into the evolution of the phase angle θ, but all other influence of the amplitude fluctuations are neglected. Under this approximation, the correlation function (5) is simply

$$R_{\mathrm{P}O}(\tau) = a\,\langle \exp\left[i\,\theta(t) - i\,\theta(t+\tau)\right]\rangle \tag{58}$$

and can be calculated directly from the linearized Langevin equation for θ. The result is [22]

$$R_{\mathrm{P}O}(\tau) = a\,e^{-D(\tau)}, \tag{59}$$

where $D(\tau)$ is given in (41). Note that this corresponds to setting $a_0 = 1$ and $a_1 = a_2 = 0$ in the asymptotic result (43).

The spectrum of (59) is non-Lorentzian for $\delta > 0$, but the large-frequency asymptote is

$$F_{\mathrm{P}O} \sim \frac{2}{\omega^2} + O\left(\omega^{-3}\right) \qquad \text{as } \omega \to \infty, \tag{60}$$

i.e., a factor of 2 lower than the correct form (53). This discrepancy is due to the neglect of the direct effect of amplitude fluctuations upon the spectrum of E within the phase oscillator approximation; the correct values of a_0, a_1, and a_2 from (44) are required for the full description of the spectrum.

The *Voigt lineshape*, obtained by convoluting a Lorentzian and a Gaussian in frequency space, is sometimes fitted to experimental oscillation data [25, 26]. We note that in the phase-oscillator approximation the correlation function (59) is the exponential of the function $D(\tau)$; if this function is assumed to be replaced by the first two terms of its Taylor series about $\tau = 0$, the resulting correlation function

$$R_{\mathrm{V}}(\tau) = a\,e^{-\frac{1}{a}|\tau| - \delta^2\tau^2} \tag{61}$$

corresponds to a Voigt spectral lineshape. This result can also be derived directly from the linearized phase equation (57) if the amplitude deviation ρ is assumed to be a (frozen) random variable chosen from the Gaussian distribution (34). This approximation completely ignores the temporal variation of $\rho(t)$ as given in (56), and is valid only in the limit $\delta \to \infty$.

4.2. Below threshold: $a \ll -1$. When the pumping parameter a is negative, the origin $r = 0$ is the attracting steady solution of the deterministic system (2). In order to find asymptotic solutions valid for $a \ll -1$, we therefore rescale lengths to closely examine the neighborhood of the origin:

$$r = \frac{1}{\sqrt{|a|}}\rho, \tag{62}$$

where ρ is of order one and is restricted to the interval $[0, \infty)$. The two-dimensional area integral transforms as

$$\int_0^{2\pi}\int_0^{\infty} r\,dr\,d\theta \longrightarrow \frac{1}{|a|}\int_0^{2\pi}\int_0^{\infty} \rho\,d\rho\,d\theta, \tag{63}$$

and the stationary probability density is

$$P_\infty(\rho) = \frac{|a|}{2\pi} \exp\left(-\frac{\rho^2}{2}\right). \tag{64}$$

Retaining the most significant terms (and all δ-dependent terms) in (14) as $|a| \to \infty$ yields the asymptotic equation

$$\frac{\partial Q}{\partial \tau} + \frac{a}{\rho}\frac{\partial}{\partial \rho}\left(\rho^2 Q\right) + \left(a\delta + \frac{1}{a}\delta\rho^2\right)\frac{\partial Q}{\partial \theta} + \frac{a}{\rho}\frac{\partial}{\partial \rho}\left(\rho\frac{\partial Q}{\partial \rho}\right) + \frac{a}{\rho^2}\frac{\partial^2 Q}{\partial \theta^2} = 0, \tag{65}$$

with initial condition

$$Q(\tau = 0) = \frac{\sqrt{|a|}}{2\pi} e^{i\theta} \rho\, e^{-\frac{1}{2}\rho^2}. \tag{66}$$

Motivated by work in the mixing of scalar fields in vortex fluid flows [27, 28], we seek a solution of (65) in the form

$$Q = \frac{\sqrt{|a|}}{2\pi} e^{i\theta} \rho \exp\left[-g(\tau)\rho^2 + h(\tau)\right]. \tag{67}$$

This is an *exact* solution of the equation if g and h satisfy the ordinary differential equations

$$\begin{aligned} \frac{1}{|a|}\frac{dg}{d\tau} &= 2g - 4g^2 - i\frac{\delta}{a^2}, \\ \frac{1}{|a|}\frac{dh}{d\tau} &= -8g + 3 + i\delta, \end{aligned} \tag{68}$$

with initial conditions $g(0) = 1/2$, $h(0) = 0$. The solutions of this system may be obtained in closed form, and the correlation function resulting from the integration (19) is

(69)

$$\begin{aligned} R(\tau) &= \frac{1}{|a|}\int_0^\infty \rho^3 \exp\left[-g(\tau)\rho^2 + h(\tau)\right]\, d\rho \\ &= \frac{1}{2|a|g^2(\tau)} \exp\left[h(\tau)\right] \\ &= \frac{2\left(1-4\mu\right)^2}{|a|\left(1-2\mu\right)^4} \exp\left[\left(-1 + i\delta + 8\mu\right)|a\tau|\right] \left\{1 - \frac{4\mu^2}{(1-2\mu)^2}\exp\left[\left(-2+8\mu\right)|a\tau|\right]\right\}^{-2}, \end{aligned}$$

where $\mu = \left(1 - \sqrt{1 - 4i\delta/a^2}\right)/4$.

In contrast to the above-threshold case in section 4.1, an analytical formula for the spectrum corresponding to this correlation function has not been found. However, numerical integration of (20) using (69) can be used to find the asymptotic spectrum for $a \ll -1$, valid for arbitrarily large values of δ. This spectrum is plotted for $a = -3$ as a dashed line in part (a) of Figures 1, 3, and 4. The requirement $a \ll -1$ for validity of the asymptotics does not really hold at the chosen value of a, and the fit to the numerical spectrum is poor near the peak, though better than the SR lineshape away from the peak. However, for $a = -10$, Figure 8 shows an excellent match between the asymptotic spectrum and the numerics—here we have taken $\delta = 200$ (so that $\delta/a^2 = 2$) in order to highlight the non-Lorentzian lineshape. Note that the non-Lorentzian effects in (69) depend on the parameter combination δ/a^2; this leads us to define the border $\delta/a^2 = 1/2$ for negative a in Figure 5.

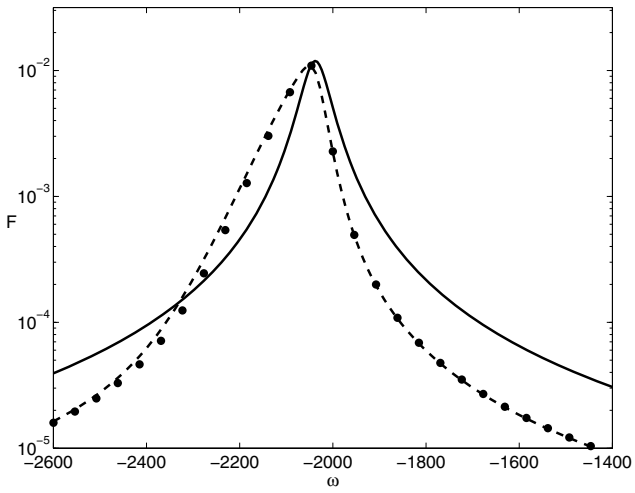


FIG. 8. *Spectrum $F(\omega)$ for $a = -10$ and $\delta = 200$. See Figure 1 for description of line types.*

5. Conclusions. We have studied the effects of additive white noise on the spectral lineshape of a dynamical system described by the Hopf normal form (1). Particular emphasis is laid on the role of the amplitude-phase coupling (detuning) parameter δ. Non-Lorentzian lineshapes occur when δ is nonnegligible, in contrast to the heretofore most studied case of $\delta = 0$.

We use asymptotic methods to find closed form solutions for the laser amplitude correlation function, valid for any δ value, when the pump parameter a is both above (equation (43)) and below (equation (69)) threshold. In the former case an analytic formula for the spectrum is also given. The asymptotics are formally valid for $|a| \gg 1$, but we find they match the numerical solution well even when the magnitude of a is as small as 5. In contrast to previous work on this problem, no assumption on the size of δ is made, so that the formulas are valid for arbitrarily large δ values. The spectrum becomes non-Lorentzian as δ increases, and we indicate (Figure 5) the parameter regions where single-eigenvalue methods fail to accurately describe the lineshape. We look at existing approximation methods (section 4.1.2) and show that they are special limiting cases of our results.

Both asymptotic and numerical solutions are based on the auxiliary function approach developed in section 2. It should be stressed that (despite the use of convenient nomenclature from the laser physics literature) the results presented are quite general and apply to any noisy dynamical system near a Hopf bifurcation point. We anticipate that the results and methods presented here will be of interest both for their applicability to high-δ lasers, and as test cases for single-eigenvalue methods developed to predict spectra in high-dimensional oscillators for electronic and chemical applications [5, 8, 20]. It would also be interesting to apply the asymptotic methods used here to more detailed models of lasers which supplement (1) with equations for the laser gain and carrier density, leading to relaxation oscillation sidebands in the spectrum [29].

Appendix. The Hopf normal form equations (2) contain only three parameters: a, Ω, and δ. In this appendix we describe in detail the derivation of (2), beginning

with the standard deterministic Hopf normal form presented in [1]:

$$\begin{aligned} \frac{d\tilde{r}}{d\tilde{t}} &= \left(d\mu + \tilde{a}\tilde{r}^2\right)\tilde{r}, \\ \frac{d\theta}{d\tilde{t}} &= \omega + c\mu + b\tilde{r}^2. \end{aligned} \tag{70}$$

(Here we use tildes to distinguish between certain parameters and dimensionless versions of the same name.) The parameters $\tilde{a}$, b, c, and d are all typically of order unity, with $\tilde{a} < 0$, and can in theory be found from the properties of the dynamical system at the bifurcation point $\mu = 0$. The bifurcation parameter is μ, and for analysis "near" the bifurcation point, the magnitude of μ is assumed to be small: $|\mu| \ll 1$. The differential equation for the evolution of the complex quantity $\tilde{E} = \tilde{r}\exp(i\theta)$ follows from (70); the effect of random fluctuations is modeled by adding a white noise term $\tilde{\Gamma}$:

$$\frac{d\tilde{E}}{d\tilde{t}} = i\left(\omega + c\mu\right)\tilde{E} + \left(d\mu + (\tilde{a} + ib)\left|\tilde{E}\right|^2\right)\tilde{E} + \tilde{\Gamma}, \tag{71}$$

with

$$\tilde{\Gamma} = \tilde{\xi}_x(t) + i\tilde{\xi}_y(t) \tag{72}$$

being a complex noise of intensity κ:

$$\left\langle \tilde{\xi}_x(t)\tilde{\xi}_x(t')\right\rangle = \left\langle \tilde{\xi}_y(t)\tilde{\xi}_y(t')\right\rangle = 2\kappa\delta(t-t'), \qquad \left\langle \tilde{\xi}_x(t)\tilde{\xi}_y(t')\right\rangle = 0. \tag{73}$$

As further discussed below, the balance between the noise intensity κ and the (small-magnitude) bifurcation parameter μ is crucial to understanding the effects of noise upon the dynamics.

To nondimensionalize the system, we write the Fokker–Planck equation for (71) in the polar coordinates $\tilde{r}$ and θ:

$$\frac{\partial P}{\partial \tilde{\tau}} + \frac{1}{\tilde{r}}\frac{\partial}{\partial \tilde{r}}\left[\left(d\mu + \tilde{a}\tilde{r}^2\right)\tilde{r}^2 P\right] + \frac{\partial}{\partial\theta}\left[\left(\omega + c\mu + b\tilde{r}^2\right)P\right] - \frac{\kappa}{\tilde{r}}\frac{\partial}{\partial\tilde{r}}\left[\tilde{r}\frac{\partial P}{\partial\tilde{r}}\right] - \frac{\kappa}{\tilde{r}^2}\frac{\partial^2 P}{\partial\theta^2} = 0. \tag{74}$$

Choosing a length scale L and a time scale T as references, $\tilde{r}$ and $\tilde{\tau}$ may be written in terms of the dimensionless variables r and τ used in (2):

$$\tilde{r} = Lr, \qquad \tilde{\tau} = T\tau. \tag{75}$$

The Fokker–Planck equation (74) then becomes

$$\begin{aligned} &\frac{\partial P}{\partial \tau} + \frac{1}{r}\frac{\partial}{\partial r}\left[\left(Td\mu + TL^2\tilde{a}r^2\right)r^2 P\right] + \frac{\partial}{\partial\theta}\left[\left(T\omega + Tc\mu + TL^2 br^2\right)P\right] - \frac{\kappa T}{L^2}\frac{1}{r}\frac{\partial}{\partial r}\left[r\frac{\partial P}{\partial r}\right] \\ &\quad - \frac{\kappa T}{L^2}\frac{1}{r^2}\frac{\partial^2 P}{\partial\theta^2} = 0. \end{aligned} \tag{76}$$

We can now choose the length and time scales L and T in order to reduce the number of dimensionless parameters. The conventional choice in the laser literature (see, e.g.,

Chapter 12 of [2]) is to set the diffusion coefficients $\kappa T/L^2$ to unity, and the parameter combination $TL^2\tilde{a}$ to -1 (recall that $\tilde{a}$ is negative). These choices determine T and L in terms of the parameter $\tilde{a}$ and the noise intensity κ, and yield the nondimensional Fokker–Planck equation:

(77)
$$\frac{\partial P}{\partial \tau} + \frac{1}{r}\frac{\partial}{\partial r}\left[\left(a - r^2\right) r^2 P\right] + \frac{\partial}{\partial \theta}\left[\left(\Omega + \delta\left(a - r^2\right)\right) P\right] - \frac{1}{r}\frac{\partial}{\partial r}\left[r\frac{\partial P}{\partial r}\right] - \frac{1}{r^2}\frac{\partial^2 P}{\partial \theta^2} = 0,$$

as used in (6). The parameters a, Ω, and δ appearing here are thus given in terms of the original system parameters as

(78)
$$a = d\mu\sqrt{\frac{-1}{\kappa\tilde{a}}}, \quad \Omega = \sqrt{\frac{-1}{\kappa\tilde{a}}}\left(\omega + c\mu - \frac{bd\mu}{\tilde{a}}\right), \quad \delta = \frac{b}{\tilde{a}}.$$

Note the appearance of the bifurcation parameter μ and the noise intensity κ in the definition of the parameter a. As stated above, the magnitude of μ must be small, $|\mu| \ll 1$, for analysis near the Hopf bifurcation point. However, the noise intensity κ may also be very small, and the resulting parameter a may therefore have arbitrarily large magnitude even for small $|\mu|$ if the noise intensity κ goes to zero. Note also that the amplitude-phase coupling parameter δ is independent of μ and depends only on the properties of the dynamical system at the bifurcation point (as determined by the parameters b and $\tilde{a}$ in the normal form (70)).

The derivation of the normal form equations at a bifurcation point relies on the transformation of the dependent and/or independent variables in the original system. The assumption made above that the noise terms enter additively into (71) is therefore somewhat simplistic—when the noise in the original dynamical system is subject to the normal form transformation, it may very well appear in (71) as, for example, multiplicative noise [13], or with unequal intensities in the real and imaginary directions. Nevertheless, the generality of the normal form is such that the study of the simplest case, i.e., additive isotropic white noise, is still instructive. Moreover, additive isotropic noise is often adopted as a phenomenological model when the details of the noise sources or the underlying dynamical system are not known; this is the case for the semiclassical laser equations as derived in Chapter 12 of [2], for example. It is hoped that the new results of this paper (especially on the importance of the amplitude-phase coupling parameter δ) for additive isotropic noise will stimulate further work on the combined effects on spectral lineshapes with nonzero δ and multiplicative and/or nonisotropic noise.

Acknowledgments. The authors gratefully acknowledge helpful discussions with Dr. Guillaume Huyet and Prof. M. P. Kennedy.

REFERENCES

[1] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, New York, Berlin, 1989.
[2] H. Risken, *The Fokker–Planck Equation*, 2nd ed., Springer, New York, 1989.
[3] K. Wiesenfeld, *Period doubling bifurcations: What good are they?*, in Noise in Nonlinear Dynamical Systems, vol. 2, F. Moss and P. V. E. McClintock, eds., Cambridge University Press, Cambridge, UK, 1989, pp. 145–175.
[4] K. Wiesenfeld, *Noisy precursors of nonlinear instabilities*, J. Statist. Phys., 38 (1985), pp. 1071–1097.

[5] A. Demir, A. Mehrotra, and J. Roychowdhury, *Phase noise in oscillators: A unifying theory and numerical methods for characterization*, IEEE Trans. Circuits Systems I, 47 (2000), pp. 655–674.

[6] F. X. Kaertner, *Analysis of white and $f^{-\alpha}$ noise in oscillators*, Int. J. Circuit Theory Appl., 18 (1990), pp. 485–519.

[7] G. J. Coram, *A simple 2-D oscillator to determine the correct decomposition of perturbations into amplitude and phase noise*, IEEE Trans. Circuits Systems I, 48 (2001), pp. 896–898.

[8] W. Vance and J. Ross, *Fluctuations near limit cycles in chemical reaction systems*, J. Chem. Phys., 105 (1996), pp. 479–487.

[9] A. Hajimiri and T. H. Lee, *A general theory of phase noise in electrical oscillators*, IEEE J. Solid-State Circuits, 33 (1998), pp. 179–194.

[10] C. H. Henry, *Theory of the linewidth of semiconductor lasers*, IEEE J. Quantum Elec., QE18 (1982), pp. 259–264.

[11] F. Baras, M. M. Mansour, and C. Van den Broeck, *Asymptotic properties of coupled nonlinear Langevin equations in the limit of weak noise.* II: *Transition to a limit cycle*, J. Statist. Phys., 28 (1982), pp. 577–587.

[12] F. Baras and M. M. Mansour, *Microscopic simulations of chemical instabilities*, Adv. Chem. Phys., 100 (1997), pp. 393–474.

[13] R. Graham, *Hopf bifurcation with fluctuating control parameter*, Phys. Rev. A., 25 (1982), pp. 3234–3258.

[14] K. Mallick and P. Marcq, *Stability analysis of a noise-induced Hopf bifurcation*, Euro. Phys. J. B, 36 (2003), pp. 119–128.

[15] V. Altares and G. Nicolis, *Stochastically forced Hopf bifurcation: Approximate Fokker–Planck equation in the limit of short correlation times*, Phys. Rev. A., 37 (1988), pp. 3630–3633.

[16] J. Olarrea and F. J. de al Rubia, *Stochastic Hopf bifurcation: The effect of colored noise on the bifurcation interval*, Phys. Rev. E., 53 (1996), pp. 268–271.

[17] A. Longtin, *Noise-induced transitions at a Hopf bifurcation in a 1st order delay-differential equation*, Phys. Rev. A, 44 (1991), pp. 4801–4813.

[18] K. Seybold and H. Risken, *On the theory of a detuned single mode laser near threshold*, Z. Physik, 267 (1974), pp. 323–330.

[19] C. Birocheau, Z. Toffano, and A. Destrez, *Linewidth evolution in semiconductor lasers throughout threshold*, Ann. Télécommun., 49 (1994), pp. 607–618.

[20] P. Gaspard, *Trace formula for noisy flows*, J. Statist. Phys., 106 (2002), pp. 57–96.

[21] M. I. Dykman, R. Mannella, P. V. E. McClintock, S. M. Soskin, and N. G. Stocks, *Noise-induced spectral narrowing in nonlinear oscillators*, Europhys. Lett., 13 (1990), pp. 691–696.

[22] S. N. Dixit, P. Zoller, and P. Lambropoulos, *Non-Lorentzian laser line shapes and the reversed peak asymmetry in double optical resonance*, Phys. Rev. A, 21 (1980), pp. 1289–1296.

[23] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[24] A. Pikovsky, M. Rosenblum, and J. Kurths, *Synchronization*, Cambridge University Press, Cambridge, UK, 2001.

[25] S. Viciani, M. Gabrysch, F. Marin, F. Monti di Sopra, M. Moser, and K. Heinz Gulden, *Lineshape of a vertical cavity surface emitting laser*, Optics Commun., 206 (2002), pp. 89–97.

[26] F. Herzel, *An analytical model for the power spectral density of a voltage-controlled oscillator and its analogy to the laser linewidth theory*, IEEE Trans. Circuits Systems I, 45 (1998), pp. 904–908.

[27] J. P. Gleeson, O. M. Roche, J. West, and A. Gelb, *Modelling annular micromixers*, SIAM J. Appl. Math., 64 (2004), pp. 1294–1310.

[28] K. Bajer, A. P. Bassom, and A. D. Gilbert, *Accelerated diffusion in the centre of a vortex*, J. Fluid. Mech., 437 (2001), pp. 395–411.

[29] M. P. van Exter, W. A. Hamel, J. P. Woerdman, and B. R. P. Zeijlmans, *Spectral signature of relaxation oscillations in semiconductor lasers*, 28 (1992), pp. 1470–1478.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1689–1702

PHASE TRANSITIONS AND CHANGE OF TYPE IN LOW-TEMPERATURE HEAT PROPAGATION*

KATARZYNA SAXTON† AND RALPH SAXTON‡

Abstract. Classical heat pulse experiments have shown heat to propagate in waves through crystalline materials at temperatures close to absolute zero. With increasing temperature, these waves slow down and finally disappear, to be replaced by diffusive heat propagation. Several features surrounding this phenomenon are examined in this work. The model used switches between an internal parameter (or extended thermodynamics) description and a classical (linear or nonlinear) Fourier law setting. This leads to a hyperbolic-parabolic change of type, which allows wavelike features to appear beneath the transition temperature and diffusion above. We examine the region around and immediately below the transition temperature, where dissipative effects are insignificant.

Key words. phase transition, hyperbolic-parabolic, change of type

AMS subject classifications. 35M10, 35Q72, 35R35

DOI. 10.1137/040612208

1. Introduction. The analysis in this paper is based on a low-temperature heat propagation model described in [9] and [10]. The model is based on experimental results of [3], [2], [5], and [6], which provide evidence of second sound, i.e., hyperbolic, or wavelike, thermal effects where Fourier's law fails, in very pure crystals of sodium fluoride and bismuth.

Significantly, these features appear only at certain temperatures below which the materials reach their peak thermal conductivities (at approximately 18.5 K and 4.5 K for NaF and Bi, respectively). No wavelike behavior is found in NaF and Bi at higher temperatures, where only diffusive heat propagation is observed. Further, the speed, U_E, at which small amplitude thermal waves propagate is a decreasing function of temperature in the region where the waves can be detected, after which the diffusion process dominates. This hyperbolic region appears separated from the diffusive region by a "critical" temperature, ϑ_λ, at which $U_E = 0$ [1]. The aim of this paper is to understand the dynamics of regular solutions having temperatures close to that of the phase transition. We begin, in section 2, by describing a phenomenological one-dimensional model which uses an internal variable behaving as an order parameter. In section 3, we will examine properties of the phase transition, and in section 4, we obtain conditions under which this class of solutions remain smooth. Some explicit cases are, finally, examined in section 5.

2. Preliminaries. We briefly describe our model and refer to [10] (see also [9]) for further details concerning the thermodynamics of materials with internal parameters. In the present context, two forms of heat transmission—diffusive propagation at high temperatures and wavelike propagation at low temperatures—are separated by a phase transition at a critical temperature, $\vartheta_\lambda > 0$. At temperatures above ϑ_λ, we

*Received by the editors July 24, 2004; accepted for publication (in revised form) March 2, 2006; published electronically July 17, 2006.
http://www.siam.org/journals/siap/66-5/61220.html
†Department of Mathematics and Computer Science, Loyola University, New Orleans, LA 70118 (saxton@loyno.edu). The work of this author was partially supported by NSF grant DMS-0104508.
‡Department of Mathematics, University of New Orleans, New Orleans, LA 70148 (rsaxton@math.uno.edu). The work of this author was partially supported by NSF grant DMS-0104489.

employ the equations for heat flow through a one-dimensional rigid solid, consisting of balance of energy and Fourier's law,

$$\varepsilon(\vartheta)_t + q_x = 0, \tag{2.1}$$

$$q = -k(\vartheta)\vartheta_x, \tag{2.2}$$

where $\varepsilon(\vartheta)$ represents internal energy, $\varepsilon'(\vartheta) = c_v(\vartheta)$ is the specific heat at constant volume, q denotes heat flux, and $k(\vartheta)$ is the heat conductivity.

At temperatures between $\vartheta = 0$ (absolute) and $\vartheta = \vartheta_\lambda$, experimental results indicate that the constitutive description of q given by (2.2) is inadequate [2]. This then requires the use of an extended form of thermodynamics, in which we employ an internal parameter, p, to appropriately model observations. The internal parameter satisfies a particular form of evolution equation (see [10], [9]), and heat flow is described below ϑ_λ by the equations

$$\varepsilon(\vartheta)_t + q_x = 0, \tag{2.3}$$

$$p_t = g_1(\vartheta)\vartheta_x + g_2(\vartheta)p, \tag{2.4}$$

$$q = -\alpha(\vartheta)p, \tag{2.5}$$

where $g_1(\vartheta) \geq 0 \geq g_2(\vartheta)$ and $\alpha(\vartheta) \geq 0$ are material functions. The second law of thermodynamics imposes the restriction that $\alpha(\vartheta) = \psi_{20}\vartheta^2 g_1(\vartheta)$, where the constant $\psi_{20} > 0$ comes from the Helmholtz free energy, ψ, which has the form $\psi = \psi_1(\vartheta) + \frac{1}{2}\psi_{20}\vartheta p^2$. While satisfying the second law of thermodynamics, the model has internal energy depending only on temperature [9]. In effect, (2.3)–(2.5) permit q to depend on the history of the temperature gradient.

The following constitutive relations will be used for g_1 and g_2 (see [10]):

$$g_1(\vartheta) = g_{10}(\vartheta)(\vartheta_\lambda - \vartheta)^r, \qquad g_{10}(\vartheta) > 0, \tag{2.6}$$

$$g_2(\vartheta) = g_{20}(\vartheta)(\vartheta_\lambda - \vartheta)^{2r}, \qquad g_{20}(\vartheta) < 0, \tag{2.7}$$

where $0 \leq \vartheta \leq \vartheta_\lambda$. Here $g_{10}, g_{20} \in C[0, \vartheta_\lambda]$ and $r \in (0, 1)$. The form of these functions can be derived from experimental data on the wave speed, $U_E(\vartheta)$, the heat conductivity, $K(\vartheta)$, and the specific heat, $c_v(\vartheta)$, as we now describe.

The characteristic equation for (2.3)–(2.5) is given by

$$c_v(\vartheta)\lambda^2 + \lambda\alpha'(\vartheta)p - \alpha(\vartheta)g_1(\vartheta) = 0. \tag{2.8}$$

If one considers waves propagating into an undisturbed state $\vartheta =$ constant, $p = 0$ ($q = 0$), this provides an expression for U_E,

$$\lambda^2 = U_E^2 \equiv \frac{\alpha(\vartheta)g_1(\vartheta)}{c_v(\vartheta)} = \psi_{20}\vartheta^2\frac{g_1^2(\vartheta)}{c_v(\vartheta)}. \tag{2.9}$$

Given experimental measurements of $U_E(\vartheta)$ and $c_v(\vartheta)$, this specifies $g_1(\vartheta)$. It is observed that second sound is a decreasing function of temperature, which we allow to reach zero [1] at $\vartheta = \vartheta_\lambda$ (cf. (2.6)).

In order to find $g_2(\vartheta)$, we use measurements of heat conductivity made in near-stationary states, for which fast processes are considered to be minor, $p_t \approx 0$.[1] In this case, the difference between solutions to (2.3)–(2.5) and the diffusion equation

$$\varepsilon(\vartheta)_t - (K(\vartheta)\vartheta_x)_x = 0, \tag{2.10}$$

where $K(\vartheta) = -\frac{\alpha(\vartheta)g_1(\vartheta)}{g_2(\vartheta)} = -\frac{U_E^2(\vartheta)c_v(\vartheta)}{g_2(\vartheta)} > 0$, becomes small. Requiring $K(\vartheta)$ to remain finite as $\vartheta \to \vartheta_\lambda -$ then leads to the form of $g_2(\vartheta)$ in (2.7).

The behavior of the specific heats of Bi and NaF is typically considered to be continuous in temperature, and we will assume this here, with

$$c_v(\vartheta) \sim c_\lambda, \qquad |\vartheta_\lambda - \vartheta| \ll 1, \qquad c_\lambda > 0, \tag{2.11}$$

where "$\sim$" denotes leading order behavior. Otherwise c_v is considered to be described by a continuous function which obeys Debye's law, $c_v(\vartheta) \sim \vartheta^3$, as $\vartheta \to 0$.

Particular forms of (2.6) and (2.7) chosen to fit available data for crystals of high purity NaF and Bi can be found in [10]. Together with (2.11), these result in U_E having the general form

$$U_E^2(\vartheta) \sim U_0^2(\vartheta)(\vartheta_\lambda - \vartheta)^{2r} \tag{2.12}$$

for $\vartheta \le \vartheta_\lambda$, with U_0, a continuous function of ϑ, found experimentally.

For convenience, we introduce the following change of variables:

$$e = \varepsilon(\vartheta) - \varepsilon(\vartheta_\lambda), \quad \vartheta = \varepsilon^{-1}(e + \varepsilon_\lambda), \quad \text{and } \varepsilon_\lambda = \varepsilon(\vartheta_\lambda). \tag{2.13}$$

We may rewrite (2.1), (2.2) in terms of $e > 0$ and q as

$$e_t + q_x = 0, \tag{2.14}$$

$$q = -d(e)e_x, \tag{2.15}$$

where $d(e) = k(\vartheta)/c_v(\vartheta),\ k(\vartheta) > 0$. When $e < 0$, equations (2.3), (2.4) become

$$e_t + q_x = 0, \tag{2.16}$$

$$q_t + \frac{h'(e)}{h(e)} q q_x = f(e)\{q + D(e)e_x\}, \tag{2.17}$$

where[2]

$$f(e) = g_2(\vartheta), \quad h(e) = \alpha(\vartheta), \quad D(e) = \frac{K(\vartheta)}{c_v(\vartheta)}. \tag{2.18}$$

[1] In this limit, however, thermal waves do not necessarily propagate at the speed dictated by the qualitatively approximate parabolic equation (2.10), but still at a characteristic velocity, λ, given by (2.9). We also distinguish $p_t \approx 0$ (for which $g_1(\vartheta), g_2(\vartheta) \neq 0$) from $p_t = 0$ (where $g_1(\vartheta) = g_2(\vartheta) = 0$). The former is an assumption concerning the dynamics, such as the time asymptotic behavior which may arise due to damping (for instance, as seen in [4], and which still preserves hyperbolic features such as finite speed of propagation) in the original system. In the latter case, ϑ is assumed to have reached the transition temperature, $\vartheta = \vartheta_\lambda$.

[2] For physical reasons [2], we will here assume that $k(\vartheta_\lambda) = K(\vartheta_\lambda)$ (where $q = -K(\vartheta)\vartheta_x = -D(e)e_x$), which occurs in the limit $p_t \approx 0$ leading to (2.10). $K(\vartheta)$ is sometimes known as the *quasistatic* heat conductivity.

Continuous initial data, $e_0(x)$ or $\vartheta_0(x)$, will be defined such that

$$e(\vartheta(x,0)) = e_0(x), \qquad e_0(0) = 0, \tag{2.19}$$

with

$$xe_0(x) > 0 \quad \text{for } x \neq 0, \tag{2.20}$$

where

$$\vartheta(x,0) = \vartheta_0(x), \qquad \vartheta_0(0) = \vartheta_\lambda. \tag{2.21}$$

Given the observed sharp decay in the speed of heat pulse propagation, nonlinear effects are involved. Since the system (2.16), (2.17) is quasilinear and hyperbolic (cf. (2.3)–(2.5)), it is possible to account for this decay, but it also becomes possible for shocks to form in finite times [7], [8], [10] in temperatures below ϑ_λ. The present analysis examines the situation under which solutions taking values at temperatures on both sides of ϑ_λ should remain smooth. As such, it can be regarded as a first step in analyzing experiments using large amplitude temperature pulses crossing into phases that involve dissipation.

3. Properties of phase transitions. Let Γ denote a curve $x = \varphi(t),\ t \geq 0$, such that $e(\varphi(t),t) = 0$ ($\vartheta(\varphi(t),t) = \vartheta_\lambda$) and let $V \subset \mathbb{R}^2_+ = \{(x,t) \in \mathbb{R}^2, t \geq 0\}$ be a neighborhood of Γ. Set $V = U_- \cup \Gamma \cup U_+$, where the regions U_- and U_+ correspond to $e < 0$ and $e > 0$, respectively. Heat propagation is then governed by (2.14), (2.15) in U_- and by (2.16), (2.17) in U_+.

We denote limits of a function u from the left and right of Γ, as $x \to \varphi(t)$, by $u^-(t) = u(\varphi(t)-,t)$ and $u^+(t) = u(\varphi(t)+,t)$, respectively, and denote the jump of u across Γ by $[u](t) = u^+(t) - u^-(t)$. Let $PC^1(Q)$ denote the class of piecewise differentiable functions on $Q \subset \mathbb{R}^2$. By assuming $\vartheta \in PC^1(\mathbb{R}^2_+)$, e becomes continuous across Γ.

Let $s = \dot{\varphi}$. Using (2.14) and (2.16) together with the jump condition across Γ,

$$-s[e] + [q] = 0, \tag{3.1}$$

demonstrates that q is continuous across Γ. This allows us to define

$$q_\flat(t) \equiv q(\varphi(t),t) = -\lim_{x \to \varphi(t)+} (d(e)e_x) = -k(\vartheta_\lambda)\vartheta_x^+(t). \tag{3.2}$$

In the following, where we wish to categorize s and to obtain a relationship between $\vartheta_x^+(t)$ and $\vartheta_x^-(t)$, ϑ and q are considered to be smooth in $V\backslash\Gamma$ (at least locally). For convenience, we next set $\psi_{20} = 1$.

LEMMA 3.1. *Let $\vartheta_x^+ > 0$. Then $s\vartheta_x^- = 0$, and $q \in C^1(V)$ if $s = 0$.*

Proof. Let us write (2.17) in the form

$$qq_x = \frac{h(e)}{h'(e)}(f(e)\{q + D(e)e_x\} - q_t). \tag{3.3}$$

Using (2.6)–(2.13), (2.18) in U_- for $e \sim 0$ ($\vartheta \sim \vartheta_\lambda$) gives

$$c_\upsilon(\vartheta) \sim c_\lambda, \tag{3.4}$$

$$h(e) \sim \vartheta_\lambda^2 g_{10}(\vartheta_\lambda)\left(\frac{1}{c_\lambda}|e|\right)^r, \tag{3.5}$$

$$f(e) \sim g_{20}(\vartheta_\lambda)\left(\frac{1}{c_\lambda}|e|\right)^{2r}, \tag{3.6}$$

and

$$D(e) \sim -\frac{(\vartheta_\lambda g_{10}(\vartheta_\lambda))^2}{c_\lambda g_{20}(\vartheta_\lambda)}. \tag{3.7}$$

Thus, since $D(e)e_x = D(e)c_v(\vartheta)\vartheta_x$, (3.3) implies

$$qq_x \sim -\frac{1}{r}eq_t \to 0 \quad \text{as } e \to 0^-. \tag{3.8}$$

Using (3.2) for $\vartheta_x^+ > 0$ implies that $q_x^- = 0$ and consequently $e_t^- = 0$, by (2.16). The definition of Γ implies that $e_t^- + se_x^- = 0$, and so $se_x^- = 0$, whence $s\vartheta_x^- = 0$.

Finally, if $s = 0$, the definition of Γ also implies that $e_t^+ = 0$, in which case (2.14) implies $q_x^+ = 0$.

Assuming that solutions depend continuously on the initial data (2.21) locally in time in V, we now obtain the following.

COROLLARY 3.2. *Let $\vartheta_0'^{\pm} = \lim_{x\to 0\pm}\vartheta_0'(x) > 0$. Then there exists $\tau > 0$ such that $s = 0$ for $t \in (0,\tau)$.*

Proof. This follows immediately from the lemma, since $\vartheta_x^-(t), \vartheta_x^+(t) > 0$, through continuous dependence, over some interval $t \in (0,\tau)$.

Remark. Since $\varphi(0) = 0$ from (2.19), in this case $\varphi(t) \equiv 0$ for all $t \in (0,\tau)$.

The next results provide a connection between the left and right states of Γ and show that the condition $s = 0$ is controlled only by initial data corresponding to these states and by the solution to the diffusion equation, (2.14), (2.15), and (2.19).

LEMMA 3.3. *Let $\vartheta_0'^-, \vartheta_0'^+ > 0$ and $s = 0$. Then $\vartheta_x^+(t) > 0 \Rightarrow \vartheta_x^-(t) > 0$, $t \in (0,\tau)$.*

Proof. Let $\gamma_\alpha \subset U_-$ denote the line segment $x = \alpha < 0$, $t \in (0,\tau)$, which lies parallel to Γ. For α sufficiently small, $q|_{\gamma_\alpha}$ can be bounded above, strictly, by 0. This can be seen by defining a curve $\gamma_\beta \in U_- \cup \Gamma$ as follows,

$$\gamma_\beta = \left\{(x,t(x)), x \le 0 : \frac{dt}{dx} = \frac{h(e)}{h'(e)q},\ t(0) = \beta \ge 0\right\}, \tag{3.9}$$

and by writing (2.17) in the form

$$\frac{1}{2}\frac{dq^2}{dx}\bigg|_{\gamma_\beta} = \frac{h(e)}{h'(e)}f(e)\{q + D(e)e_x\} \equiv \frac{h}{h'}\mathcal{M}. \tag{3.10}$$

Using (2.7), (2.18), and (3.5), we find that

$$\frac{h(e)}{h'(e)}|_{\gamma_\alpha} \sim \frac{e}{r} \tag{3.11}$$

for $\alpha \sim 0$.

Also, $D(e)e_x = K(\vartheta)\vartheta_x|_{\gamma_\alpha} \to k(\vartheta_\lambda)\vartheta_x^-$ and $q_{|_{\gamma_\alpha}} \to q_\flat = -k(\vartheta_\lambda)\vartheta_x^+$ as $\alpha \to 0$. Since $e_{|_{\gamma_\alpha}} \to 0$ as $\alpha \to 0$, and through (3.6), it follows that as $\alpha \to 0$,

$$\mathcal{M}_{|_{\gamma_\alpha}} \sim g_{20}(\vartheta_\lambda)k(\vartheta_\lambda)\left(\frac{1}{c_\lambda}\right)^{2r}(\vartheta_x^- - \vartheta_x^+)(|e|^{2r})_{|_{\gamma_\alpha}} \sim 0. \tag{3.12}$$

As a result, by (3.10),

$$q(x,t)_{|_{\gamma_\beta}} \sim q_\flat(\beta), \tag{3.13}$$

where

$$\frac{dt}{dx}{}_{|_{\gamma_\beta}} \sim \frac{e}{rq}\,. \tag{3.14}$$

Next, the use of (2.16), (2.17), and (3.12) shows that

$$\left(q_t - \frac{r}{e} q e_t\right)\Big|_{\gamma_\alpha} \sim 0, \tag{3.15}$$

from which a rearrangement and integration give

$$q(\alpha,t) \sim q(\alpha,0)\left(\frac{e(\alpha,t)}{e_0(\alpha)}\right)^r. \tag{3.16}$$

Since $\vartheta_0'^+ > 0$ and $e_0(\alpha) < 0$, we have, locally for $x < 0$ and $t > 0$, that $e < 0$ and $q \sim q_\flat \sim -k(\vartheta_\lambda)\vartheta_x^+ < 0$. Consequently $\frac{dt}{dx}{}_{|_{\gamma_\beta}} > 0$ by (3.14), and hence there exists some $\beta > 0$ such that $q(\alpha,0) \sim q_\flat(\beta)$, with $\beta \to 0+$ as $\alpha \to 0-$. Therefore (3.16) implies

$$q_\flat(t) = q_\flat(0) \lim_{\alpha\to 0}\left(\frac{e(\alpha,t)}{e_0(\alpha)}\right)^r. \tag{3.17}$$

Given $\varepsilon'(\vartheta) = c_v(\vartheta)$ and (2.11), (2.13), and (3.4), we have

$$\lim_{\alpha\to 0}\frac{e(\alpha,t)}{e_0(\alpha)} = \lim_{\alpha\to 0}\frac{\vartheta_\lambda - \vartheta(\alpha,t)}{\vartheta_\lambda - \vartheta_0(\alpha)} = \frac{\vartheta_x^-(t)}{\vartheta_0'^-}. \tag{3.18}$$

Finally, since $q_\flat(t)/q_\flat(0) = \vartheta_x^+(t)/\vartheta_0'^+$, this implies

$$\frac{\vartheta_x^-(t)}{\vartheta_0'^-} = \left(\frac{\vartheta_x^+(t)}{\vartheta_0'^+}\right)^{1/r}, \tag{3.19}$$

which leads to the desired conclusion.

THEOREM 3.4. *Let* $\vartheta_0'^-, \vartheta_0'^+ > 0$, *and* $\vartheta_x^+(t) > 0$ *for* $t \in (0,T)$, *where* $T > 0$. *Then* $s = 0$ *for* $t \in (0,T)$.

Proof. Suppose that $t^\sharp < T$ denotes the first time for which $\vartheta_x^-(t) = 0$, so that $t^\sharp > 0$ by Corollary 3.1. Using Lemma 3.1 and continuity in t implies $s(t^\sharp) = 0$, and thus Lemma 3.2 holds for $t \in (0,t^\sharp)$. Using continuity once again then implies that $\vartheta_x^+(t^\sharp) = 0$, which violates the hypothesis $\vartheta_x^+(t) > 0$ for $t \in (0,T)$. The result follows through contradiction.

4. Smooth solutions in the transcritical region. Next we set up the problem of phase transition in the region $U_- \cup \Gamma \cup U_+$, where we now extend U_+ to all of $\mathbb{R}^2_{++} = \{(x,t) \in \mathbb{R}^2, t \geq 0,\, x > \varphi(t)\}$ and restrict U_- to temperatures close to ϑ_λ. Assuming $\vartheta_0'^+, \vartheta_0'^- > 0$ means that we may set $\varphi(t) = 0$ for $t > 0$, given $\varphi(0) = 0$, from the results of the previous section.

The transcritical phase transition problem reduces to solving

$$e_t - (d(e)e_x)_x = 0 \tag{4.1}$$

in U_+, by (2.14) (2.15). In U_-, equations (2.16), (2.17) become

$$e_t + q_x = 0, \tag{4.2}$$

$$q_t + \frac{rq}{e} q_x = 0, \tag{4.3}$$

where we have used (3.11) to obtain (4.3). We recall that across Γ, where now

$$e(0,t) = 0, \quad t > 0, \tag{4.4}$$

both e and q are continuous. Initial data for (4.1)–(4.3) are given by

$$e(x,0) = e_0(x), \quad \text{with } xe_0(x) > 0 \text{ for } x \neq 0, \qquad \text{and } e_0(0) = 0. \tag{4.5}$$

In U_+, $e(x,t)$ is a solution to (4.1) satisfying the Dirichlet boundary condition (4.4), together with (4.5). This solution determines $q_\flat(t)$,

$$q_\flat(t) = q(0,t) = -\lim_{x \to 0+} (d(e(x,t))e_x(x,t)) = -k(\vartheta_\lambda)\vartheta_x^+(t), \tag{4.6}$$

as a function of the initial data $e_0(x)$, for $x > 0$.

In U_-, the pair $(e(x,t), q(x,t))$ is a solution to (4.2), (4.3) satisfying (4.4), (4.5), and (4.6).

Eigenvalues of the system (4.2), (4.3) are given by $\lambda_\alpha = 0 < \lambda_\beta = \frac{rq}{e}$. Due to the fact that q does not possess initial data, characteristic curves $\gamma_\alpha, \gamma_\beta$ are parametrized by t and x, respectively (see (3.9), (3.14)):

$$\left.\frac{dx}{dt}\right|_{\gamma_\alpha} = 0, \qquad \left. x(t;\alpha)\right|_{\gamma_\alpha} = \left. x(0;\alpha)\right|_{\gamma_\alpha} = \alpha < 0, \tag{4.7}$$

and

$$\left.\frac{dt}{dx}\right|_{\gamma_\beta} = \frac{e}{rq}, \qquad \left. t(x;\beta)\right|_{\gamma_\beta} > 0, \ \left. t(0;\beta)\right|_{\gamma_\beta} = \beta > 0. \tag{4.8}$$

This system has Riemann invariants, $|e|^r/q$ and q, which satisfy

$$\left.\frac{|e|^r}{q}\right|_{\gamma_\alpha} = \text{ constant} \tag{4.9}$$

and

$$q_{|\gamma_\beta} = \text{ constant}. \tag{4.10}$$

In order to examine when solutions in U_- remain smooth, we use the following results.

LEMMA 4.1. *Let $(e,q) \in C^1(U_-)$, and suppose that there are constants $\delta > 0, \alpha_0 < 0$ such that $e_0(\alpha) < -\delta$ for $\alpha < \alpha_0$ and $q_\flat(\beta) > -1/\delta$ for $\beta > 0$. Then for each β small enough, there is a unique characteristic, γ_β, connecting Γ with $\mathbb{R}_- = \{(x,t) : x < 0, t = 0\}$.*

Proof. Consider a region $\mathcal{W} \subset U_-$ bounded to the left and right by the line $x = \alpha < 0$ and by Γ, and to the top and bottom by the characteristic γ_β and by $\mathbb{R}_-$.

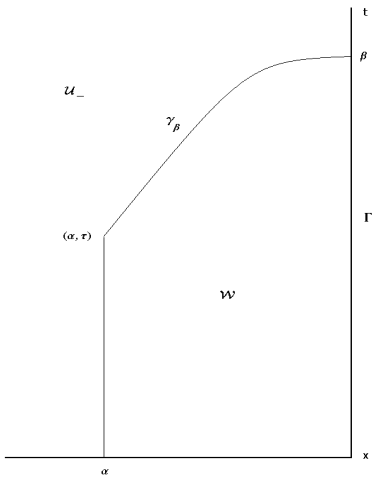


FIG. 4.1.

Let γ_β meet the line $x = \alpha$ at $(x,t) = (\alpha,\tau)$, where $\tau = \tau(\alpha,\beta) \le \beta$ by (4.8) (see Figure 4.1). We assume in the following that γ_β is entirely contained inside U_-.

Applying the divergence theorem to (4.2) on $\mathcal{W}$, and using (4.10), provides the relation

$$\int_\alpha^0 e_0(x)dx = \int_0^\beta q_\flat(t)dt - (1-r)\beta q_\flat(\beta) + \tau(1-r)q_\flat(\beta) - \int_0^\tau q(t,\alpha)dt. \tag{4.11}$$

Fixing β in (4.11) and differentiating with respect to α then gives

$$\begin{aligned} -e_0(\alpha) &= ((1-r)q_\flat(\beta) - q(\tau,\alpha))\frac{\partial\tau}{\partial\alpha} \\ &= -rq_\flat(\beta)\frac{\partial\tau}{\partial\alpha}. \end{aligned} \tag{4.12}$$

So $\frac{\partial\tau}{\partial\alpha} = \frac{e_0(\alpha)}{rq_\flat(\beta)} > \delta^2/r$ for $\alpha < \alpha_0$, which means that $\tau(\alpha,\beta)$ is bounded below, for fixed β, by a uniformly increasing function of α. Since $\tau(0,\beta) = \beta > 0$, it follows that τ must reduce to zero as α decreases from $\alpha = 0$, at which point γ_β meets $\mathbb{R}_-$.

We note that if γ_β connects Γ to $\mathbb{R}_-$, then (4.11) reduces to

$$\int_\alpha^0 e_0(x)dx = \int_0^\beta q_\flat(t)dt - (1-r)\beta q_\flat(\beta), \tag{4.13}$$

which gives a functional relation between α and β. In particular, differentiating with respect to α,

$$-e_0(\alpha)\frac{d\alpha}{d\beta} = rq_\flat(\beta) - (1-r)\beta q_\flat'(\beta) \tag{4.14}$$

shows that $\frac{d\alpha}{d\beta} < 0$ if $rq_\flat(\beta) - (1-r)\beta q_\flat'(\beta) < 0$. Thus, at least for small $\beta > 0$, one finds that $\frac{d\alpha}{d\beta} < 0$, since $q_\flat(0) < 0$ due to our assumption $\vartheta_0'^{+} > 0$.

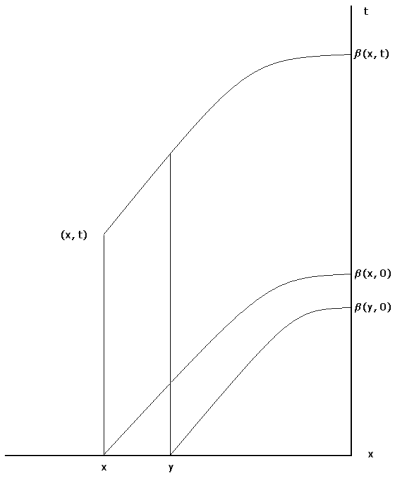


FIG. 4.2.

We will use the notation $\beta = \beta(\alpha, \tau)$ to denote the time of intersection of the characteristic curve going through $(x,t) = (\alpha, \tau)$, with Γ, i.e., $\beta = \beta(\alpha, \tau(\alpha, \beta)), \alpha < 0$. The inversion is possible whenever $\frac{\partial \tau}{\partial \beta} > 0$, which is a subject of the following results.

LEMMA 4.2. *Let* $(e,q) \in C^1(U_-)$, $q_\flat(\beta) < 0$, *and* $q_\flat'(\beta) \neq 0$. *Then* $\frac{\partial t}{\partial \beta}(x,\beta) > 0$, *and if* $\frac{\partial t}{\partial \beta} \to 0$ *as* $x \to x^* < 0$, *then* $|q_t(x, t(x,\beta))| \to \infty$.

Proof. By (4.10), we have that

$$q(x, t(x,\beta)) = q_\flat(\beta) \tag{4.15}$$

along γ_β, and so $q_{|_{\gamma_\beta}} < 0$. Differentiating relation (4.15) in β gives

$$(q_t t_\beta)_{|_{\gamma_\beta}} = q_t(0, t(0,\beta)) = q_\flat'(\beta) \tag{4.16}$$

since $t(0,\beta) = \beta$. Thus $\frac{\partial t}{\partial \beta} \to 0 \Leftrightarrow |q_t|_{|_{\gamma_\beta}} \to \infty$.

In the following we take, for simplicity, U_- to be the vertical strip $\{(x,t) \in \mathbb{R}^2_+ : x^* < x < 0\}$ for some $x^* < 0$. With regard to the theorem, we remark that the two crystalline materials, bismuth and sodium fluoride, have values of the material constant, r, lying between zero and one-half [10].

THEOREM 4.3. *Let* $\beta_1 \neq \beta_2$. *Then* γ_{β_1} *and* γ_{β_2} *do not intersect in* U_-, *provided that one of the following holds:*

Case 1. $r \in (0,1)$, *and either*

(i) $q_\flat'(\beta) > 0$

or

(ii) $q_\flat'(\beta) < 0$ *and* $r q_\flat(\beta) - (1-r)\beta q_\flat'(\beta) < 0$.

Case 2. $r > 1$, *and either*

(i) $q_\flat'(\beta) < 0$

or

(ii) $q_\flat'(\beta) > 0$ *and* $r q_\flat(\beta) - (1-r)\beta q_\flat'(\beta) < 0$.

Case 3. $r = 1$.

Proof. Recalling (4.9) and (4.10), for $(x,t) \in U_-$,

$$\frac{|e(x,t)|^r}{q_\flat(\beta(x,t))} = \frac{|e(x,t)|^r}{q(x,t)} = \frac{|e_0(x)|^r}{q(x,0)} = \frac{|e_0(x)|^r}{q_\flat(\beta(x,0))}, \tag{4.17}$$

which gives

$$|e(x,t)|^r = \frac{q_\flat(\beta(x,t))}{q_\flat(\beta(x,0))}|e_0(x)|^r. \tag{4.18}$$

As a result, (4.8) can be written as

$$\frac{dt}{dx}|_{\gamma_\beta} = -|q_\flat(\beta(x,t))|^{(1-r)/r}\frac{e_0(x)}{r|q_\flat(\beta(x,0))|^{1/r}}\,. \tag{4.19}$$

Integrating (4.19) along γ_β and using (4.10) again,

$$t = t(x,\beta) = \beta - |q_\flat(\beta)|^{(1-r)/r}\int_0^x \frac{e_0(y)}{r|q_\flat(\beta(y,0))|^{1/r}}dy \tag{4.20}$$

with $x* < x < 0$ and $0 < t < \beta$. Thus, differentiating with respect to $\beta = \beta(x,t)$,

$$\frac{\partial t}{\partial \beta} = 1 - \frac{1-r}{r}|q_\flat(\beta)|^{1/r-3}q_\flat(\beta)q_\flat'(\beta)\int_0^x \frac{e_0(y)}{r|q_\flat(\beta(y,0))|^{1/r}}dy. \tag{4.21}$$

Case 1(i) of the proof follows since $e_0 < 0, q_\flat < 0$, and characteristics spread with decreasing x, $\frac{\partial t}{\partial \beta} > 1$.

In order to establish Case 1(ii), let us first suppose that $(x^*,t^*) \in \partial U_-$ for $x^* < 0$, and that $t^* > 0$ is the least time at which solutions may fail to be in C^1. Next, assume that $\beta_1 < \beta_2$ are such that γ_{β_1} and γ_{β_2} intersect at (x^*,t^*), where $\gamma_{\beta_i} = \{(x,t) : t = t(x,\beta_i)\}$, $i = 1,2$. So (4.20) holds for $i = 1$, $i = 2$ at (x^*,t^*),

$$t^*(x^*,\beta_i) = \beta_i - |q_\flat(\beta_i)|^{(1-r)/r}\int_0^{x^*} \frac{e_0(y)}{r|q_\flat(\beta_i(y,0))|^{1/r}}dy, \quad i = 1,2, \tag{4.22}$$

where we note that the integral terms are identical for $i = 1,2$. Eliminating this term leads to the relation

$$t^* - \beta_2 = (t^* - \beta_1)\left(\frac{q_\flat(\beta_2)}{q_\flat(\beta_1)}\right)^{(1-r)/r}, \tag{4.23}$$

where $Q \equiv \frac{q_\flat(\beta_2)}{q_\flat(\beta_1)})^{(1-r)/r} > 1$ since $q_\flat' < 0$, by our hypothesis. However, it is easily shown that conditions (ii) imply

$$1 < \mathcal{Q} < \frac{\beta_2}{\beta_1}\,. \tag{4.24}$$

Thus $t^* > 0$ cannot exist since (4.23), (4.24) imply

$$t^* = \beta_1\frac{\beta_2/\beta_1 - Q}{1-Q} < 0. \tag{4.25}$$

Cases 2(i) and (ii) can be established as above, and Case 3 follows since (4.20) then implies $t_\beta \equiv 1$.

Remark. In addition to q_t, the terms e_t, q_x, and e_x naturally blow up where $\frac{\partial t}{\partial \beta} \to 0$, under the conditions of Theorem 4.3. This results from q remaining constant on γ_β and e remaining bounded away from zero on γ_α by (4.10) and (4.9). Equation (4.3) then implies $|q_x| \to \infty$ as $|q_t| \to \infty$, and therefore $|e_t| \to \infty$ by (4.2). On differentiating (4.9) in α, the same result follows for $|e_x|$.

Let us finally state some relationships more concisely. For any point $(x,t) \in U_-$, the solution $(e(x,t), q(x,t))$ to (4.2), (4.3) subject to (4.4), (4.5), and (4.6) takes the form

$$q(x,t) = q_\flat(\beta(x,t)), \tag{4.26}$$

$$e(x,t) = \left(\frac{q_\flat(\beta(x,t))}{q_\flat(\beta(x,0))}\right)^{1/r} e_0(x), \tag{4.27}$$

with

$$t = \beta(x,t) - \left(\frac{q_\flat(\beta(x,t))}{q_\flat(\beta(x,0))}\right)^{(1-r)/r} \beta(x,0), \tag{4.28}$$

where $\beta(x,0)$ is derived from (4.13) (with $\alpha = x$) and (4.28) comes from repeated use of (4.20) with $t > 0$ and $t = 0$ to eliminate the integral term. Using Theorem 4.3, we may invert (4.28) for fixed x in order to obtain $\beta(x,t)$ used in (4.26) and (4.27). Similarly, by using (4.20) in (4.21), one can find $\frac{\partial t}{\partial \beta}$ in terms of $\beta(x,t)$ in the useful form

$$\frac{\partial t}{\partial \beta} = 1 - \frac{1-r}{r}\frac{q_\flat'(\beta)}{q_\flat(\beta)}(\beta - t). \tag{4.29}$$

5. Some explicit smooth solutions. Having observed the role that heat flux continuity across Γ plays in solving (4.2)–(4.3), we now examine the effect on U_- of having a stationary or self-similar solution to (4.1) in U_+. The solutions are defined for U_- lying in the transcritical region.

In the case that the U_+ component of the solution is stationary, i.e., $e_t = q_x = 0$ with $q = -d(e)e_x < 0$, this implies that $q_\flat(\beta) = q_\flat(0)$ for all $\beta > 0$. Since $q|_{\gamma_\beta}$ remains constant, this means in turn that $q = q_\flat(0)$ in U_-. Consequently, since $(|e|^r/q)|_{\gamma_\alpha}$ is also constant, it follows that $e(x,t) = e_0(x)$ for arbitrary initial temperature distributions, $e_0(x)$, in U_-. As a result, the solution is everywhere stationary.

In the next case, we choose U_+ to be governed by the linear heat equation $d(e) = 1$ (assuming $c_\lambda = 1$), in which we can use, for example, any explicit solution formula for the semi-infinite interval. We set up the following example by means of a self-similar solution, in U_+, of the form $e(x,t) = e(x/\sqrt{t+1})$ and let this extend to U_- through Γ. The solution in U_+ is then represented by

$$e(x,t) = A\sqrt{\pi}\, erf\left(\frac{x}{2\sqrt{t+1}}\right), \quad A > 0, \tag{5.1}$$

giving heat flux as

$$q(x,t) = -\frac{A}{\sqrt{t+1}} e^{-x^2/4(t+1)} \quad \text{with } q_\flat(t) = -\frac{A}{\sqrt{t+1}}. \tag{5.2}$$

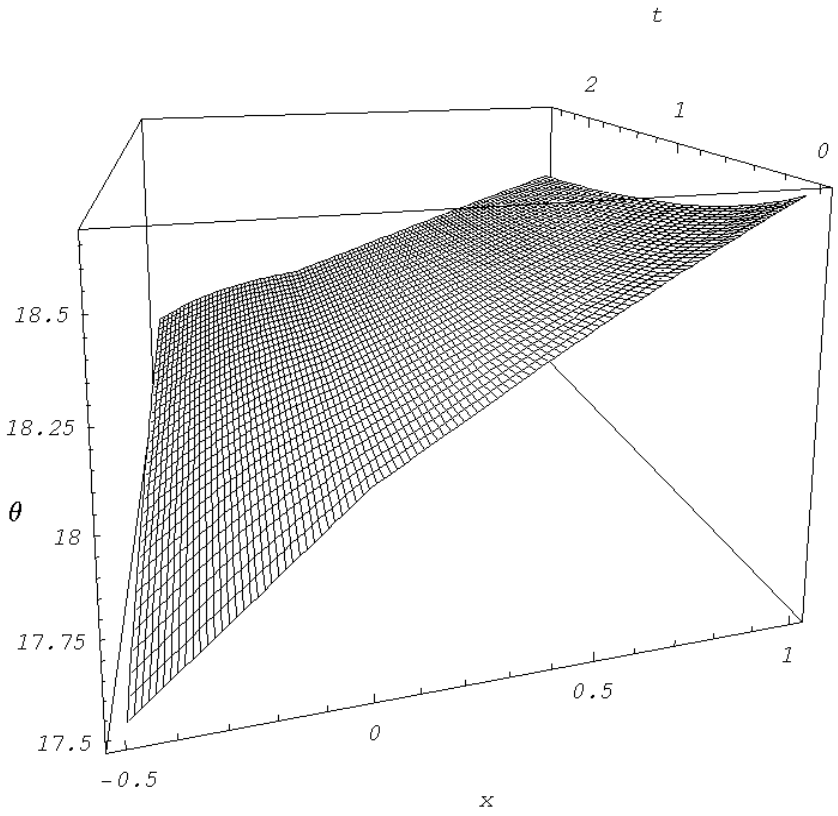


FIG. 5.1. *Temperature evolution with* $\vartheta_\lambda = 18K$, *for* $0 \le t \le 2.5\,\mu sec.$ *and* $-.5 \le x \le 1\,cm.$

Since $q_b'(t) > 0$, Theorem 4.1 guarantees that the characteristics in U_- do not intersect. We choose as initial data for e in U_-,

$$e_0(x) = x, \quad x < 0, \tag{5.3}$$

which is equivalent to $\vartheta_0(x) = x + \vartheta_\lambda$. The following connection can now be made from (4.13) (with $\alpha = x$),

$$\beta(x,0) = \left(\frac{B(x) + \sqrt{B(x)^2 - (1-r^2)}}{1+r} \right)^2 - 1, \tag{5.4}$$

where

$$B(x) = \frac{1}{4A}x^2 + 1. \tag{5.5}$$

Equation (4.28) then leads to a relation for $\beta = \beta(x,t)$,

$$t = \beta - (\beta+1)^{-(1-r)/2r}(\beta(x,0)+1)^{(1-r)/2r}\beta(x,0). \tag{5.6}$$

Taking, as a special case, $r = 1/3$ in (5.6) gives

$$\beta(x,t) = \frac{t - 1 + \sqrt{(1+t)^2 + 4H(x;1/3)}}{2}, \tag{5.7}$$

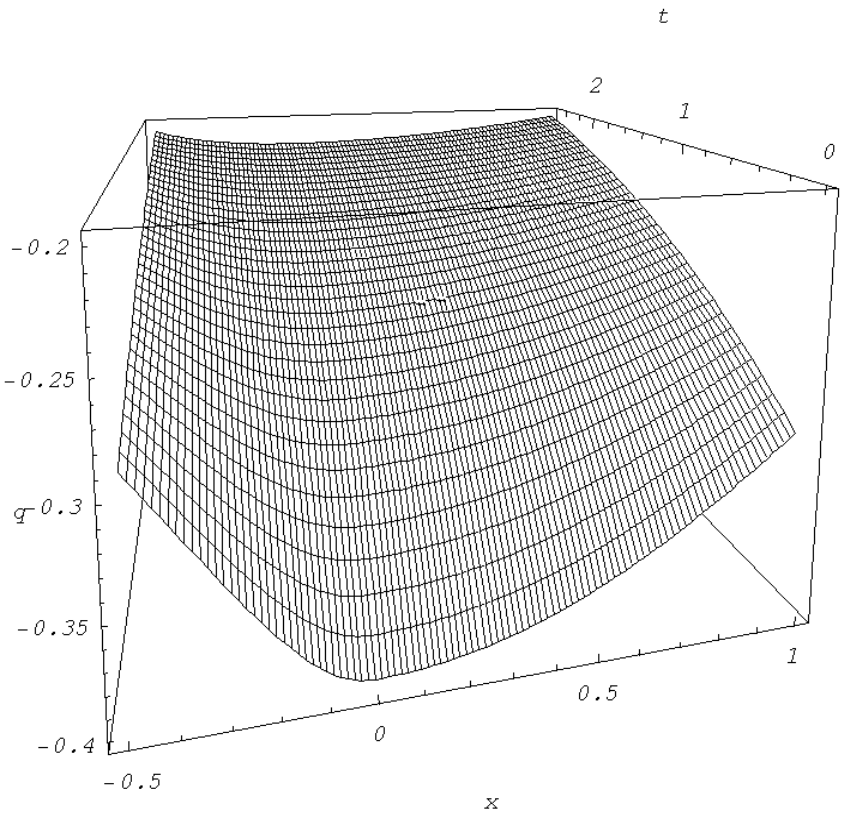


FIG. 5.2. *Heat flux evolution with* $\vartheta_\lambda = 18K$, *for* $0 \le t \le 2.5\,\mu sec.$, $-.5 \le x \le 1$ *cm, with* $-0.4 \le q < 0\,W/cm^2$.

where we have set

$$H(x;r) = (\beta(x,0)+1)^{(1-r)/2r}\beta(x,0). \tag{5.8}$$

Using (4.26), (4.27) together with (5.2), (5.3), (5.4), and (5.7) finally gives a solution for (e,q) in U_-,

$$e(x,t) = \left(\frac{3\sqrt{2}}{4}\right)^3 \left(\frac{B(x)+\sqrt{B(x)^2-8/9}}{\sqrt{t+1+\sqrt{(1+t)^2+4H(x;1/3)}}}\right)^3 e_0(x), \tag{5.9}$$

$$q(x,t) = -\sqrt{2}\frac{A}{\sqrt{t+1+\sqrt{(1+t)^2+4H(x;1/3)}}}, \tag{5.10}$$

where, from (2.11), (2.13) and the definition $\varepsilon'(\vartheta) = c_v(\vartheta)$,

$$\vartheta(x,t) = \vartheta_\lambda + e(x,t). \tag{5.11}$$

Figures 5.1 and 5.2 illustrate the behavior of $\vartheta(x,t)$ and $q(x,t)$ in $U_- \cup \Gamma \cup U_+$, using (5.1), (5.2), and (5.9)–(5.11) with $A = 2/5$.

One can see in Figure 5.2 that, despite the temperature gradient being initially larger to the left of the phase transition at $x = 0$ than to the right (see Figure 5.1), q is nevertheless greater there than at the transition itself. This illustrates distinctly "non-Fourier" behavior in U_-. It is also possible to observe in Figure 5.1 the derivative discontinuity in ϑ at $x = 0$ changing from "concave down" to "concave up" as time progresses. From (3.19), with $r = 1/3$, this change occurs when $\vartheta_x^+(t) = (\vartheta_0'^+)^3/(\vartheta_0'^-)^{1/2}$, or at about $t = 1$. With $q_\flat'(\beta) > 0$ and $r < 1$, as in the present case, it is easy to check that such a concavity change can occur only if $0 < \vartheta_0'^{\,+} < \vartheta_0'^{\,-}$.

REFERENCES

[1] H. Beck and R. Beck, *Heat-pulse propagation in dielectric solids*, Phys. Rev. B, 8 (1973), pp. 1669–1679.

[2] H. E. Jackson and C. T. Walker, *Thermal conductivity, second sound, and phonon-phonon interactions in NaF*, Phys. Rev. B, 3 (1971), pp. 1428–1439.

[3] H. E. Jackson, C. T. Walker, and T. F. Mcnelly, *Second sound in NaF*, Phys. Rev. Lett., 25 (1970), pp. 26–28.

[4] H. Li and K. Saxton, *Asymptotic behavior of solutions to quasilinear hyperbolic equations with nonlinear damping*, Quart. Appl. Math., 61 (2003), pp. 295–313.

[5] T. F. Mcnelly, S. J. Rogers, D. J. Chamin, R. J. Rollefson, W. M. Goubau, G. E. Schmidt, J. A. Krumhansi, and R. O. Pohl, *Heat pulses in NaF: Onset of second sound*, Phys. Rev. Lett., 24 (1970), pp. 100–102.

[6] V. Narayanamurti and R. C. Dynes, *Observation of second sound in bismuth*, Phys. Rev. Lett., 28 (1972), pp. 1461–1465.

[7] T. Ruggeri, A. Muracchini, and L. Seccia, *Shock waves and second sound in a rigid heat conductor: A critical temperature for NaF and Bi*, Phys. Rev. Lett., 64 (1990), pp. 2640–2643.

[8] T. Ruggeri, A. Muracchini, and L. Seccia, *Second sound and characteristic temperature in solids*, Phys. Rev. B, 54 (1996), pp. 332–339.

[9] K. Saxton, R. Saxton, and W. Kosinski, *On second sound at the critical temperature*, Quart. Appl. Math., 57 (1999), pp. 723–740.

[10] K. Saxton and R. Saxton, *Nonlinearity and memory effects in low temperature heat propagation*, Arch. Mech., 52 (2000), pp. 127–142.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1703–1726

FIBER DYNAMICS IN TURBULENT FLOWS: GENERAL MODELING FRAMEWORK*

NICOLE MARHEINEKE† AND RAIMUND WEGENER†

Abstract. The paper at hand deals with the modeling of turbulence effects on the dynamics of a long slender elastic fiber. Independent of the choice of the drag model, a general aerodynamic force concept is derived on the basis of the velocity field for the randomly fluctuating component of the flow. Its construction as a centered differentiable Gaussian field complies thereby with the requirements of the stochastic k-ϵ turbulence model and Kolmogorov's universal equilibrium theory on local isotropy.

Key words. fiber-fluid interaction, Cosserat rod, turbulence modeling, Kolmogorov's energy spectrum, double-velocity correlations, differentiable Gaussian fields

AMS subject classifications. 74F10, 76F60, 76F05, 60H40

DOI. 10.1137/050637182

1. Introduction.

The understanding of fiber-fluid interactions is of great interest for research, development, and production in textiles manufacturing. In the melt-spinning process of nonwoven materials, hundreds of individual endless fibers obtained by continuous extrusion of a melted polymer are stretched and entangled by highly turbulent air flows to finally form a web. The quality of this web and the resulting nonwoven material depends essentially on the dynamics of the fibers.

Fiber-turbulence interaction is a complex phenomenon that is governed by many factors, including the nature of the flow field, turbulent length scales, concentration, and size of fibers [13], [14]. Thin fibers decrease the turbulent intensity by increasing the apparent viscosity, whereas fibers whose thickness induces Reynolds numbers greater than some critical one intensify the turbulence due to vortex shedding [5]. Both mechanisms are strongly affected by the concentration. In the application considered here, however, the turbulence is not significantly influenced by the fibers. Hence, the turbulent flow is determined under neglect of suspended fibers, and its effect is theoretically studied on a single long slender fiber using a general drag model.

The fiber dynamics is described in section 2 by the Kirchhoff–Love equations for the motion of a Cosserat rod capable of large bending deformations. In terms of these the fiber slenderness allows the formulation of a wavelike system of nonlinear PDEs of fourth order with the algebraic constraint of inextensibility. The behavior of this system relies on the model for the external force imposed on the fiber by the turbulent flow, particularly on the choice of the air drag coefficients. The modeling of a generally valid aerodynamic force in section 4 is based on the splitting of the flow velocity into mean and fluctuation parts in the Reynolds-averaged Navier–Stokes equations. Thus, a centered differentiable Gaussian field for the randomly fluctuating component of the flow velocity is derived under the Global-from-Local Assumption of underlying locally isotropic and homogeneous turbulence, given in section 3. The

*Received by the editors August 1, 2005; accepted for publication (in revised form) March 20, 2006; published electronically July 17, 2006. This work has been supported by the Kaiserslautern Excellence Cluster *Dependable Adaptive Systems and Mathematical Modeling*.
http://www.siam.org/journals/siap/66-5/63718.html
†Fraunhofer-Institut für Techno- und Wirtschaftsmathematik (ITWM), Gottlieb-Daimler-Str., Geb. 49, D-67663 Kaiserslautern, Germany (marheineke@itwm.fhg.de, wegener@itwm.fhg.de).

construction of the initial condition for the respective local double-velocity correlation tensors satisfies thereby Kolmogorov's universal equilibrium theory as well as the local distribution of kinetic energy k and dissipation ϵ provided by the stochastic k-ϵ turbulence model. The dynamic behavior of the local correlation tensors is described by an advection equation, whose solution coincides with Taylor's hypothesis of frozen turbulence patterns. The temporal change of the global coherences is included by the averaging procedure. In section 4 the developed local velocity fluctuation fields hand their properties to the corresponding correlated local stochastic forces along the fiber. Gluing them together yields the global aerodynamic force that represents the turbulence effects on the fiber motion. Considering a wide class of feasible air drag models, the stated Global-from-Local Force Concept in combination with a linearization ansatz enables a good $\mathcal{L}^2$- and $\mathcal{L}^\infty$-approximation of the correlated force by Gaussian white noise with flow-dependent amplitude in case of a macroscopic description of the fiber.

2. Fiber dynamics. In the actual spinning process the fiber is endless, and its deposition plays a crucial role for the generation of the nonwoven material. However, as this paper focuses exclusively on the description of its dynamics due to the turbulent flow, the following considerations are restricted on a long slender elastic polymer fiber that is fixed at one end, suspended in a highly turbulent air stream. Let l denote its length and d its diameter with slenderness ratio $\delta = d/l \ll 1$. To describe its motion a one-dimensional model is derived on the dynamical Kirchhoff–Love theory for a Cosserat rod capable of large geometrically nonlinear deformations [2].

2.1. Equations of motion. Treat the fiber in the reference configuration as a body $\mathcal{B}$ given within a fixed Cartesian frame $\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3}$. Define $\mathbf{p}(\mathbf{z}, t)$ to be the position of the material point $\mathbf{z} \in \mathcal{B}$ at time t; then $\mathbf{p}(., t)$ states the actual configuration of the closure $\mathrm{cl}\mathcal{B}$ of $\mathcal{B}$ at time t. Introduce the curvilinear coordinates $\mathbf{x} := (x_1, x_2, s) \in \mathbb{R} \times \mathbb{R} \times [0, l]$ on $\mathcal{B}$ with s denoting the arc length. Then define $\tilde{\mathbf{p}}(\mathbf{x}, t) := \mathbf{p}(\tilde{\mathbf{z}}(\mathbf{x}), t)$, where $\tilde{\mathbf{z}}$ assigns $\mathbf{z} \in \mathrm{cl}\,\mathcal{B}$ to each $\mathbf{x}$. In particular, $\tilde{\mathbf{p}}(., ., s, t)$ describes the actual configuration of the cross section $\mathcal{B}(s)$ at time t.

The fiber model is now developed under the assumption that the position field $\tilde{\mathbf{p}}$ is determined by three vector-valued functions $\mathbf{r}(s, t)$, $\mathbf{d_1}(s, t)$, and $\mathbf{d_2}(s, t)$, i.e.,

$$\tilde{\mathbf{p}}(\mathbf{x}, t) = \mathbf{r}(s, t) + \mathbf{j}(\mathbf{r}(s, t), \mathbf{d_1}(s, t), \mathbf{d_2}(s, t), \mathbf{x}, t), \tag{1}$$

where the fiber line $\mathbf{r}(s, t)$ might be interpreted as the actual configuration of the center line at time t and the orthonormal directors $\mathbf{d_1}(s, t)$ and $\mathbf{d_2}(s, t)$ state the orientation of the actual configuration of $\mathcal{B}(s)$ at time t. Additionally, let $\mathbf{d_3}(s, t) = \mathbf{d_1}(s, t) \times \mathbf{d_2}(s, t)$. In terms of these functions, the feasible deformations of the fiber, e.g., flexure κ_1, κ_2, torsion τ, shear w_1, w_2, and dilatation w_3, are then expressed using the relations $\partial_s \mathbf{d_i} = \mathbf{b} \times \mathbf{d_i}$, $\mathbf{b} = \kappa_1 \mathbf{d_1} + \kappa_2 \mathbf{d_2} + \tau \mathbf{d_3}$, and $\partial_s \mathbf{r} = \sum_{i=1}^{3} w_i \mathbf{d_i}$. Thus,

$$\kappa_1 = -\mathbf{d_2} \cdot \partial_s \mathbf{d_3}, \quad \kappa_2 = -\mathbf{d_3} \cdot \partial_s \mathbf{d_1}, \quad \tau = -\mathbf{d_1} \cdot \partial_s \mathbf{d_2},$$
$$w_i = \partial_s \mathbf{r} \cdot \mathbf{d_i}.$$

According to Bernoulli's hypothesis that cross sections never experience warping as a consequence of deformation, the function $\mathbf{j}$ of (1) can moreover be prescribed by $\mathbf{j}(\mathbf{r}, \mathbf{d_1}, \mathbf{d_2}, \mathbf{x}, t) = x_1 \mathbf{d_1} + x_2 \mathbf{d_2}$. Assuming the reference configuration $\mathcal{B}$ to be a homogeneous (with respect to the density distribution) cylindrical body with circular cross sections of constant radius, the linear and angular impulse-momentum laws for

$\mathcal{B}$ read [2]

$$\partial_s \mathbf{q} + \mathbf{f} = \rho A \partial_{tt} \mathbf{r}, \tag{2}$$

$$\partial_s \mathbf{m} + \partial_s \mathbf{r} \times \mathbf{q} + \mathbf{l} = \rho I \sum_{i=1}^{2} (\partial_{tt} \mathbf{d_i} \times \mathbf{d_i}). \tag{3}$$

Here, ρ denotes density, $A = \pi d^2/4$ cross-sectional area, and $I = \pi d^4/64$ the moment of inertia. Closing the system by means of constitutive laws for inner force $\mathbf{q}$ and moment $\mathbf{m}$ as well as given outer line force $\mathbf{f}$ and moment $\mathbf{l}$, the Kirchhoff–Love equations (2) and (3) yield the description for fiber line and directors $\mathbf{r}$, $\mathbf{d_1}$, $\mathbf{d_2}$. The orthonormality of $\mathbf{d_1}$ and $\mathbf{d_2}$ thus reduces the number of unknowns to six. As no outer moment is acting on the fiber, $\mathbf{l} = \mathbf{0}$.

Constitutive laws for elastic materials look in general like

$$\mathbf{m} = \mathbf{M}(\kappa_1, \kappa_2, \tau, w_1, w_2, w_3, s), \qquad \mathbf{q} = \mathbf{Q}(\kappa_1, \kappa_2, \tau, w_1, w_2, w_3, s).$$

We apply here, in particular, Bernoulli–Euler beam theory that the inner moment $\mathbf{m}$ arises due to bending and torsion,

$$\mathbf{m} = EI(\kappa_1 \mathbf{d_1} + \kappa_2 \mathbf{d_2}) + GJ\tau \mathbf{d_3}, \tag{4}$$

with Young's modulus E, shear modulus G, and polar moment of inertia $J = \pi d^4/32$. Moreover, we interpret $\mathbf{q}$ as a vectorial Lagrangian multiplier and impose instead of a material law for $\mathbf{q}$ the following constraints on $\mathbf{d_3}$ and $\partial_s \mathbf{r}$:

$$\mathbf{d_3} = \frac{\partial_s \mathbf{r}}{\|\partial_s \mathbf{r}\|_2}, \qquad \|\partial_s \mathbf{r}\|_2 = 1. \tag{5}$$

This excludes shear and extensional deformation from the model. The restrictions are reasonable for a long slender fiber because shear and elongation are negligibly small in comparison to bending.

Apart from this, the slenderness enables a further simplification of system (2), (3). Nondimensionalizing (3) yields $\partial_s \mathbf{m} + \partial_s \mathbf{r} \times \mathbf{q} = \delta^2 D \sum_{i=1}^{2} (\partial_{tt}^2 \mathbf{d_i} \times \mathbf{d_i})$ with negligibly small right-hand side as the slenderness ratio δ satisfies $\delta \ll 1$ and $D = \mathcal{O}(1)$. Setting the right-hand side to zero, i.e.,

$$\partial_s \mathbf{m} + \partial_s \mathbf{r} \times \mathbf{q} = \mathbf{0}, \tag{6}$$

and using (5), we obtain $\partial_s \tau = 0$. Consequently, the torsion over the whole fiber equals the introduced torsion at the ends, $\tau = \tau_0$. Rewriting (4) thus gives $\mathbf{m} = EI(\mathbf{d_3} \times \partial_s \mathbf{d_3}) + GJ\tau_0 \mathbf{d_3}$, where $\partial_s \mathbf{d_3}$ represents the curvature vector $\partial_{ss} \mathbf{r} = \kappa \mathbf{n}$ with $\kappa = \sqrt{\kappa_1^2 + \kappa_2^2}$ and $\mathbf{n}$ a normal vector. Splitting the inner force $\mathbf{q}$ into tangential and normal parts with respect to the fiber position yields

$$\begin{aligned} \mathbf{q} &= (\mathbf{q} \cdot \mathbf{d_3})\, \mathbf{d_3} + \mathbf{d_3} \times (\mathbf{q} \times \mathbf{d_3}) \\ &\overset{(6)}{=} (\mathbf{q} \cdot \mathbf{d_3} + EI\,(\partial_{ss} \mathbf{d_3} \cdot \mathbf{d_3}))\, \mathbf{d_3} - EI\,(\mathbf{d_3} \cdot \mathbf{d_3})\, \partial_{ss} \mathbf{d_3} + GJ\tau_0\, \mathbf{d_3} \times \partial_s \mathbf{d_3}. \end{aligned}$$

Defining

$$\begin{aligned} T :=\ & \mathbf{q} \cdot \mathbf{d_3} + EI\,(\partial_{ss} \mathbf{d_3} \cdot \mathbf{d_3}) \\ \overset{(5)}{=}\ & \underbrace{\mathbf{q} \cdot \partial_s \mathbf{r}}_{\text{tension}} - \underbrace{EI\,\|\partial_{ss} \mathbf{r}\|_2^2}_{\text{curvature due to bending}} \end{aligned}$$

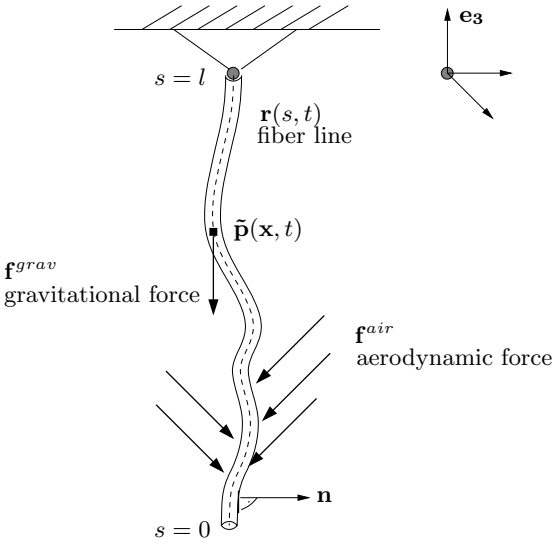


FIG. 1. *Fiber dynamics caused by external forces.*

as a modified tractive force, $\mathbf{q}$ depends exclusively on fiber line $\mathbf{r}$ and scalar Lagrangian multiplier T, and thus two more degrees of freedom vanish, which is consistent with the removing of the unknown directors $\mathbf{d_i}$. Plugging

$$\partial_s \mathbf{q} \overset{(5)}{=} \partial_s(T\,\partial_s \mathbf{r}) - EI\;\partial_{ssss}\mathbf{r} + GJ\tau_0\,\partial_s\mathbf{r} \times \partial_{sss}\mathbf{r}$$

into (2), the dynamics of a freely swinging fiber that is fixed at one end (cf. Figure 1) is described by

$$\begin{aligned} \rho A\,\partial_{tt}\mathbf{r}(s,t) &= \partial_s[T(s,t)\,\partial_s\mathbf{r}(s,t)] - EI\,\partial_{ssss}\mathbf{r}(s,t) + GJ\tau_0\,\partial_s\mathbf{r}(s,t) \times \partial_{sss}\mathbf{r}(s,t) \\ &\quad + \mathbf{f}^{grav} + \mathbf{f}^{air}(\mathbf{r}(.),s,t), \qquad (7) \\ \|\partial_s\mathbf{r}(s,t)\|_2 &= 1, \qquad (8) \end{aligned}$$

for $(s,t) \in (0,l) \times \mathbb{R}^+$ with Dirichlet conditions at the fixed end ($s = l$) and Neumann at the free end ($s = 0$),

$$\begin{aligned} \mathbf{r}(l,t) &= \mathbf{0}, & \partial_{ss}\mathbf{r}(0,t) &= \mathbf{0}, \\ \partial_s\mathbf{r}(l,t) &= \mathbf{e_3}, & \partial_{sss}\mathbf{r}(0,t) &= \mathbf{0}, \\ & & T(0,t) &= 0, \end{aligned}$$

as well as appropriate initial conditions ($t = 0$), e. g.,

$$\mathbf{r}(s,0) = (s-l)\,\mathbf{e_3}, \qquad \partial_t\mathbf{r}(s,0) = \mathbf{0}.$$

The Neumann conditions might be interpreted as natural boundary conditions, and the ending $s = 0$ is free of stress. Thus, neither outer moment nor force are acting on it. Moreover, $T(0,t)$ viewed as a tractive force vanishes. The Lagrangian multiplier $T(s,t)$ is thereby related to the algebraic constraint (8) of conservation of length. The behavior of our fiber system (7), (8)—but definitely also of the original one

(2), (3)—is strongly affected by the external line forces that arise due to gravitational ($\mathbf{f}^{grav} = \rho A\,\mathbf{g}$) and aerodynamic ($\mathbf{f}^{air}$) forces. We prescribe the aerodynamic force as a function depending on arc length s, time t, and additionally on the fiber line $\mathbf{r} : [0, l] \times \mathbb{R}_0^+ \to \mathbb{R}^3$ in a functional sense.

In this work, we initially introduce no twisting at the fiber ends, $\tau_0 = 0$, such that (7) simplifies to a wavelike system of nonlinear PDEs of fourth order, if the feasible functional dependence of the aerodynamic force is localized on the fiber point, e.g., $\mathbf{f}^{air}(\mathbf{r}(.), s, t) = \mathbf{f}^{air}(\mathbf{r}(s,t), \partial_s\mathbf{r}(s,t), \partial_t\mathbf{r}(s,t), s, t)$.

2.2. Air drag. The description of the fiber dynamics in a turbulent flow relies essentially on the model for the aerodynamic force $\mathbf{f}^{air}$ that is imposed on the fiber by the fluid. Neglecting the fiber influence on the flow, a dimensionless air drag coefficient c^{drag} based on Reynolds (Re), Mach, and Froude numbers can be associated with $\mathbf{f}^{air}$ [16]. If just frictional and inertial forces occur in the flow around the fiber, c^{drag} is particularly determined by

$$c^{drag} = \frac{\|\mathbf{f}^{air}\|_2}{0.5\,\rho^{air}\,d\,\|\mathbf{v}\|_2^2},$$

with air density ρ^{air}, fiber diameter d, and relative velocity between fluid flow and fiber $\mathbf{v} = \mathbf{u} - \partial_t\mathbf{r}$. Thus, the magnitude of the line force is proportional to Bernoulli's dynamic pressure $p = 0.5\rho^{air}\|\mathbf{v}\|_2^2$ acting along d. In general, we characterize a feasible air drag model by a function $\mathbf{f} : \mathbb{R}^3 \times \mathbb{R}^2 \to \mathbb{R}^3$, depending on a given velocity and a normalized direction. In this context, the aerodynamic force of (7) reads as

$$\mathbf{f}^{air}(\mathbf{r}(.), s, t) = \mathbf{f}(\mathbf{u}(\mathbf{r}(s,t), t) - \partial_t\mathbf{r}(s,t), \partial_s\mathbf{r}(s,t)), \tag{9}$$

where the flow velocity $\mathbf{u} : \mathbb{R}^3 \times \mathbb{R}_0^+ \to \mathbb{R}^3$ acts as outer input parameter to the fiber problem. However, as this instantaneous flow velocity is not available from a stochastic description of a turbulent flow, we derive a concept for a random Gaussian aerodynamic force in this work. Note that this concept utilizes exclusively the functional relation $\mathbf{f}$ and is hence generally applicable to a wide class of air drag models.

3. Model for a velocity fluctuation field. Consider the flow to be subsonic, highly turbulent, with small pressure gradients and Mach number Ma $< 1/3$. Then it can be modeled as an incompressible Newtonian fluid using the incompressible Navier–Stokes equations (NSE). Solving NSE by means of direct numerical simulation (DNS) gives the exact velocity field needed for the determination of the force of (9). However, DNS presupposes the resolution of all vortices ranging from the large energy-bearing ones of length l_T to the smallest, viscously determined Kolmogorov vortices of size η with $l_\mathrm{T}/\eta = \mathrm{Re}^{3/4}$ [18]. Therefore, the number of grid points that are required for the refinement of a three-dimensional domain is proportional to $\mathrm{Re}^{9/4}$. Despite the existence of recent high speed computers, DNS is thus still restricted to simple, small Reynolds number flow. Large eddy simulation (LES) as a combination of DNS and stochastic turbulence models offers an alternative. Applying a low-pass filter on NSE, only the vortices of large scales are resolved directly, whereas the small vortices are taken into account by a stochastic approximation of their effect on the larger ones [15]. However, for the relevant flow regime under consideration, LES also requires enormous computational capacity, due to very long run time and high memory demands. Because of neglect of the fiber influence on the flow, which

leads to the decoupling of fiber and flow computation, the Reynolds number is here specified by the machine geometry and not by the fiber diameter. For the resulting high Reynolds number flow, the stochastic turbulence models represent a reasonable compromise between accuracy and computational efficiency [6]. They are based on the Reynolds-averaged NSE (RANS), where the instantaneous velocity $\mathbf{u} : \mathbb{R}^3 \times \mathbb{R}_0^+ \to \mathbb{R}^3$ is expressed as the sum of a mean $\bar{\mathbf{u}}$ and a fluctuating part $\mathbf{u}'$:

$$\mathbf{u}(\mathbf{x}, t) = \bar{\mathbf{u}}(\mathbf{x}, t) + \mathbf{u}'(\mathbf{x}, t).$$

Applying in particular the standard k-ϵ model [12] yields a deterministic description of mean velocity $\bar{\mathbf{u}} : \mathbb{R}^3 \times \mathbb{R}_0^+ \to \mathbb{R}^3$, turbulent kinetic energy $k : \mathbb{R}^3 \times \mathbb{R}_0^+ \to \mathbb{R}^+$, and dissipation rate $\epsilon : \mathbb{R}^3 \times \mathbb{R}_0^+ \to \mathbb{R}^+$. Hereby, the variables k and ϵ might be interpreted as parameters of an $\mathbb{R}^3$-valued differentiable random field representing the fluctuations $(\mathbf{u}'_{\mathbf{x},t}, (\mathbf{x}, t) \in \mathbb{R}^3 \times \mathbb{R}_0^+)$:

$$k(\mathbf{x}, t) = \frac{1}{2}\, \mathbb{E}[\mathbf{u}'(\mathbf{x}, t) \cdot \mathbf{u}'(\mathbf{x}, t)], \tag{10}$$

$$\epsilon(\mathbf{x}, t) = \nu\, \mathbb{E}[\nabla \mathbf{u}'(\mathbf{x}, t) : \nabla \mathbf{u}'(\mathbf{x}, t)]. \tag{11}$$

To conform to the notation of probability theory and turbulence literature, note that the mean $\mathbb{E}[\mathbf{u}']$ equals the averaged quantity $\overline{\mathbf{u}'}$. Constructing a suitable fluctuation field requires the analysis of the turbulent behavior of the flow, which is characterized by means of statistical quantities, i.e., double-velocity correlations revealing spatial and temporal relations within a domain.

DEFINITION 1 (velocity fluctuation field). *Let* $(\Omega, \mathcal{A}, \mathrm{P})$ *be a probability space. The* velocity fluctuation field *of a turbulent flow is said to be a centered* $\mathbb{R}^3$*-valued random field* $(\mathbf{\Phi}_{\mathbf{x},t}, (\mathbf{x}, t) \in \mathbb{R}^3 \times \mathbb{R}_0^+)$ *with* $\mathbf{\Phi}_{\mathbf{x},t} \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathrm{P})$*. Its correlation tensor reads*

$$\mathbf{\Gamma}(\mathbf{x}, t, \mathbf{y}, \tau) = \mathbb{E}[\mathbf{\Phi}(\mathbf{x}, t) \otimes \mathbf{\Phi}(\mathbf{y}, \tau)].$$

Classifying turbulence, we face shear turbulence in practice. Although it can be simulated via RANS models, this kind of flow is hardly understood. Physical interpreting and mathematical handling of the statistical quantities is extremely difficult. Therefore, it is helpful to consider approximations like homogeneous and/or isotropic turbulent flows. Isotropy obviously has a hypothetical character, but knowledge of its characteristics forms a fundamental basis for the study of actual, anisotropic turbulent flows. Certain theoretical considerations concerning the energy transfer through the eddy-size spectrum from the larger to the smaller eddies (i.e., forward-scatter) lead to the conclusion that the fine structure of anisotropic turbulent flows is almost isotropic (Kolmogorov's local isotropy hypothesis [8]). Thus, many features of isotropic turbulence apply to phenomena in actual turbulence that are mainly determined by the fine-scale structure. Even if we consider the anisotropic large-scale structure of an actual turbulence, it is possible to treat such a turbulence, for purposes of a first approximation, as isotropic. The differences are mostly sufficiently small [10]. However, effects like back-scatter are not included. As velocity fluctuations in an isotropic flow are Gaussian [7], we restrict ourselves to Gaussian flows that are uniquely determined by their correlation tensor. This motivates the following assumption.

GLOBAL-FROM-LOCAL ASSUMPTION. *Let* $(\Omega, \mathcal{A}, \mathrm{P})$ *be a probability space. Let* $\{(\mathbf{w}^{\mathbf{y},\tau}_{\mathbf{x},t}, (\mathbf{x}, t) \in \mathbb{R}^3 \times \mathbb{R}_0^+),\ (\mathbf{y}, \tau) \in \mathbb{R}^3 \times \mathbb{R}_0^+\}$ *be a family of local velocity fluctuation fields that correspond to spatially and temporally homogeneous, isotropic, and incompressible Gaussian flows with respect to the points* $(\mathbf{y}, \tau)$*. Let* $\tilde{\boldsymbol{\gamma}}^{\mathbf{y},\tau} : (\mathbb{R}^3 \times \mathbb{R}_0^+)^2 \to$

$\mathbb{R}^{3\times 3}$ *denote their respective correlation tensors. For each local field the quantities* $k = k(\mathbf{y},\tau)$, $\epsilon = \epsilon(\mathbf{y},\tau)$, *and* $\bar{\mathbf{u}} = \bar{\mathbf{u}}(\mathbf{y},\tau)$ *are taken as constant. Then we assume that our actual global fluctuation field* $\mathbf{u}'$ *can be constructed as*

$$\mathbf{u}'(\mathbf{x},t) = \langle \mathbf{w}^{\mathbf{y},\tau}(\mathbf{x},t)\rangle_{M(\mathbf{x},t)}, \tag{12}$$

with $M(\mathbf{x},t) = \{(\mathbf{y},\tau) \in \mathbb{R}^3 \times \mathbb{R}_0^+ \mid \|\mathbf{x}-\mathbf{y}-\bar{\mathbf{u}}(\mathbf{x},t)(t-\tau)\|_2 \le l_{\mathrm{T}} \wedge |t-\tau| \le t_{\mathrm{T}}\}$, $|M(\mathbf{x},t)| = \int_{M(\mathbf{x},t)} d\mathbf{y}\, d\tau$, *and turbulent large-scale length* l_{T} *and time* t_{T}. *The brackets* $\langle\cdot\rangle$ *represent the Gaussian average that is uniquely prescribed by expectation and covariance (resp.) correlations according to*

$$\begin{aligned}
\mathbb{E}[\mathbf{u}'(\mathbf{x},t)] &= \frac{1}{|M(\mathbf{x},t)|}\int_{M(\mathbf{x},t)} \mathbb{E}[\mathbf{w}^{\mathbf{y},\tau}(\mathbf{x},t)]\; d\mathbf{y}\, d\tau = \mathbf{0},\\
\mathrm{Cov}(\mathbf{u}'(\mathbf{x_1},t_1),\mathbf{u}'(\mathbf{x_2},t_2)) &= \frac{1}{\sqrt{|M(\mathbf{x_1},t_1)||M(\mathbf{x_2},t_2)|}}\int_{M(\mathbf{x_1},t_1)\cap M(\mathbf{x_2},t_2)} \tilde{\boldsymbol{\gamma}}^{\mathbf{y},\tau}(\mathbf{x_1},t_1,\mathbf{x_2},t_2)\; d\mathbf{y}\, d\tau\\
&= \boldsymbol{\Gamma}'(\mathbf{x_1},t_1,\mathbf{x_2},t_2).
\end{aligned}$$

Note that the terminology used will be explained in the course of this section. The construction rule (12) enables the realization of a globally inhomogeneous and anisotropic turbulent flow on the basis of a very limited number of data stemming from general turbulence theory and specific, case-dependent k-ϵ simulations. So far, it is not clear at all whether such a differentiable turbulent field exists. The underlying local fluctuation fields $\mathbf{w}^{\mathbf{y},\tau}$, which can be interpreted as fine-scale structure of the turbulence, satisfy Kolmogorov's local isotropy hypothesis as well as the local kinetic energy k and dissipation ϵ distribution of the k-ϵ model. Averaging their statistical parameters over a region M where the local stochastic quantities differ only slightly glues them together to the global fluctuation field $\mathbf{u}'$, the anisotropic large-scale structure. The respective global quantities $k_{\mathbf{u}'}$ and $\epsilon_{\mathbf{u}'}$ are thus prescribed as averages of the hardly varying local ones. This is indicated by using the turbulent large-scale length l_{T} and time t_{T}. Presuming global homogeneity, the global and local quantities coincide and obey (10), (11), as desired.

In the following, we deal with the generation of the centered local fluctuation fields by modeling their correlation tensors. Therefore we skip the superscripts denoting the respective points. To determine the temporal behavior of the correlations, we first construct an initial condition for the correlation tensor satisfying the assumptions of homogeneity and isotropy as well as the requirements of the k-ϵ model and Kolmogorov's energy spectrum (sections 3.1–3.4). This initial condition meets the smoothness demands and guarantees the differentiability of the actual global field. Then we formulate an advection equation for the dynamics, whose solution coincides with Taylor's hypothesis of frozen turbulence (section 3.5). In section 3.6 we finally formulate the global fluctuation field as Ito-integral over the local fields, which yields the positive definite correlation tensor proposed in the Global-from-Local Assumption.

3.1. Locally homogeneous isotropic turbulence.

Definition 2 (homogeneous turbulence). *Let* $(\Omega,\mathcal{A},\mathrm{P})$ *be a probability space. Let* $(\boldsymbol{\Phi}_{\mathbf{x},t},(\mathbf{x},t) \in \mathbb{R}^3\times\mathbb{R}_0^+)$ *be a velocity fluctuation field with correlation tensor* $\boldsymbol{\Gamma}$. *A turbulent flow is said to be* spatially homogeneous *if* $\boldsymbol{\Gamma}$ *is invariant regarding spatial translations, i.e.,*

$$\boldsymbol{\Gamma}(\mathbf{x},t,\mathbf{y},\tau) = \boldsymbol{\Gamma}(\mathbf{x}-\mathbf{a},t,\mathbf{y}-\mathbf{a},\tau) \quad \forall\, \mathbf{a}\in\mathbb{R}^3.$$

A turbulent flow is said to be temporally homogeneous *if* $\mathbf{\Gamma}$ *is invariant regarding time shifts, i.e.,*

$$\mathbf{\Gamma}(\mathbf{x},t,\mathbf{y},\tau)=\mathbf{\Gamma}(\mathbf{x},t-a,\mathbf{y},\tau-a)\quad\forall\, a\in\mathbb{R}.$$

Definition 3 (isotropic turbulence). *Let* $(\Omega,\mathcal{A},\mathrm{P})$ *be a probability space. Let* $(\mathbf{\Phi}_{\mathbf{x},t},(\mathbf{x},t)\in\mathbb{R}^3\times\mathbb{R}_0^+)$ *be a velocity fluctuation field with correlation tensor* $\mathbf{\Gamma}$*. A turbulent flow is said to be* isotropic *if* $\mathbf{\Gamma}$ *is invariant regarding rotations and reflections, i.e.,*

$$\mathbf{\Gamma}(\mathbf{x},t,\mathbf{y},t)=\mathbf{S}\;\mathbf{\Gamma}(\mathbf{S}^{-1}\mathbf{x},t,\mathbf{S}^{-1}\mathbf{y},t)\;\mathbf{S}^t\quad\forall\,\mathbf{S}\in\mathcal{O}(3). \tag{13}$$

The correlation tensor $\tilde{\boldsymbol{\gamma}}$ corresponding to a local fluctuation field $\mathbf{w}$, $\tilde{\boldsymbol{\gamma}}(\mathbf{x},t,\mathbf{y},\tau)=\mathbb{E}[\mathbf{w}(\mathbf{x},t)\otimes\mathbf{w}(\mathbf{y},\tau)]$, depends only on the spatial and temporal difference of its arguments due to homogeneity. Thus, we define

$$\boldsymbol{\gamma}(\mathbf{z},\varsigma)=\tilde{\boldsymbol{\gamma}}(\mathbf{x}+\mathbf{z},t+\varsigma,\mathbf{x},t). \tag{14}$$

To derive the structure of the initial correlation tensor in the following, we focus now on

$$\boldsymbol{\gamma}_0(\mathbf{z})=\boldsymbol{\gamma}(\mathbf{z},0)\quad\text{ or }\quad\tilde{\boldsymbol{\gamma}}_0(\mathbf{x},\mathbf{y})=\tilde{\boldsymbol{\gamma}}(\mathbf{x},t,\mathbf{y},t).$$

Properties of the Initial Correlation Tensor. *The correlation tensor corresponding to a homogeneous isotropic turbulent flow has the following properties:*

$$\boldsymbol{\gamma}_0(\mathbf{z})=\boldsymbol{\gamma}_0(-\mathbf{z}), \tag{15}$$

$$\boldsymbol{\gamma}_0(\mathbf{z})\quad\textit{is symmetric}, \tag{16}$$

$$\boldsymbol{\gamma}_0(\mathbf{0})=c\mathbf{I},\quad c\neq 0, \tag{17}$$

$$\boldsymbol{\gamma}_0(\mathbf{z})\quad\textit{has two different eigenvalues:}\quad c_1(z)\ \textit{in}\ \tfrac{\mathbf{z}}{z}\ \textit{and}\ c_2(z)\ \textit{in the respective normal plane,} \tag{18}$$

$$\boldsymbol{\gamma}_0(\mathbf{z})=\frac{c_1(z)-c_2(z)}{z^2}\mathbf{z}\otimes\mathbf{z}+c_2(z)\mathbf{I},\quad z=\|\mathbf{z}\|_2. \tag{19}$$

Hereby, the symmetry of $\boldsymbol{\gamma}_0$, (16), results directly from its definition and the permutability of the arguments (15) that is concluded from the translation and reflection invariance. Applying additionally rotation invariance yields (17) and (18). The general form (19) is deduced from the spectral theorem using the eigenvalues of (18).

The one-dimensional functions c_1 and $c_2\in\mathcal{C}^\infty(\mathbb{R}_0^+)$ can be interpreted as longitudinal and lateral correlations [10]. In general $c_1\neq c_2$, but for $z\to 0$ we have $c_2(z)\to c_1(z)\to c$, with c given in (17).

As a turbulent flow contains a continuous spectrum of scales, it is convenient to introduce the spectral density $\mathbf{M}$ depending on the wave vector $\boldsymbol{\kappa}$. Assuming absolute Lebesgue-continuity of the spectrum of the underlying fluctuation velocity field [4], the spectral density $\mathbf{M}$ is the Fourier transform of the correlation tensor $\boldsymbol{\gamma}_0$,

$$\mathbf{M}(\boldsymbol{\kappa})=\frac{1}{(2\pi)^3}\int_{\mathbb{R}^3}e^{-i\mathbf{z}\cdot\boldsymbol{\kappa}}\boldsymbol{\gamma}_0(\mathbf{z})\,d\mathbf{z}. \tag{20}$$

Then, the spectral energy distribution (energy spectrum) E is defined by

$$E(\kappa)=\frac{1}{2}\kappa^2\int_{S^2}\mathrm{tr}(\mathbf{M}(\kappa\mathbf{e}))\,d\mathbf{e}, \tag{21}$$

with $\kappa = \|\boldsymbol{\kappa}\|_2$, unit sphere S^2, and unit vector $\mathbf{e} \in S^2$.

PROPERTIES OF THE SPECTRAL DENSITY. *The spectral density corresponding to a homogeneous isotropic turbulent flow has the following properties:*

$$\mathbf{M}(\boldsymbol{\kappa}) = \frac{e_1(\kappa) - e_2(\kappa)}{\kappa^2} \boldsymbol{\kappa} \otimes \boldsymbol{\kappa} + e_2(\kappa)\mathbf{I}, \tag{22}$$

$$\mathrm{tr}\mathbf{M}(\kappa) = \frac{1}{2\pi} \frac{E(\kappa)}{\kappa^2}. \tag{23}$$

Due to the Fourier relation (20), $\mathbf{M}$ inherits the isotropic property (13) from $\boldsymbol{\gamma}_0$ and therefore has an analogous representation with the one-dimensional spectral functions $e_1,\ e_2 \in \mathcal{C}^\infty(\mathbb{R}_0^+)$. The connection (23) between trace $\mathrm{tr}\mathbf{M}$ and energy spectrum E can be concluded from (21). Because of isotropy the sphere integral becomes $\int_{S^2} \mathrm{tr}(\mathbf{M}(\kappa\mathbf{e}))d\mathbf{e} = 4\pi\,\mathrm{tr}\mathbf{M}(\kappa)$.

In our case of an incompressible local flow field $\mathbf{w}$, the presented characteristics and dependencies of correlation and spectral functions can be simplified, which halves the number of unknowns and results in a well-structured Sine–Fourier relation between c_1 and E.

INFLUENCE OF INCOMPRESSIBILITY ON CORRELATIONS AND SPECTRAL DENSITY. *Assuming incompressibility, the following relation for the correlation functions c_1 and $c_2 : \mathbb{R}_0^+ \to \mathbb{R}$ is valid:*

$$c_1(z) + \frac{z}{2}\partial_z c_1(z) = c_2(z). \tag{24}$$

Moreover, the spectral functions e_1 and $e_2 : \mathbb{R}_0^+ \to \mathbb{R}$ are given by

$$e_1(\kappa) = 0, \qquad e_2(\kappa) = \frac{1}{4\pi} \frac{E(\kappa)}{\kappa^2}. \tag{25}$$

Relation (24) is concluded from the incompressibility using

$$\mathbf{0} = \mathbb{E}[(\nabla_{\mathbf{x}} \cdot \mathbf{w}(\mathbf{x},t))\,\mathbf{w}(\mathbf{y},t)] = \nabla_{\mathbf{x}} \cdot \tilde{\boldsymbol{\gamma}}_0(\mathbf{x},\mathbf{y}) \overset{\mathbf{z}:=\mathbf{x}-\mathbf{y}}{=} \nabla_{\mathbf{z}} \cdot \boldsymbol{\gamma}_0(\mathbf{z})$$

and substituting (19). Analogously to the correlation functions, the number of unknown spectral functions can be reduced to one. In particular, (25) is deduced by combining $\mathbf{0} = \nabla_{\mathbf{z}} \cdot \boldsymbol{\gamma}_0(\mathbf{z}) = i \int_{\mathbb{R}^3} e^{i\boldsymbol{\kappa}\cdot\mathbf{z}}\,\mathbf{M}(\boldsymbol{\kappa})\boldsymbol{\kappa}\,d\boldsymbol{\kappa}$ and thus $\mathbf{M}(\boldsymbol{\kappa})\boldsymbol{\kappa} = \mathbf{0}$ for all $\boldsymbol{\kappa} \in \mathbb{R}^3$ with (23).

For an incompressible isotropic and homogeneous turbulent flow, the correlation tensor $\boldsymbol{\gamma}_0$ of second order can thus be expressed by the single one-dimensional correlation function c_1. In particular,

$$\mathrm{tr}\boldsymbol{\gamma}_0(z) = 3c_1(z) + z\,\partial_z c_1(z) = \frac{1}{z^2}\,\partial_z\,(z^3\,c_1(z)). \tag{26}$$

Consequently, the whole local fluctuation velocity field is uniquely determined by c_1, whose relation to E will be useful for the further realization of the initial correlations.

RELATION BETWEEN CORRELATION FUNCTION AND ENERGY SPECTRUM. *Let c_1 be the correlation function, and E the energy spectrum corresponding to an homogeneous, isotropic, and incompressible turbulent flow. This implies their finiteness over the whole definition range. Then the following relations are valid:*

$$c_1(z) = \frac{2}{z^3} \int_0^\infty \frac{1}{\kappa}\,\partial_\kappa \left(\frac{E(\kappa)}{\kappa}\right) \sin(\kappa z)\,d\kappa, \tag{27}$$

$$E(\kappa) = \frac{\kappa}{\pi} \int_0^\infty \frac{1}{z}\,\partial_z\,(z^3\,c_1(z))\,\sin(\kappa z)\,dz. \tag{28}$$

The Sine–Fourier relations follow from the respective connections of c_1 and E to the Fourier transforms γ_0 and $\mathbf{M}$. Plugging (26) and (23) into

$$\mathrm{tr}\boldsymbol{\gamma}_0(z) = \int_0^\infty \int_0^{2\pi} \int_{-1}^{1} e^{i\kappa z q}\kappa^2 \,\mathrm{tr}\mathbf{M}(\kappa)\, dq\, d\phi\, d\kappa = \frac{4\pi}{z}\int_0^\infty \kappa\,\mathrm{tr}\mathbf{M}(\kappa)\sin(\kappa z)\, d\kappa$$

gives $\partial_z(z^3\, c_1(z)) = 2z\int_0^\infty E(\kappa)/\kappa\, \sin(\kappa z)\, d\kappa$ and consequently, after some algebraic manipulations, (27).

Further Decisive Coherences. *Further relevant relations between longitudinal correlation function c_1 and energy spectrum E are formulated as*

$$c_1(0) = \frac{2}{3}\int_0^\infty E(\kappa) d\kappa, \tag{29}$$

$$\partial_{zz} c_1(0) = -\frac{2}{15}\int_0^\infty E(\kappa)\kappa^2 d\kappa. \tag{30}$$

By means of partial integration, (27) can be rewritten as

$$c_1(z) = 2\int_0^\infty E(\kappa)\frac{\sin(\kappa z) - \kappa z\cos(\kappa z)}{k^3 z^3}\, d\kappa,$$

from which L'Hospital directly yields (29) and (30).

Finally, the differentiability of a homogeneous Gaussian flow can be concluded from

$$\int_{\mathbb{R}^3} (\ln(1+\kappa))^\alpha\ \kappa^{2p}\,\mathbf{M}(\boldsymbol{\kappa}) d\boldsymbol{\kappa} < \boldsymbol{\infty} \quad \text{for } \alpha > 3. \tag{31}$$

According to [11], equation (31) ensures the existence of an almost surely p-times sample differentiable modification, which we equate to the considered flow for purposes of an intuitive notation. As for our isotropic incompressible local flow field, the differentiability can thus be formulated as a requirement on the decay of the energy spectrum E by rewriting the volume integral of (31) with the help of (22) and (25) as

$$\int_0^\infty \int_{S^2} (\ln(1+\kappa))^\alpha\, \kappa^{2p}\, \frac{E(\kappa)}{4\pi}\, (\mathbf{I} - \mathbf{e}\otimes\mathbf{e})\, d\mathbf{e}\, d\kappa = \int_0^\infty \frac{2}{3}\,(\ln(1+\kappa))^\alpha\, \kappa^{2p}\, E(\kappa) d\kappa\ \mathbf{I}. \tag{32}$$

3.2. Parameters from the k-ϵ model. The kinetic turbulent energy k and the dissipation rate ϵ stemming from the k-ϵ turbulence model act as parameters for the differentiable local fluctuation fields. Presupposing an isotropic, homogeneous, and incompressible Gaussian flow, they can be expressed in terms of the correlation c_1 (resp., the energy function E).

With $\mathbb{E}[\mathbf{w}(\mathbf{x},t)\cdot\mathbf{w}(\mathbf{x},t)] = \mathrm{tr}\boldsymbol{\gamma}_0(0) \overset{(26)}{=} 3\, c_1(0)$, we obtain

$$k = \frac{1}{2}\mathbb{E}[\mathbf{w}(\mathbf{x},t)\cdot\mathbf{w}(\mathbf{x},t)] = \frac{3}{2}c_1(0) \overset{(29)}{=} \int_0^\infty E(\kappa) d\kappa. \tag{33}$$

As for ϵ, we consider $\mathbb{E}[\nabla\mathbf{w}(\mathbf{x},t)\otimes\nabla\mathbf{w}(\mathbf{y},t)] = \nabla_{\mathbf{x}}\nabla_{\mathbf{y}}\tilde{\boldsymbol{\gamma}}_0(\mathbf{x},\mathbf{y}) = -\nabla_{\mathbf{z}}\nabla_{\mathbf{z}}\boldsymbol{\gamma}_0(\mathbf{z})$ with $\mathbf{z} = \mathbf{x} - \mathbf{y}$. Thus, the dissipation reads as

$$\epsilon = \nu\,\mathbb{E}[\nabla\mathbf{w}(\mathbf{x},t) : \nabla\mathbf{w}(\mathbf{x},t)] = -\nu\,\nabla_{\mathbf{z}}\cdot\nabla_{\mathbf{z}}\mathrm{tr}(\boldsymbol{\gamma}_0(\mathbf{z}))|_{\mathbf{z}=\mathbf{0}} = -3\nu\,\partial_{zz}\mathrm{tr}\boldsymbol{\gamma}_0(z)|_{z=0},$$

and with (26) and the differentiability of c_1,

$$\epsilon = -15\nu\,\partial_{zz}c_1(0) \overset{(30)}{=} 2\nu\int_0^\infty E(\kappa)\kappa^2 d\kappa. \tag{34}$$

The even extension $c_1(z) = c_1(-z)$ for $z \le 0$ in combination with the Fourier relation (27) results in a global differentiability of c_1 on $\mathbb{R}$ such that its odd derivatives vanish at $z = 0$. Therefore, the parameters k and ϵ describe the behavior of c_1 for small z by a Taylor expansion up to fourth order,

$$c_1(z) = \frac{2}{3}k - \frac{1}{30}\frac{\epsilon}{\nu}z^2 + \mathcal{O}(z^4).$$

3.3. Kolmogorov's energy spectrum. For the construction of the complete correlation function c_1 we need additional physical information about the flow, which can be gained from the energy spectrum E. The energy spectrum of isotropic turbulence was a well-studied topic of research during the last century (see references in [8, 10]). In particular, Kolmogorov's work (1941) was trendsetting. Based on dimensional analysis, he derived not only the characteristic ranges but also the typical run of the spectrum which agree with later physical concepts and experiments [1]. In the following, we briefly state Kolmogorov's 5/3-law and his hypothesis of local isotropy.

By (21), the energy spectrum depends on the wave number κ. Moreover, observing that turbulence is strongly driven by the large eddies, E can certainly be expected to be a function of the length l_T of the larger energy-containing eddies and the mean strain rate feeding the turbulence through direct interaction between mean flow and large eddies. Since turbulence is dissipative in the mean, it should additionally depend on ν and ϵ. Assuming a wide separation of energy (κ_e) and dissipation (κ_d) scales, Kolmogorov formulated the following.

Universal equilibrium theory (see [8]).

1. *If $\kappa_e < \kappa_d$, there exists a range for wave numbers $\kappa > \kappa_e$ in which the turbulence is in a statistical equilibrium and exclusively determined by dissipation ϵ and kinematic viscosity ν. This equilibrium state is universal; i.e., it occurs in isotropic as well as anisotropic turbulence. (This is the local isotropy hypothesis.)*
2. *If $\kappa_e \ll \kappa_d$, there exists an inertial subrange for wave numbers $\kappa_e < \kappa < \kappa_d$ in which the energy spectrum is just a function of dissipation ϵ and wave number κ.*

By means of dimensional analysis the first hypothesis leads to the Kolmogorov scales for length η, time t_K, and velocity v_K:

$$\eta = \left(\frac{\nu^3}{\epsilon}\right)^{1/4}, \qquad t_\mathrm{K} = \left(\frac{\nu}{\epsilon}\right)^{1/2}, \qquad v_\mathrm{K} = (\nu\,\epsilon)^{1/4},$$

with characteristic wave number $\kappa_\mathrm{K} = \eta^{-1} \approx \kappa_d$. The Kolmogorov length η is the smallest characteristic turbulence length. The second hypothesis yields Kolmogorov's 5/3-law,

$$E(\kappa) = C_\mathrm{K}\,\epsilon^{2/3}\kappa^{-5/3}, \qquad \kappa_e < \kappa < \kappa_d, \tag{35}$$

with Kolmogorov constant C_K. Here, $C_\mathrm{K} = 0.5$ is supposed to be an appropriate estimate according to the experiments of Yeung and Zhou [19].

The form of the energy spectrum sketched in Figure 2 is also designed by Batchelor and Proudman [3]. They derived that $E(\kappa) \sim \kappa^4$ for $\kappa \to 0$, whereas Heisenberg [9] deduced $E(\kappa) \sim \kappa^{-7}$ for $\kappa \to \infty$.

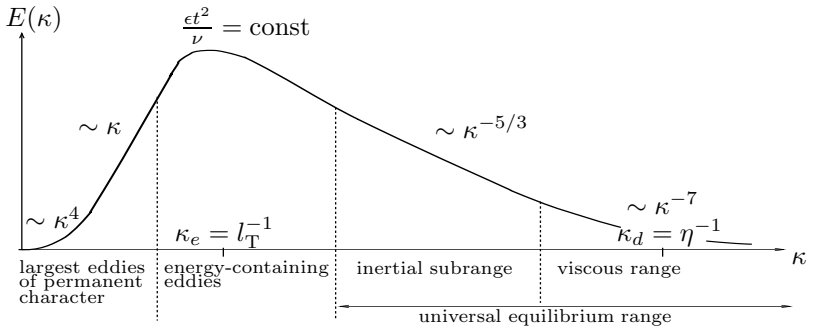


FIG. 2. *Sketch of energy spectrum for isotropic turbulence.*

3.4. Initial local correlations. Having provided the mathematical and physical fundamental ideas, we now model the initial correlation tensor of a local, homogeneous, $\mathcal{L}^2$-continuous, and differentiable Gaussian fluctuation field that satisfies the k-ϵ model and Kolmogorov's 5/3-law. For this purpose, we introduce an admissible underlying spectral energy distribution function.

MODEL FOR THE INITIAL LOCAL CORRELATION TENSOR. *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space. Let $(\mathbf{w}_{\mathbf{x},t}, (\mathbf{x},t) \in \mathbb{R}^3 \times \mathbb{R}_0^+)$ be the Gaussian velocity fluctuation field of an isotropic, homogeneous, and incompressible turbulent flow with $\mathbf{w}_{\mathbf{x},t} \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{P})$. Let kinetic energy k and dissipation rate ϵ be constant. Construct the initial correlation function $c_1 \in \mathcal{C}^\infty(\mathbb{R}_0^+)$,*

$$c_1(z) = \frac{2}{z^3} \int_0^\infty \frac{1}{\kappa}\, \partial_\kappa \left(\frac{E(\kappa)}{\kappa} \right) \sin(\kappa z)\, d\kappa,$$

by choosing $E \in \mathcal{C}^2(\mathbb{R}_0^+)$ as

$$E(\kappa) = \begin{cases} K\kappa_1^{-5/3} \sum_{j=4}^{6} a_j \, (\frac{\kappa}{\kappa_1})^j, & \kappa < \kappa_1, \\ K\kappa^{-5/3}, & \kappa_1 \le \kappa \le \kappa_2, \\ K\kappa_2^{-5/3} \sum_{j=7}^{9} b_j \, (\frac{\kappa}{\kappa_2})^{-j}, & \kappa > \kappa_2, \end{cases} \tag{36}$$

where κ_1 and κ_2 are implicitly given by

$$\int_0^\infty E(\kappa)\, d\kappa = k \quad \text{and} \quad \int_0^\infty E(\kappa)\kappa^2\, d\kappa = \frac{\epsilon}{2\nu}. \tag{37}$$

The parameters are fixed as $a_4 = 230/9$, $a_5 = -391/9$, $a_6 = 170/9$, $b_7 = 209/9$, $b_8 = -352/9$, $b_9 = 152/9$, $K = C_{\mathrm{K}}\, \epsilon^{2/3}$, $C_{\mathrm{K}} = 0.5$, and viscosity ν.

Then, $(\mathbf{w}_{\mathbf{x},t}, (\mathbf{x},t) \in \mathbb{R}^3 \times \mathbb{R}_0^+)$ is differentiable and fulfills the requirements of the Kolmogorov's 5/3-law as well as these of the k-ϵ model,

$$k = \frac{1}{2}\, \mathbb{E}[\mathbf{w}(\mathbf{x},t) \cdot \mathbf{w}(\mathbf{x},t)], \qquad \epsilon = \nu\, \mathbb{E}[\nabla \mathbf{w}(\mathbf{x},t) : \nabla \mathbf{w}(\mathbf{x},t)].$$

According to (27), the presented nonnegative function E satisfies the requirements on a spectral energy distribution function. Furthermore, it coincides with the run of Kolmogorov's energy spectrum (35). The differentiability of the local flow field, i.e., $(\ln(1+\kappa))^\alpha\, \kappa^2\, E(\kappa) \in \mathcal{L}^1(\mathbb{R}_0^+)$ for $\alpha > 3$ (cf. (32)), is ensured by the constructed

decay of $E(\kappa) \sim \kappa^{-7}$ for $\kappa \to \infty$. The information coming from the k-ϵ model is finally included in the defined moments of E on the basis of (33) and (34).

Alternatively, smoother variants of the piecewise composed energy spectrum are also imaginable for adapted regularity parameters a_i and b_j. However, E given by (36), (37) turns out to successfully satisfy our demands.

Summing up, the high Reynolds number flow under consideration is characterized by the flow quantities k, ϵ, ν in combination with the model for the energy spectrum, equations (36), (37). The flow quantities particularly determine the size of the characteristic energy ranges (energy bearing, inertial, and viscous) in Figure 2 by specifying κ_1 and κ_2 for fixed regularity parameters a_i and b_j. Note that, outside of the flow regime, the use of the k-ϵ turbulence model with the stated energy spectrum is questionable. However, the derivation of the initial local correlation tensor and the following general force concept require just the description of an appropriate spectral energy distribution that could alternatively be obtained by another turbulence model, e.g., LES.

3.5. Dynamics of local correlations. The dynamics of a local correlation tensor $\boldsymbol{\gamma}$ might be described by an advection equation according to the observation that the decay of the mean properties is rather slow with respect to the time scale of the fluctuating fine-scale structures

$$\partial_\varsigma \boldsymbol{\gamma}(\mathbf{z}, \varsigma) + \bar{\mathbf{u}} \cdot \nabla_{\mathbf{z}} \boldsymbol{\gamma}(\mathbf{z}, \varsigma) = 0.$$

Its solution,

$$\boldsymbol{\gamma}(\mathbf{z}, \varsigma) = \boldsymbol{\gamma}_0(\mathbf{z} - \bar{\mathbf{u}}\varsigma), \tag{38}$$

coincides with Taylor's hypothesis of frozen turbulence [17]; i.e., fluctuations arise due to so-called frozen turbulence patterns that are transported by the mean flow without changing their structure.

Equation (38) completes the construction of the local correlation tensor $\boldsymbol{\gamma}$. Consequently, we deal here locally with homogeneous, isotropic, incompressible turbulence moving with the mean flow velocity $\bar{\mathbf{u}}$, whose spectral energy distribution E fulfills the demands of the k-ϵ model as well as of Kolmogorov's universal equilibrium theory.

3.6. Construction of global turbulence. The Global-from-Local Assumption (12) prescribes the actual global fluctuation field $(\mathbf{u}'_{\mathbf{x},t}, (\mathbf{x}, t) \in \mathbb{R}^3 \times \mathbb{R}^+_0)$ on the basis of the family of underlying parameterized local fields $\{(\mathbf{w}^{\mathbf{y},\tau}_{\mathbf{x},t}, (\mathbf{x}, t) \in \mathbb{R}^3 \times \mathbb{R}^+_0), (\mathbf{y}, \tau) \in \mathbb{R}^3 \times \mathbb{R}^+_0\}$. However, so far the positive definiteness of its proposed correlation tensor $\boldsymbol{\Gamma}'$ is not proved, but it might be concluded from an explicit formulation of $\mathbf{u}'$.

Explicit Formulation of the Global Fluctuation Field. *Let the global fluctuation field* $(\mathbf{u}'_{\mathbf{x},t}, (\mathbf{x}, t) \in \mathbb{R}^3 \times \mathbb{R}^+_0)$ *be given as an Ito-integral over the family of the local fields* $\{(\mathbf{w}^{\mathbf{y},\tau}_{\mathbf{x},t}, (\mathbf{x}, t) \in \mathbb{R}^3 \times \mathbb{R}^+_0), (\mathbf{y}, \tau) \in \mathbb{R}^3 \times \mathbb{R}^+_0\}$:

$$\mathbf{u}'(\mathbf{x}, t) = \frac{1}{\sqrt{|M(\mathbf{x}, t)|}} \int_{M(\mathbf{x},t)} \mathbf{w}^{\mathbf{y},\tau}(\mathbf{x}, t)\, d\mathcal{W}_{\mathbf{y},\tau}, \tag{39}$$

$$M(\mathbf{x}, t) = \{(\mathbf{y}, \tau) \in \mathbb{R}^3 \times \mathbb{R}^+_0 \mid \|\mathbf{x} - \mathbf{y} - \bar{\mathbf{u}}(\mathbf{x}, t)(t - \tau)\|_2 \le l_{\mathrm{T}} \wedge |t - \tau| \le t_{\mathrm{T}}\},$$

where $(\mathcal{W}_{\mathbf{y},\tau}, (\mathbf{y}, t) \in \mathbb{R}^3 \times \mathbb{R}^+_0)$ *represents a Wiener process (Brownian motion). Then the field of* (39) *satisfies the probability distribution, expectation, and covariance structure of the averaging procedure* $\langle\cdot\rangle$ *in the Global-from-Local Assumption.*

The global field results from linear superpositions of joint Gaussians and is thus also Gaussian. Due to the permutability of expectation and integration with respect to space and time following from Fubini's theorem, it inherits the centered property from the local fields, so that the constructed $(\mathbf{u}'_{\mathbf{x},t}, (\mathbf{x},t) \in \mathbb{R}^3 \times \mathbb{R}_0^+)$ satisfies the definition of a turbulent Gaussian flow. Additionally, it is differentiable. Its correlation tensor reads

$$\begin{aligned}\boldsymbol{\Gamma}'(\mathbf{x_1},t_1,\mathbf{x_2},t_2) &= \frac{1}{\sqrt{|M(\mathbf{x_1},t_1)|\,|M(\mathbf{x_2},t_2)|}} \\ &\quad \cdot \mathbb{E}\left[\int_{M(\mathbf{x_1},t_1)} \mathbf{w}^{\mathbf{y_1},\tau_1}(\mathbf{x_1},t_1)\,d\mathcal{W}_{\mathbf{y_1},\tau_1} \otimes \int_{M(\mathbf{x_2},t_2)} \mathbf{w}^{\mathbf{y_2},\tau_2}(\mathbf{x_2},t_2)\,d\mathcal{W}_{\mathbf{y_2},\tau_2}\right]. \end{aligned} \tag{40}$$

By means of the Ito-calculus, the expectation of the dyadic product of the integrals can be expressed by

$$\begin{aligned}&\mathbb{E}\left[\int_{M(\mathbf{x_1},t_1)} \mathbf{w}^{\mathbf{y_1},\tau_1}(\mathbf{x_1},t_1)\,d\mathcal{W}_{\mathbf{y_1},\tau_1} \otimes \int_{M(\mathbf{x_2},t_2)} \mathbf{w}^{\mathbf{y_2},\tau_2}(\mathbf{x_2},t_2)\,d\mathcal{W}_{\mathbf{y_2},\tau_2}\right] \\ &\qquad = \mathbb{E}\left[\int_{M(\mathbf{x_1},t_1)\cap M(\mathbf{x_2},t_2)} \mathbf{w}^{\mathbf{y},\tau}(\mathbf{x_1},t_1) \otimes \mathbf{w}^{\mathbf{y},\tau}(\mathbf{x_2},t_2)\,d\mathbf{y}\,d\tau\right].\end{aligned}$$

Plugging this relation into (40), we obtain the proposed covariance of the Global-from-Local Assumption for the in general inhomogeneous, anisotropic global flow

$$\boldsymbol{\Gamma}'(\mathbf{x_1},t_1,\mathbf{x_2},t_2) = \frac{1}{\sqrt{|M(\mathbf{x_1},t_1)|\,|M(\mathbf{x_2},t_2)|}} \int_{M(\mathbf{x_1},t_1)\cap M(\mathbf{x_2},t_2)} \tilde{\boldsymbol{\gamma}}^{\mathbf{y},\tau}(\mathbf{x_1},t_1,\mathbf{x_2},t_2)\;d\mathbf{y}\;d\tau.$$

Due to its derivation from the random field of (39), $\boldsymbol{\Gamma}'$ is undoubtedly a positive definite function, which is necessary for the numerical realization of $\mathbf{u}'$.

The global quantities for kinetic energy $k_{\mathbf{u}'}$ and dissipation rate $\epsilon_{\mathbf{u}'}$ are the averages over a region M where the local, RANS-based quantities k and ϵ differ only slightly. This region is determined by means of the turbulent large-scale length l_T and time t_T and under regard of the advective influence of the mean flow in (38):

$$\begin{aligned}k_{\mathbf{u}'}(\mathbf{x},t) &= \frac{1}{|M(\mathbf{x},t)|}\int_{M(\mathbf{x},t)} k(\mathbf{y},\tau)\;d\mathbf{y}\;d\tau, \\ \epsilon_{\mathbf{u}'}(\mathbf{x},t) &= \frac{1}{|M(\mathbf{x},t)|}\int_{M(\mathbf{x},t)} \epsilon(\mathbf{y},\tau)\;d\mathbf{y}\;d\tau.\end{aligned}$$

In case of global homogeneity we achieve, in particular, the conformity of the global and local statistic quantities.

Despite weakening the conditions on the global turbulent flow, $\boldsymbol{\Gamma}'$ still keeps the correlation structure of the local fields. Let λ_T be the turbulent fine-scale length; then $\boldsymbol{\gamma}_0^{\mathbf{y},\tau}(\mathbf{x_1}-\mathbf{x_2}-\bar{\mathbf{u}}(\mathbf{y},\tau)(t_1-t_2)) \approx \mathbf{0}$ for $\|\mathbf{x_1}-\mathbf{x_2}-\bar{\mathbf{u}}(\mathbf{y},\tau)(t_1-t_2)\|_2 > \lambda_\mathrm{T}$, $(\mathbf{y},\tau) \in \mathbb{R}^3 \times \mathbb{R}_0^+$. Gluing the local correlations together yields $\boldsymbol{\Gamma}'(\mathbf{x_1},t_1,\mathbf{x_2},t_2) \approx \mathbf{0}$, even if $M(\mathbf{x_1},t_1) \cap M(\mathbf{x_2},t_2) \neq \emptyset$ as $\lambda_\mathrm{T} \ll l_\mathrm{T}$. Thus, $\boldsymbol{\Gamma}'$ states no wrong, absurd correlations.

4. General aerodynamic force concept. In the course of this section, the aerodynamic force that is acting on the fiber is modeled on top of the RANS-based

description for the turbulent flow. Thus, we introduce the mean relative velocity $\bar{\mathbf{v}}(s,t) = \bar{\mathbf{u}}(\mathbf{r}(s,t),t) - \partial_t \mathbf{r}(s,t)$. Then,

$$\tilde{\mathbf{f}}^{air}(s,t) = \mathbf{f}(\bar{\mathbf{v}}(s,t) + \mathbf{u}'(\mathbf{r}(s,t),t), \partial_s \mathbf{r}(s,t)) \tag{41}$$

prescribes a stochastic force $(\tilde{\mathbf{f}}^{air}_{s,t}, (s,t) \in [0,l]\times\mathbb{R}^+_0)$ as a (generally nonlinear) function on the derived global fluctuation field $\mathbf{u}'$. However, the efficient numerical handling of this inhomogeneous construct (41) seems to be hopeless because of its complexity. Thus, we follow the Global-from-Local ansatz once more.

Global-from-Local Force Concept. *Let $\mathbf{f} : \mathbb{R}^3 \times \mathbb{R}^2 \to \mathbb{R}^3$ be an arbitrarily chosen air drag model. Let $\{(\mathbf{g}^{\sigma,\tau}_{s,t}, (s,t) \in [0,l]\times\mathbb{R}^+_0), (\sigma,\tau) \in [0,l]\times\mathbb{R}^+_0\}$ be a family of homogeneous local aerodynamic forces that are imposed by local Gaussian velocity fluctuation fields on the locally linear fiber around the respective fiber points (σ,τ). Then, the global aerodynamic force is constructed as a Gaussian,*

$$\mathbf{f}^{air}(\mathbf{r}(.),s,t) = \langle \mathbf{g}^{\sigma,\tau}(s,t)\rangle_{N(\mathbf{r}(.),s,t)}, \tag{42}$$

with

$$\begin{aligned} N(\mathbf{r}(.),s,t) = \{&(\sigma,\tau) \in [0,l]\times\mathbb{R}^+_0 \,| \\ &\|\mathbf{r}(s,t) - \mathbf{r}(\sigma,\tau) - \bar{\mathbf{u}}(\mathbf{r}(s,t),t)(t-\tau)\|_2 \le l_{\mathrm{T}} \wedge |t-\tau| \le t_{\mathrm{T}}\}, \end{aligned} \tag{43}$$

and mean flow velocity $\bar{\mathbf{u}}$, turbulent large-scale length l_{T}, and time t_{T} as well as averaging brackets $\langle\cdot\rangle$ defined analogously to (12).

In analogy to the velocity fluctuations in section 3, this concept (42) realizes a Gaussian global aerodynamic force along the fiber on the basis of a family of homogeneous local random forces. Focusing on the construction of these forces, correlated local forces are deduced from the restriction of our derived Gaussian local velocity fields on the fiber in section 4.1. The proposed linearization approach of section 4.2 enables their approximation by Gaussian white noise with flow-dependent amplitude for a macroscopic description of the fiber. Therefore, $\mathcal{L}^2$- and $\mathcal{L}^\infty$-similarity estimates are stated in section 4.3. In section 4.4 we finally present the corresponding correlated global aerodynamic force and its uncorrelated asymptotic limit.

4.1. Correlated local force. Define the family $\{(\mathbf{g}^{\sigma,\tau}_{s,t}, (s,t) \in [0,l]\times\mathbb{R}^+_0), (\sigma,\tau) \in [0,l]\times\mathbb{R}^+_0\}$ of local aerodynamic forces by

$$\mathbf{g}^{\sigma,\tau}(s,t) = \mathbf{f}(\bar{\mathbf{v}}(\sigma,\tau) + \mathbf{w}^{\sigma,\tau}_f(s,t), \partial_s \mathbf{r}(\sigma,\tau)), \tag{44}$$

$$\mathbf{w}^{\sigma,\tau}_f(s,t) = \mathbf{w}^{\mathbf{r}(\sigma,\tau),\tau}(\mathbf{r}(\sigma,\tau) + (s-\sigma)\partial_s \mathbf{r}(\sigma,\tau) + (t-\tau)\partial_t \mathbf{r}(\sigma,\tau), t). \tag{45}$$

Presupposing a linear fiber around the point (σ,τ), the centered local Gaussian velocity fluctuation fields of section 3 keep their homogeneous correlation structure for their respective restrictions on the fiber in (45):

$$\begin{aligned} \mathbb{E}[\mathbf{w}^{\sigma,\tau}_f(s_1,t_1) \otimes \mathbf{w}^{\sigma,\tau}_f(s_2,t_2)] &= \boldsymbol{\gamma}^{\mathbf{r}(\sigma,\tau),\tau}_0((s_1-s_2)\partial_s \mathbf{r}(\sigma,\tau) - (t_1-t_2)\bar{\mathbf{v}}(\sigma,\tau)) \\ &= \boldsymbol{\gamma}^{\sigma,\tau}_f(s_1-s_2, t_1-t_2). \end{aligned} \tag{46}$$

Locally, for small spatial and temporal differences, the assumption of fiber linearity is reasonable, whereas for large ones, $\boldsymbol{\gamma}^{\sigma,\tau}_f \approx \mathbf{0}$ anyway due to the decay of the correlations. By means of the transformation theorem of random variables, the homogeneous property is handed on $\mathbf{g}^{\sigma,\tau}$ for all feasible drag models $\mathbf{f}$ in (44). Indeed, the chosen

drag model determines the probability distributions of $\mathbf{g}^{\sigma,\tau}$ that are in general not Gaussian. Averaging over the prescribed homogeneous local forces along the fiber in (42) results in a correlated global aerodynamic force that represents the turbulence effects on the fiber motion in (7). Be aware that the stated Global-from-Local Force Concept generates here a functional dependence between $\mathbf{f}^{air}$ and $\mathbf{r}$, so that the fiber dynamics is not modeled by a system of PDEs as in the deterministic flow case of (9).

4.2. Linearization approach. The numerical realization of the correlated Gaussian global aerodynamic force $\mathbf{f}^{air}$ depends crucially on the determination of the probability distributions of $\mathbf{g}^{\sigma,\tau}$, particularly on the computation of the integrals for expectation and covariance according to the definition of the averaging brackets $\langle\cdot\rangle$; cf. (12). The degree of difficulty is thereby mainly determined by the air drag model. For practical reasons, we hence propose a linearization ansatz for $\mathbf{g}^{\sigma,\tau}$ that yields Gaussian local forces

$$\begin{aligned}\mathbf{g}^{\sigma,\tau}(s,t) &= \mathbf{f}(\bar{\mathbf{v}}(\sigma,\tau)+\mathbf{w}_f^{\sigma,\tau}(s,t),\partial_s\mathbf{r}(\sigma,\tau))\\ &\approx \mathbf{f}(\bar{\mathbf{v}}(\sigma,\tau),\partial_s\mathbf{r}(\sigma,\tau))+\mathbf{L}^{\mathbf{f}}(\sigma,\tau)\ \mathbf{w}_f^{\sigma,\tau}(s,t)\\ &= \mathbf{g}_{cc}^{\sigma,\tau}(s,t),\end{aligned}\tag{47}$$

where the linear operator $\mathbf{L}^{\mathbf{f}}$ is induced by the air drag model $\mathbf{f}$. The finite-dimensional distributions of $\mathbf{g}_{cc}^{\sigma,\tau}$ are uniquely given by expectation and covariance,

$$\begin{aligned}\mathbb{E}[\mathbf{g}_{cc}^{\sigma,\tau}(s,t)] &= \mathbf{f}(\bar{\mathbf{v}}(\sigma,\tau),\partial_s\mathbf{r}(\sigma,\tau)) = \boldsymbol{\mu}^{\sigma,\tau},\\ \mathrm{Cov}(\mathbf{g}_{cc}^{\sigma,\tau}(s_1,t_1),\mathbf{g}_{cc}^{\sigma,\tau}(s_2,t_2)) &= \mathbf{L}^{\mathbf{f}}(\sigma,\tau)\ \boldsymbol{\gamma}_f^{\sigma,\tau}(s_1-s_2,t_1-t_2)\ (\mathbf{L}^{\mathbf{f}}(\sigma,\tau))^t\\ &= \boldsymbol{\Gamma}_{\mathbf{g},cc}^{\sigma,\tau}(s_1-s_2,t_1-t_2),\end{aligned}\tag{48}$$

whose evaluation is directly deduced from the centered, homogeneous Gaussian $\mathbf{w}_f^{\sigma,\tau}$; see (46).

4.3. Limit to uncorrelated local force. The correlated Gaussian local forces $\mathbf{g}_{cc}^{\sigma,\tau}$ contain all turbulent coherences explicitly in their covariance function $\boldsymbol{\Gamma}_{\mathbf{g},cc}^{\sigma,\tau}$: $[0,l]\times\mathbb{R}_0^+\to\mathbb{R}^{3\times3}$ of (48). Alternatively, uncorrelated generalized Gaussian local forces $\mathbf{g}_{uc}^{\sigma,\tau}$ might be introduced whose flow-dependent amplitude represents the mean turbulent coherences. Their covariance functions read

$$\boldsymbol{\Gamma}_{\mathbf{g},uc}^{\sigma,\tau}(s,t)=\int_{\mathbb{R}^2}\boldsymbol{\Gamma}_{\mathbf{g},cc}^{\sigma,\tau}(\xi,\varsigma)\ d\xi\,d\varsigma\ \ \delta_0(s)\,\delta_0(t),\tag{49}$$

with the real one-dimensional Dirac function δ_0. If their effects, i.e., their correlations, are compared on a macroscopic fiber scale that includes the whole covariance structure of $\mathbf{g}_{cc}^{\sigma,\tau}$, the family of the uncorrelated forces $\mathbf{g}_{uc}^{\sigma,\tau}$ is a good approximation for that of the correlated $\mathbf{g}_{cc}^{\sigma,\tau}$. In the following, $\mathcal{L}^2$- and $\mathcal{L}^\infty$-estimates for their similarity take center stage.

Define the family $\{((\mathbf{g}_{uc}^{\sigma,\tau})_{s,t},(s,t)\in[0,l]\times\mathbb{R}_0^+),(\sigma,\tau)\in[0,l]\times\mathbb{R}_0^+\}$ of local uncorrelated aerodynamic forces by

$$\mathbf{g}_{uc}^{\sigma,\tau}(s,t)=\mathbf{f}(\bar{\mathbf{v}}(\sigma,\tau),\partial_s\mathbf{r}(\sigma,\tau))+\mathbf{L}^{\mathbf{f}}(\sigma,\tau)\ \mathbf{z}^{\sigma,\tau}(s,t),\tag{50}$$

$$\mathbf{z}^{\sigma,\tau}(s,t)=\mathbf{D}^{\sigma,\tau}\,\mathbf{p}^{\sigma,\tau}(s,t).\tag{51}$$

The centered uncorrelated local velocity fluctuation fields $(\mathbf{z}_{s,t}^{\sigma,\tau},(s,t)\in[0,l]\times\mathbb{R}_0^+)$ along the fiber are particularly given by Gaussian white noise $(\mathbf{p}_{s,t}^{\sigma,\tau},(s,t)\in[0,l]\times\mathbb{R}_0^+)$

with flow-dependent amplitude

$$\mathbf{D}^{\sigma,\tau} = \sqrt{\int_{\mathbb{R}^2} \boldsymbol{\gamma}_f^{\sigma,\tau}(\xi,\varsigma)\ d\xi\, d\varsigma} \tag{52}$$

that contains the integral correlations of $\mathbf{w}_f^{\sigma,\tau}$. The existence of $\mathbf{D}^{\sigma,\tau}$ presupposes the linear independence of the fiber tangent $\partial_s\mathbf{r}(\sigma,\tau)$ and the relative velocity $\bar{\mathbf{v}}(\sigma,\tau)$, as can be concluded from the definition of $\boldsymbol{\gamma}_f^{\sigma,\tau}$ in (46). The velocity correlations are then described by

$$\begin{aligned}\mathbb{E}[\mathbf{z}^{\sigma,\tau}(s_1,t_1)\otimes\mathbf{z}^{\sigma,\tau}(s_2,t_2)] &= (\mathbf{D}^{\sigma,\tau})^2\,\delta_0(s_1-s_2)\,\delta_0(t_1-t_2)\\ &= \boldsymbol{\delta}_f^{\sigma,\varsigma}(s_1-s_2,t_1-t_2),\end{aligned} \tag{53}$$

with the real one-dimensional Dirac function δ_0. In this sense, $\boldsymbol{\delta}_f^{\sigma,\tau}$ is the uncorrelated analogue to $\boldsymbol{\gamma}_f^{\sigma,\tau}$ and induces the desired integral dependence (49) between $\boldsymbol{\Gamma}_{\mathbf{g},uc}^{\sigma,\tau}$ and $\boldsymbol{\Gamma}_{\mathbf{g},cc}^{\sigma,\tau}$ due to the linear construction in (50) and (47).

Focusing on an arbitrarily chosen fiber point (σ,τ), we skip the superscripts of the quantities in the following and deduce a formulation for the force amplitude $\mathbf{D}$ in terms of the manageable energy spectrum E of (21). Therefore, we presume the linear independence of fiber tangent $\mathbf{t} = \partial_s\mathbf{r}$ and mean relative velocity $\bar{\mathbf{v}}$, so that they induce the intuitive choice of a right-hand orthonormal basis, i.e., $\mathbf{t}$, $\mathbf{n} = (\bar{\mathbf{v}} - (\bar{\mathbf{v}}\cdot\mathbf{t})\mathbf{t})/\|\bar{\mathbf{v}} - (\bar{\mathbf{v}}\cdot\mathbf{t})\mathbf{t}\|_2$, $\mathbf{b} = \mathbf{t}\times\mathbf{n}$.

Relation between Local Fiber Correlations and Spectral Quantities. *Assume* $\mathbf{t}$ *and* $\bar{\mathbf{v}}$ *to be linearly independent. Let* $\boldsymbol{\gamma}_f(\xi,\varsigma) = \boldsymbol{\gamma}_0(\xi\mathbf{t} - \varsigma\bar{\mathbf{v}})$, $(\xi,\varsigma)\in\mathbb{R}^2$, *be the local velocity correlation tensor along the fiber. Then, its negative Fourier transform* $\mathbf{m} = \mathcal{F}_{\boldsymbol{\gamma}_f}$ *is expressed by the spectral density* $\mathbf{M}$ *of* (20)*:*

$$\mathbf{m}(\lambda_1,\lambda_2) = \int_{\mathbb{R}^3} \mathbf{M}(\boldsymbol{\kappa})\ \delta_0(\lambda_1 - \mathbf{t}\cdot\boldsymbol{\kappa})\ \delta_0(\lambda_2 + \bar{\mathbf{v}}\cdot\boldsymbol{\kappa})\ d\boldsymbol{\kappa}. \tag{54}$$

The integral correlations are prescribed by $\mathbf{m}(0,0) = (2\pi)^{-2}\int\boldsymbol{\gamma}_f(\xi,\varsigma)\,d\xi\,d\varsigma = \mathcal{F}_{\boldsymbol{\delta}_f}$. *In particular,*

$$\mathbf{m}(0,0) = \frac{1}{2\pi\bar{v}_\mathrm{n}}\int_0^\infty \frac{E(\kappa)}{\kappa^2}d\kappa\ \mathbf{P}_{\mathbf{t},\mathbf{n}}, \tag{55}$$

where $\mathbf{P}_{\mathbf{t},\mathbf{n}} := \mathbf{t}\otimes\mathbf{t} + \mathbf{n}\otimes\mathbf{n}$ *denotes the projector onto the plane spanned by* $\mathbf{t}$ *and* $\mathbf{n}$*, and* $\bar{v}_\mathrm{n} := \bar{\mathbf{v}}\cdot\mathbf{n}$.

Inserting the Fourier relation (20) for $\boldsymbol{\gamma}_0$ and $\mathbf{M}$ into the definition of $\mathbf{m}$ and evaluating the two-dimensional integral over the exponential function yields relation (54). Using isotropy and incompressibility of $\mathbf{M}$, (22) and (25), the dependence on the energy spectrum follows:

$$\mathbf{m}(\lambda_1,\lambda_2) = \frac{1}{4\pi}\int_{\mathbb{R}^3}\frac{E(\kappa)}{\kappa^2}\left(\mathbf{I} - \frac{1}{\kappa^2}\boldsymbol{\kappa}\otimes\boldsymbol{\kappa}\right)\ \delta_0(\lambda_1 - \mathbf{t}\cdot\boldsymbol{\kappa})\ \delta_0(\lambda_2 + \bar{\mathbf{v}}\cdot\boldsymbol{\kappa})\ d\boldsymbol{\kappa}, \tag{56}$$

with $\kappa = \|\boldsymbol{\kappa}\|_2$. Consider the matrix $\mathbf{m}^{\mathbf{t},\mathbf{n},\mathbf{b}}$ that represents the tensor $\mathbf{m}$ in the $(\mathbf{t},\bar{\mathbf{v}})$-induced basis, and substitute $\mathbf{t}\cdot\boldsymbol{\kappa} = \kappa_\mathrm{t}$, $\mathbf{n}\cdot\boldsymbol{\kappa} = \kappa_\mathrm{n}$, and $\mathbf{b}\cdot\boldsymbol{\kappa} = \kappa_\mathrm{b}$. Integration over κ_t and κ_n then gives $\mathbf{m}^{\mathbf{t},\mathbf{n},\mathbf{b}}(0,0) = \int_0^\infty E(\kappa)/\kappa^2 d\kappa/(2\pi\bar{v}_\mathrm{n})\ \mathrm{diag}(1,1,0)$ and, with the spectral theorem on the eigenvalues, the invariant form (55) of the tensor.

RELATION BETWEEN FORCE AMPLITUDE AND ENERGY SPECTRUM. *Let* $\mathbf{D}$ *be the force amplitude and* E *the energy spectrum corresponding to a homogeneous, isotropic, and incompressible local velocity fluctuation field. Then the following relation holds:*

$$\mathbf{D} = \sqrt{\frac{2\pi}{\bar{v}_{\mathrm{n}}} \int_0^\infty \frac{E(\kappa)}{\kappa^2} d\kappa}\ \mathbf{P}_{\mathbf{t},\mathbf{n}}. \tag{57}$$

Relation (57) results directly from (52) and (55). It allows the interesting observation that the uncorrelated local velocity fluctuation field $\mathbf{z}$ of (51) has no component in the binormal direction $\mathbf{b}$ of the fiber. The reason for this behavior is the incompressibility of the underlying flow field, since

$$\mathbf{P_b} \int_{\mathbb{R}^2} \boldsymbol{\gamma}_f(\xi,\varsigma)\ d\xi\, d\varsigma = \mathbf{P_b} \int_0^\infty z c_2(z)\, dz = \mathbf{P_b} \int_0^\infty e_1(\kappa)\, d\kappa = \mathbf{0}, \qquad \mathbf{P_b} = \mathbf{b} \otimes \mathbf{b},$$

due to (25) or, respectively, to (24) and partial integration.

Proceeding with the general similarity estimates for the correlated and an uncorrelated local force, it is sufficient to study the effects of the centered local velocity fluctuation fields on a macroscopic fiber scale because of their linear relation, (47), (50). For this purpose, we consider the respective macroscopic velocity fields that are gained from spatially and temporally smoothing along the fiber, and compare their correlation tensors.

Let $\mathbf{w}_f$ and $\mathbf{z}$ be a correlated and an uncorrelated local velocity fluctuation field. The introduction of the normalized spatial and temporal smoothing functions $G_{\boldsymbol{\alpha}} : \mathbb{R}^2 \to \mathbb{R}$, $\boldsymbol{\alpha} = (\alpha_s, \alpha_t) \in (\mathbb{R}_0^+)^2$, enables then the definition of two families of macroscopic velocity fields along the fiber:

$$\mathbf{W}_{\boldsymbol{\alpha}}(s,t) = \int G_{\boldsymbol{\alpha}}(s-\phi, t-\psi)\ \mathbf{w}_f(\phi,\psi)\ d\psi\ d\phi, \tag{58}$$

$$\mathbf{Z}_{\boldsymbol{\alpha}}(s,t) = \int G_{\boldsymbol{\alpha}}(s-\phi, t-\psi)\ \ \mathbf{z}(\phi,\psi)\ \ d\psi\ d\phi, \tag{59}$$

with their correlation tensors

$$\boldsymbol{\Gamma}_{\mathbf{W}_{\boldsymbol{\alpha}}}(\xi,\varsigma) = \int H_{\boldsymbol{\alpha}}(\xi-\phi, \varsigma-\psi)\ \boldsymbol{\gamma}_f(\phi,\psi)\ d\psi\, d\phi, \tag{60}$$

$$\boldsymbol{\Gamma}_{\mathbf{Z}_{\boldsymbol{\alpha}}}(\xi,\varsigma) = \int H_{\boldsymbol{\alpha}}(\xi-\phi, \varsigma-\psi)\ \boldsymbol{\delta}_f(\phi,\psi)\ d\psi\, d\phi, \tag{61}$$

where $H_{\boldsymbol{\alpha}}(\xi,\varsigma) = \int G_{\boldsymbol{\alpha}}(\xi-\phi, \varsigma-\psi)\ G_{\boldsymbol{\alpha}}(\phi,\psi)\ d\psi\, d\phi$ are also normalized smoothing functions. Taking the convolution keeps the properties of the local fields so that $\mathbf{W}_{\boldsymbol{\alpha}}$ and $\mathbf{Z}_{\boldsymbol{\alpha}}$ are Gaussian, centered, and homogeneous for all smoothing parameters $\boldsymbol{\alpha} \in (\mathbb{R}_0^+)^2$. The Gaussian property follows thereby directly from the linear superposition of joint Gaussians. The centered and homogeneous properties are deduced by using the permutability of expectation and integration according to Fubini's theorem.

CHOICE OF SMOOTHING OPERATORS. *Let the smoothing functions* $G_{\boldsymbol{\alpha}} : \mathbb{R}^2 \to \mathbb{R}$, $\boldsymbol{\alpha} = (\alpha_s, \alpha_t) \in (\mathbb{R}_0^+)^2$, *be defined as products of spatial and temporal characteristic functions*

$$G_{\boldsymbol{\alpha}}(\xi,\varsigma) = \alpha_s \alpha_t\ \chi_{[\frac{-1}{2\alpha_s}, \frac{1}{2\alpha_s}]}(\xi)\, \chi_{[\frac{-1}{2\alpha_t}, \frac{1}{2\alpha_t}]}(\varsigma). \tag{62}$$

Then, $H_{\boldsymbol{\alpha}} : \mathbb{R}^2 \to \mathbb{R}$ are given by the products of the hat functions, and their respective negative Fourier transforms are

$$H_{\boldsymbol{\alpha}}(\xi,\varsigma) = \alpha_s\,\alpha_t\ (1-|\alpha_s\,\xi|)\,(1-|\alpha_t\,\varsigma|)\ \chi_{[\frac{-1}{\alpha_s},\frac{1}{\alpha_s}]}(\xi)\,\chi_{[\frac{-1}{\alpha_t},\frac{1}{\alpha_t}]}(\varsigma),$$

$$\mathcal{F}_{H_{\boldsymbol{\alpha}}}(\kappa_1,\kappa_2) = \mathcal{F}_{H_{\mathbf{1}}}\left(\frac{\kappa_1}{\alpha_s},\frac{\kappa_2}{\alpha_t}\right) = \frac{1}{\pi^2}\ \frac{1-\cos(\kappa_1/\alpha_s)}{(\kappa_1/\alpha_s)^2}\ \frac{1-\cos(\kappa_2/\alpha_t)}{(\kappa_2/\alpha_t)^2}. \tag{63}$$

The relation between $\mathcal{F}_{H_{\boldsymbol{\alpha}}}$ and $\mathcal{F}_{H_{\mathbf{1}}}$ results directly from their definition by using $H_{\boldsymbol{\alpha}}(\xi,\varsigma) = \alpha_s\,\alpha_t\,H_{\mathbf{1}}(\alpha_s\,\xi,\alpha_t\,\varsigma)$.

The derivation of the similarity estimates depends decisively on the behavior of the symmetric, nonnegative, differentiable function $\mathcal{E} : \mathbb{R}^2 \to \mathbb{R}_0^+$ that is defined by means of the energy spectrum E,

$$\mathcal{E}(\kappa_1,\kappa_2) := \int_{\mathbb{R}} \frac{E(\|(\kappa_1,\kappa_2,l)\|_2)}{(\kappa_1,\kappa_2,l)^2}\ dl. \tag{64}$$

It is radially decaying with maximum in the origin, i.e., $\max_{\boldsymbol{\kappa}} \mathcal{E}(\kappa_1,\kappa_2) = \mathcal{E}(0,0)$ and $g(\kappa) := \mathcal{E}(\kappa,a\kappa), \kappa \in \mathbb{R}_0^+$, strictly monotonically decreasing for $a \in \mathbb{R}$.

Similarity Estimates. *Choose the smoothing functions $G_{\boldsymbol{\alpha}} : \mathbb{R}^2 \to \mathbb{R}$, $\boldsymbol{\alpha} = (\alpha_s,\alpha_t) \in (\mathbb{R}_0^+)^2$, of* (62) *for the definition of the families of macroscopic velocity fields according to* (58) *and* (59)*. Then the following estimates hold:*

$\mathcal{L}^2$-similarity:

$$\begin{aligned}\mathcal{I}_{\mathcal{L}^2} &:= \|\boldsymbol{\Gamma}_{\mathbf{W}_{\boldsymbol{\alpha}}} - \boldsymbol{\Gamma}_{\mathbf{Z}_{\boldsymbol{\alpha}}}\|_{\mathcal{L}^2(l^2(\mathbb{R}^2))} \\ &\le \frac{\sqrt{\alpha_s\,\alpha_t}}{\sqrt{6}\,\pi\,\bar{v}_{\mathrm{n}}}\sqrt{\mathcal{S}^2\left(\alpha_s^2\left(1+\frac{\bar{v}_{\mathrm{t}}^2}{\bar{v}_{\mathrm{n}}^2}\right)+\frac{\alpha_t^2}{\bar{v}_{\mathrm{n}}^2}\right)+\frac{8\mathcal{E}_0^2}{3\pi}\left(\alpha_s^3+\frac{\alpha_t^3}{(\bar{v}_{\mathrm{n}}+|\bar{v}_{\mathrm{t}}|)^3}\right)}.\end{aligned} \tag{65}$$

$\mathcal{L}^\infty$-similarity:

$$\begin{aligned}\mathcal{I}_{\mathcal{L}^\infty} &:= \|\boldsymbol{\Gamma}_{\mathbf{W}_{\boldsymbol{\alpha}}} - \boldsymbol{\Gamma}_{\mathbf{Z}_{\boldsymbol{\alpha}}}\|_{\mathcal{L}^\infty(l^2(\mathbb{R}^2))} \\ &\le \frac{\sqrt{2}\,\alpha_s\,\alpha_t}{\pi^2\,\bar{v}_{\mathrm{n}}}\left[\mathcal{S}\left(\alpha_s\left(1+\frac{\bar{v}_{\mathrm{t}}}{\bar{v}_{\mathrm{n}}}\right)\left(\frac{c}{2}+\ln\left(\frac{1}{\alpha_s}\right)\right)+\frac{\alpha_t}{\bar{v}_{\mathrm{n}}}\left(\frac{c}{2}+\ln\left(\frac{\bar{v}_{\mathrm{n}}+|\bar{v}_{\mathrm{t}}|}{\alpha_t}\right)\right)\right)\right. \\ &\qquad \left.+\mathcal{E}_0\left(\alpha_s+\frac{\alpha_t}{\bar{v}_{\mathrm{n}}+|\bar{v}_{\mathrm{t}}|}\right)\right],\end{aligned} \tag{66}$$

where

$$\|\boldsymbol{\Gamma}\|_{\mathcal{L}^2(l^2(\mathbb{R}^2))} := \left(\int_{\mathbb{R}^2} \boldsymbol{\Gamma}(\xi,\varsigma) : \boldsymbol{\Gamma}(\xi,\varsigma)\ d\xi\,d\varsigma\right)^{1/2},$$

$$\|\boldsymbol{\Gamma}\|_{\mathcal{L}^\infty(l^2(\mathbb{R}^2))} := \sup_{(\xi,\varsigma)\in\mathbb{R}^2}(\boldsymbol{\Gamma}(\xi,\varsigma) : \boldsymbol{\Gamma}(\xi,\varsigma))^{1/2}.$$

The quantities $\mathcal{E}_0 = \mathcal{E}(0,0)$ and $\mathcal{S} = \sup_{\boldsymbol{\kappa}\in[0,1]^2}\|\nabla_{\boldsymbol{\kappa}}\mathcal{E}(\kappa_1,\kappa_2)\|_2$ are defined by the energy moment of (64)*. Moreover, $\bar{v}_{\mathrm{t}} = \bar{\mathbf{v}}\cdot\mathbf{t}$, $\bar{v}_{\mathrm{n}} = \bar{\mathbf{v}}\cdot\mathbf{n}$, and $c = \int_0^1(1-\cos\iota)/\iota\,d\iota$.*

Proof. (1) $\mathcal{L}^2$-similarity. The norm in $\mathcal{L}^2(l^2(\mathbb{R}^2))$ is conserved under the Fourier transformation according to the Plancherel theorem as the operator : induces a scalar product in the l^2-space. Using the fact that the Fourier transform of a convolution equals the product of the respective Fourier transforms then gives

$$\begin{aligned}\|\boldsymbol{\Gamma}_{\mathbf{W}_{\boldsymbol{\alpha}}} - \boldsymbol{\Gamma}_{\mathbf{Z}_{\boldsymbol{\alpha}}}\|^2_{\mathcal{L}^2(l^2(\mathbb{R}^2))} &= (2\pi)^2\,\|(\mathcal{F}_{\boldsymbol{\gamma}_f} - \mathcal{F}_{\boldsymbol{\delta}_f})\,\mathcal{F}_{H_{\boldsymbol{\alpha}}}\|^2_{\mathcal{L}^2(l^2(\mathbb{R}^2))} \\ &= (2\pi)^2\int_{\mathbb{R}^2}\|\mathbf{m}(\lambda_1,\lambda_2)-\mathbf{m}(0,0)\|^2_{l^2}\,\mathcal{F}^2_{H_{\boldsymbol{\alpha}}}(\lambda_1,\lambda_2)\ d\lambda_1 d\lambda_2.\end{aligned} \tag{67}$$

With (56) and

$$\left(\mathbf{I}-\frac{1}{\boldsymbol{\kappa}^2}\boldsymbol{\kappa}\otimes\boldsymbol{\kappa}\right):\left(\mathbf{I}-\frac{1}{\boldsymbol{\iota}^2}\boldsymbol{\iota}\otimes\boldsymbol{\iota}\right)=1+\frac{(\boldsymbol{\kappa}\cdot\boldsymbol{\iota})^2}{\boldsymbol{\kappa}^2\boldsymbol{\iota}^2}\leq 2, \tag{68}$$

we obtain

$$\begin{aligned}\|\mathbf{m}(\lambda_1,\lambda_2)-\mathbf{m}(0,0)\|_{l^2}^2=&\frac{1}{(4\pi)^2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{E(\|\boldsymbol{\kappa}\|_2)}{\boldsymbol{\kappa}^2}\frac{E(\|\boldsymbol{\iota}\|_2)}{\boldsymbol{\iota}^2}\left(1+\frac{(\boldsymbol{\kappa}\cdot\boldsymbol{\iota})^2}{\boldsymbol{\kappa}^2\boldsymbol{\iota}^2}\right)\\&\cdot(\delta_0(\lambda_1-\mathbf{t}\cdot\boldsymbol{\kappa})\delta_0(\lambda_2+\bar{\mathbf{v}}\cdot\boldsymbol{\kappa})-\delta_0(\mathbf{t}\cdot\boldsymbol{\kappa})\delta_0(\bar{\mathbf{v}}\cdot\boldsymbol{\kappa}))\\&\cdot(\delta_0(\lambda_1-\mathbf{t}\cdot\boldsymbol{\iota})\delta_0(\lambda_2+\bar{\mathbf{v}}\cdot\boldsymbol{\iota})-\delta_0(\mathbf{t}\cdot\boldsymbol{\iota})\delta_0(\bar{\mathbf{v}}\cdot\boldsymbol{\iota}))\;d\boldsymbol{\kappa}\,d\boldsymbol{\iota}.\end{aligned}$$

Inserting this relation into (67) and integrating over λ_1 and λ_2 cancels two Dirac functions. The other two vanish after choosing the $(\mathbf{t},\bar{\mathbf{v}})$-induced basis. Applying (64) and (68) yields, with $\bar{v}_\mathrm{t}=\bar{\mathbf{v}}\cdot\mathbf{t}$ and $\bar{v}_\mathrm{n}=\bar{\mathbf{v}}\cdot\mathbf{n}=\|\bar{\mathbf{v}}-(\bar{\mathbf{v}}\cdot\mathbf{t})\mathbf{t}\|_2>0$,

$$\begin{aligned}\mathcal{I}_{\mathcal{L}^2}^2\leq&\frac{1}{2\bar{v}_\mathrm{n}}\left[\int_{\mathbb{R}^2}\left(\mathcal{E}^2(\kappa_1,\kappa_2)-2\mathcal{E}(\kappa_1,\kappa_2)\mathcal{E}(0,0)\right)\;\mathcal{F}_{H_{\boldsymbol{\alpha}}}^2(\kappa_1,-(\bar{v}_\mathrm{t}\kappa_1+\bar{v}_\mathrm{n}\kappa_2))\;d\kappa_1\,d\kappa_2\right.\\&\left.+\int_{\mathbb{R}^2}\frac{1}{\bar{v}_\mathrm{n}}\mathcal{E}^2(0,0)\;\mathcal{F}_{H_{\boldsymbol{\alpha}}}^2(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2\right]\\=&\frac{\alpha_s\alpha_t}{2\bar{v}_\mathrm{n}^2}\int_{\mathbb{R}^2}\left(\mathcal{E}\left(\alpha_s\iota_1,\frac{1}{\bar{v}_\mathrm{n}}(\alpha_t\iota_2-\alpha_s\bar{v}_\mathrm{t}\iota_1)\right)-\mathcal{E}(0,0)\right)^2\mathcal{F}_{H_1}^2(\iota_1,\iota_2)\,d\iota_1\,d\iota_2.\end{aligned} \tag{69}$$

The latter calculation is based on the substitution $\kappa_1=\alpha_s\iota_1$, $\bar{v}_\mathrm{t}\kappa_1+\bar{v}_\mathrm{n}\kappa_2=\alpha_t\iota_2$, $\lambda_1=\alpha_s\iota_1$, $\lambda_2=\alpha_t\iota_2$ and on the properties of the even smoothing functions (63). Positivity and radial decay of $\mathcal{E}$ induce the splitting of the integral in (69),

$$\mathcal{I}_{\mathcal{L}^2}^2\leq\frac{\alpha_s\alpha_t}{2\bar{v}_\mathrm{n}^2}\;(J_U+J_{\mathbb{R}^2\setminus U}), \tag{70}$$

with regard to the domain decomposition $\mathbb{R}^2=U\cup(\mathbb{R}^2\setminus U)$, where

$$U:=\{(\iota_1,\iota_2)\;|\;\iota_1\in[-\alpha_s^{-1},\alpha_s^{-1}]\;\wedge\;\iota_2\in[-\alpha_t^{-1}(\bar{v}_\mathrm{n}+|\bar{v}_\mathrm{t}|),\alpha_t^{-1}(\bar{v}_\mathrm{n}+|\bar{v}_\mathrm{t}|)]\}.$$

The energy difference in J_U can be estimated by means of its differentiability, for $(\mathcal{E}(\alpha_s\iota_1,\bar{v}_\mathrm{n}^{-1}(\alpha_t\iota_2-\alpha_s\bar{v}_\mathrm{t}\iota_1))-\mathcal{E}(0,0))^2\leq S^2\,\|(\alpha_s\iota_1,\bar{v}_\mathrm{n}^{-1}(\alpha_t\iota_2-\alpha_s\bar{v}_\mathrm{t}\iota_1))\|_2^2$. Thus,

$$J_U\leq\mathcal{S}^2\int_U\left(\alpha_s^2\left(1+\frac{\bar{v}_\mathrm{t}^2}{\bar{v}_\mathrm{n}^2}\right)\iota_1^2+\frac{\alpha_t^2}{\bar{v}_\mathrm{n}^2}\iota_2^2-2\alpha_s\alpha_t\frac{\bar{v}_\mathrm{t}}{\bar{v}_\mathrm{n}^2}\iota_1\iota_2\right)\;\mathcal{F}_{H_1}^2(\iota_1,\iota_2)\;d\iota_1\,d\iota_2.$$

The odd term vanishes by the integration. Using the equivalence of the integrand in the four quadrants, we obtain with (63)

$$\begin{aligned}J_U\leq&\,4\mathcal{S}^2\int_0^{\alpha_t^{-1}(\bar{v}_\mathrm{n}+|\bar{v}_\mathrm{t}|)}\int_0^{\alpha_s^{-1}}\left(\alpha_s^2\left(1+\frac{\bar{v}_\mathrm{t}^2}{\bar{v}_\mathrm{n}^2}\right)\iota_1^2+\frac{\alpha_t^2}{\bar{v}_\mathrm{n}^2}\iota_2^2\right)\;\mathcal{F}_{H_1}^2(\iota_1,\iota_2)\;d\iota_1\,d\iota_2\\\leq&\,\frac{\mathcal{S}^2}{3\pi^2}\left(\alpha_s^2\left(1+\frac{\bar{v}_\mathrm{t}^2}{\bar{v}_\mathrm{n}^2}\right)+\frac{\alpha_t^2}{\bar{v}_\mathrm{n}^2}\right),\end{aligned} \tag{71}$$

where the compact integration domain is replaced by $(\mathbb{R}_0^+)^2$. On the other hand, the energy difference in $J_{\mathbb{R}^2\setminus U}$ can be estimated by its maximum $\mathcal{E}_0$ due to the strict decay

of $\mathcal{E}$. The equivalence of the integrand in the quadrants leads then to

$$J_{\mathbb{R}^2\setminus U} \le 4\mathcal{E}_0^2 \left(\int_0^\infty \int_{\alpha_s^{-1}}^\infty \mathcal{F}_{H_1}^2(\iota_1,\iota_2)\, d\iota_1\, d\iota_2 + \int_{\alpha_t^{-1}(\bar v_{\mathrm{n}}+|\bar v_{\mathrm{t}}|)}^\infty \int_0^{\alpha_s^{-1}} \mathcal{F}_{H_1}^2(\iota_1,\iota_2)\, d\iota_1\, d\iota_2 \right)$$

(72)

$$\le \frac{8\mathcal{E}_0^2}{9\pi^3} \left(\alpha_s^3 + \frac{\alpha_t^3}{(\bar v_{\mathrm{n}}+|\bar v_{\mathrm{t}}|)^3} \right),$$

where the integration interval of ι_1 in the second summand is replaced by $\mathbb{R}_0^+$. Inserting (71) and (72) into (70) yields the $\mathcal{L}^2$-estimate.

(2) $\mathcal{L}^\infty$-similarity. The $\mathcal{L}^\infty$-estimate is derived in analogy to the $\mathcal{L}^2$-estimate. Consider therefore

$$\begin{aligned}
\|(\mathbf{\Gamma}_{\mathbf{W}_{\boldsymbol\alpha}} - \mathbf{\Gamma}_{\mathbf{Z}_{\boldsymbol\alpha}})(\boldsymbol\sigma)\|_{l^2}^2 &= \left\| \int_{\mathbb{R}^2} e^{i\boldsymbol\lambda\cdot\boldsymbol\sigma} \left(\mathbf{m}(\boldsymbol\lambda) - \mathbf{m}(\mathbf{0})\right) \mathcal{F}_{H_{\boldsymbol\alpha}}(\boldsymbol\lambda)\, d\boldsymbol\lambda \right\|_{l^2}^2 \\
&= \frac{1}{(4\pi)^2} \int_{\mathbb{R}^2}\int_{\mathbb{R}^2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} e^{i(\boldsymbol\lambda+\boldsymbol\mu)\cdot\boldsymbol\sigma}\, \mathcal{F}_{H_{\boldsymbol\alpha}}(\boldsymbol\lambda)\, \mathcal{F}_{H_{\boldsymbol\alpha}}(\boldsymbol\mu) \frac{E(\|\boldsymbol\kappa\|_2)}{\boldsymbol\kappa^2} \frac{E(\|\boldsymbol\iota\|_2)}{\boldsymbol\iota^2} \left(1 + \frac{(\boldsymbol\kappa\cdot\boldsymbol\iota)^2}{\boldsymbol\kappa^2\boldsymbol\iota^2}\right) \\
&\qquad \cdot \left(\delta_0(\lambda_1 - \mathbf{t}\cdot\boldsymbol\kappa)\delta_0(\lambda_2 + \bar{\mathbf{v}}\cdot\boldsymbol\kappa) - \delta_0(\mathbf{t}\cdot\boldsymbol\kappa)\delta_0(\bar{\mathbf{v}}\cdot\boldsymbol\kappa)\right) \\
&\qquad \cdot \left(\delta_0(\mu_1 - \mathbf{t}\cdot\boldsymbol\iota)\delta_0(\mu_2 + \bar{\mathbf{v}}\cdot\boldsymbol\iota) - \delta_0(\mathbf{t}\cdot\boldsymbol\iota)\delta_0(\bar{\mathbf{v}}\cdot\boldsymbol\iota)\right) d\boldsymbol\kappa\, d\boldsymbol\iota\, d\boldsymbol\lambda\, d\boldsymbol\mu,
\end{aligned}$$

according to (56) and (68). Following the calculations of the $\mathcal{L}^2$-estimate gives

$$\begin{aligned}
&\|(\mathbf{\Gamma}_{\mathbf{W}_{\boldsymbol\alpha}} - \mathbf{\Gamma}_{\mathbf{Z}_{\boldsymbol\alpha}})(\boldsymbol\sigma)\|_{l^2}^2 \\
&\le \frac{\alpha_s^2\,\alpha_t^2}{8\pi^2\,\bar v_{\mathrm{n}}^2} \left(\int_{\mathbb{R}^2} e^{i(\alpha_s\iota_1, -\alpha_t\iota_2)\cdot\boldsymbol\sigma} \left(\mathcal{E}\left(\alpha_s\iota_1, \frac{1}{\bar v_{\mathrm{n}}}(\alpha_t\iota_2 - \alpha_s\bar v_{\mathrm{t}}\iota_1)\right) - \mathcal{E}(0,0)\right) \mathcal{F}_{H_1}(\boldsymbol\iota) d\boldsymbol\iota \right)^2 \\
&\le \frac{\alpha_s^2\,\alpha_t^2}{8\pi^2\,\bar v_{\mathrm{n}}^2} \left(\int_{\mathbb{R}^2} \left| \mathcal{E}\left(\alpha_s\iota_1, \frac{1}{\bar v_{\mathrm{n}}}(\alpha_t\iota_2 - \alpha_s\bar v_{\mathrm{t}}\iota_1)\right) - \mathcal{E}(0,0)\right| \mathcal{F}_{H_1}(\boldsymbol\iota)\, d\boldsymbol\iota \right)^2.
\end{aligned}$$

Repeating then the splitting ansatz for the integral and the estimation arguments for the energy difference in J_U and $J_{\mathbb{R}^2\setminus U}$ yields

$$\begin{aligned}
J_U &\le \frac{4\mathcal{S}}{\pi^2} \int_0^{\alpha_t^{-1}(\bar v_{\mathrm{n}}+|\bar v_{\mathrm{t}}|)} \int_0^{\alpha_s^{-1}} \left(\alpha_s \left(1 + \frac{|\bar v_{\mathrm{t}}|}{\bar v_{\mathrm{n}}}\right) \iota_1 + \frac{\alpha_t}{\bar v_{\mathrm{n}}}\iota_2 \right) \mathcal{F}_{H_1}(\iota_1,\iota_2)\, d\iota_1\, d\iota_2 \\
&\le \frac{4\mathcal{S}}{\pi^2} \int_0^\infty \frac{1-\cos\iota}{\iota^2} d\iota \left(\alpha_s \left(1 + \frac{|\bar v_{\mathrm{t}}|}{\bar v_{\mathrm{n}}}\right) \left(\int_0^1 \frac{1-\cos\iota}{\iota} d\iota + \int_1^{\alpha_s^{-1}} \frac{1-\cos\iota}{\iota} d\iota \right) \right. \\
&\qquad \left. + \frac{\alpha_t}{\bar v_{\mathrm{n}}} \left(\int_0^1 \frac{1-\cos\iota}{\iota} d\iota + \int_1^{\alpha_t^{-1}(\bar v_{\mathrm{n}}+|\bar v_{\mathrm{t}}|)} \frac{1-\cos\iota}{\iota} d\iota \right) \right) \\
&\le \frac{2\mathcal{S}}{\pi} \left(\alpha_s \left(1 + \frac{\bar v_{\mathrm{t}}}{\bar v_{\mathrm{n}}}\right) \left(c + 2\ln\left(\frac{1}{\alpha_s}\right)\right) + \frac{\alpha_t}{\bar v_{\mathrm{n}}} \left(c + 2\ln\left(\frac{\bar v_{\mathrm{n}} + |\bar v_{\mathrm{t}}|}{\alpha_t}\right)\right)\right),
\end{aligned}$$

with $c = \int_0^1 (1-\cos\iota)/\iota\, d\iota$ and

$$\begin{aligned}
J_{\mathbb{R}^2\setminus U} &\le 4\mathcal{E}_0 \left(\int_0^\infty \int_{\alpha_s^{-1}}^\infty \mathcal{F}_{H_1}(\iota_1,\iota_2)\, d\iota_1\, d\iota_2 + \int_{\alpha_t^{-1}(\bar v_{\mathrm{n}}+|\bar v_{\mathrm{t}}|)}^\infty \int_0^{\alpha_s^{-1}} \mathcal{F}_{H_1}(\iota_1,\iota_2)\, d\iota_1\, d\iota_2 \right) \\
&\le \frac{4\mathcal{E}_0}{\pi} \left(\alpha_s + \frac{\alpha_t}{(\bar v_{\mathrm{n}} + |\bar v_{\mathrm{t}}|)} \right). \qquad \square
\end{aligned}$$

In the limit $\alpha_i \to 0$, $i = s,t$, the support of the smoothing function $G_{\boldsymbol{\alpha}}$ tends to be the whole $\mathbb{R}^2$. This is unrealistic, as the fiber length l prescribes a natural upper bound for the spatial smoothing parameter α_s. Thus, $\alpha_s = l_{\mathrm{T}}/l$ is certainly a reasonable value for the macroscopic smoothing of the turbulent flow effects on the fiber. The temporal flow and fiber scales are related to the spatial ones by the respective velocities $\bar{\mathbf{u}}$ and $\partial_t \mathbf{r}$. The choice of $\alpha_t = \alpha_s \|\partial_t \mathbf{r}\|_2 / \|\bar{\mathbf{u}}\|_2$ seems likely. Consequently, the actual quality of the similarity estimates (65) and (66) is determined by the scales of the considered fiber-flow problem. Moreover, it depends crucially on the relation between fiber direction $\mathbf{t}$ and mean relative velocity $\bar{\mathbf{v}}$. Be aware that the estimates do not hold for linear dependence, because the amplitude $\mathbf{D}$ of the underlying uncorrelated velocity fluctuation field $\mathbf{z}$ is then not defined; cf. (51), (57). However, these events might be viewed as elements of a nullset, since the perturbing influence of the turbulence and the fiber inertia prevents the fiber from moving continuously within the mean streamlines.

4.4. Correlated and uncorrelated global force. After having provided the correlated local forces $\mathbf{g}_{cc}^{\sigma,\tau}$ and their uncorrelated asymptotic limits $\mathbf{g}_{uc}^{\sigma,\tau}$, we conclude this section with the statement of the corresponding global forces. According to the Global-from-Local Force Concept (42) and the linearization approach of (47), the correlated and uncorrelated global aerodynamic forces read

$$\begin{aligned} \mathbf{f}_{cc}^{air}(\mathbf{r}(.),s,t) &= \langle \mathbf{g}_{cc}^{\sigma,\tau}(s,t)\rangle_{N(\mathbf{r}(.),s,t)} \\ &= \langle \mathbf{f}(\bar{\mathbf{v}}(\sigma,\tau),\partial_s \mathbf{r}(\sigma,\tau))\rangle_{N(\mathbf{r}(.),s,t)} + \langle \mathbf{L}^{\mathbf{f}}(\sigma,\tau)\ \mathbf{w}_f^{\sigma,\tau}(s,t)\rangle_{N(\mathbf{r}(.),s,t)}, \end{aligned} \tag{73}$$

$$\begin{aligned} \mathbf{f}_{uc}^{air}(\mathbf{r}(.),s,t) &= \langle \mathbf{g}_{uc}^{\sigma,\tau}(s,t)\rangle_{N(\mathbf{r}(.),s,t)} \\ &= \langle \mathbf{f}(\bar{\mathbf{v}}(\sigma,\tau),\partial_s \mathbf{r}(\sigma,\tau))\rangle_{N(\mathbf{r}(.),s,t)} + \langle \mathbf{L}^{\mathbf{f}}(\sigma,\tau)\ \mathbf{z}^{\sigma,\tau}(s,t)\rangle_{N(\mathbf{r}(.),s,t)}, \end{aligned} \tag{74}$$

where

$$\begin{aligned} \langle \mathbf{L}^{\mathbf{f}}(\sigma,\tau)\ \mathbf{z}^{\sigma,\tau}(s,t)\rangle_{N(\mathbf{r}(.),s,t)} &= \sqrt{\langle \mathbf{L}^{\mathbf{f}}(\sigma,\tau)(\mathbf{D}^{\sigma,\tau})^2(\mathbf{L}^{\mathbf{f}}(\sigma,\tau))^t\rangle_{N(\mathbf{r}(.),s,t)}}\ \mathbf{p}(s,t) \\ &= \sqrt{\frac{1}{|N(\mathbf{r}(.),s,t)|}\int_{N(\mathbf{r}(.),s,t)} \mathbf{L}^{\mathbf{f}}(\sigma,\tau)(\mathbf{D}^{\sigma,\tau})^2(\mathbf{L}^{\mathbf{f}}(\sigma,\tau))^t\, d\sigma\, d\tau}\ \mathbf{p}(s,t) \end{aligned}$$

by means of Ito-calculus and the integration rule of independent Gaussian random fields. Here, $(\mathbf{p}_{s,t}, (s,t) \in [0,l] \times \mathbb{R}_0^+)$ describes as $\mathbb{R}^3$-valued Gaussian white noise a centered homogeneous generalized Gaussian random field on a two-dimensional parameter set, i.e.,

$$\lim_{(\triangle s, \triangle t)\to \mathbf{0}} \sqrt{\triangle s \triangle t}\ \mathbf{p}(s,t) \sim \mathcal{N}(\mathbf{0},\mathbf{I}).$$

The global forces (73), (74) inherit thereby the proven approximation quality of the local forces on a macroscopic fiber scale because of the applied linear averaging procedure $\langle\cdot\rangle$.

The local flow quantities hardly ever differ in the fiber region N, since it is contained in the turbulence domain M of (12). This fact, in combination with the assumption of a locally linear fiber, motivates the skipping of the averaging procedure. For further theoretical and numerical treatment it is hence convenient to consider the following approximative forces:

$$\hat{\mathbf{f}}_{cc}^{air}(\mathbf{r}(.),s,t) = \mathbf{f}(\bar{\mathbf{v}}(s,t),\partial_s \mathbf{r}(s,t)) + \mathbf{L}^{\mathbf{f}}(s,t)\ \langle \mathbf{w}_f^{\sigma,\tau}(s,t)\rangle_{N(\mathbf{r}(.),s,t)}, \tag{75}$$

$$\hat{\mathbf{f}}_{uc}^{air}(\mathbf{r}(.),s,t) = \mathbf{f}(\bar{\mathbf{v}}(s,t),\partial_s \mathbf{r}(s,t)) + \mathbf{L}^{\mathbf{f}}(s,t)\ \mathbf{D}^{s,t}\ \mathbf{p}(s,t). \tag{76}$$

Analogously to (39), the averaging brackets in (75) can be explicitly formulated as Ito-integrals with the Wiener process/Brownian motion ($\mathcal{W}_{\sigma,\tau}, (\sigma,\tau) \in [0,l] \times \mathbb{R}_0^+$):

$$\langle \mathbf{w}_f^{\sigma,\tau}(s,t) \rangle_{N(\mathbf{r}(.),s,t)} = \frac{1}{\sqrt{|N(\mathbf{r}(.),s,t)|}} \int_{N(\mathbf{r}(.),s,t)} \mathbf{w}_f^{\sigma,\tau}(s,t)\, d\mathcal{W}_{\sigma,\tau}.$$

Whereas the functional dependence between $\hat{\mathbf{f}}_{cc}^{air}$ and $\mathbf{r}$ remains as a consequence of the realization of the correlation structure of the underlying local velocity fluctuation fields $\mathbf{w}_f^{\sigma,\tau}$, the applied simplification localizes the uncorrelated global force in (76), i.e., $\hat{\mathbf{f}}_{uc}^{air}(\mathbf{r}(.),s,t) = \hat{\mathbf{f}}_{uc}^{air}(\mathbf{r}(s,t), \partial_s\mathbf{r}(s,t),\, \partial_t\mathbf{r}(s,t), s, t)$. Thus, the resulting fiber motion is given by a wave-like system of stochastic PDEs with algebraic constraint.

5. Conclusions and outlook. Our presented Global-from-Local Force Concept in combination with the linearization approach (47) allows the approximation of the constructed correlated random aerodynamic force by Gaussian white noise with flow-dependent amplitude in the case of a macroscopic description of the fiber dynamics. The stated general results are applicable to concrete practical problems with fiber-turbulence interaction scales that yield negligibly small deviations in the $\mathcal{L}^2$- and $\mathcal{L}^\infty$-estimates of (65) and (66). Choose therefore a specific air drag model $\mathbf{f}$ and derive an appropriate linear drag operator $\mathbf{L}^{\mathbf{f}}$; then the global aerodynamic force $\hat{\mathbf{f}}_{uc}^{air}$ of (76) leads to a stochastic partial differential system with additive white noise for the fiber dynamics, (7), which can efficiently be handled numerically. For the choice of an empirically motivated, nearly quadratic drag model in a melt-spinning process, the effects of the correlated global force and its uncorrelated asymptotic limit that are imposed on the fiber by the turbulent flow are quantified and numerically compared in a subsequent paper.

REFERENCES

[1] N. Afzal and R. Narasimha, *Axisymmetric turbulent boundary layer along a circular cylinder*, J. Fluid Mech., 74 (1976), pp. 113–128.

[2] S. S. Antman, *Nonlinear Problems of Elasticity*, Springer-Verlag, New York, 1995.

[3] G. K. Batchelor and I. Proudman, *The large-scale structure of homogeneous turbulence*, Philos. Trans. Roy. Soc. London A, 248 (1956), pp. 369–405.

[4] A. J. Chorin, *Vorticity and Turbulence*, Springer-Verlag, New York, 1994.

[5] R. Core and C. Crowe, *Effect of particle size on modulating turbulent intensity*, Internat. J. Multiphase Flow, 15 (1989), pp. 279–285.

[6] J. H. Ferziger and M. Perić, *Computational Methods for Fluid Dynamics*, 3rd ed., Springer, Berlin, 2002.

[7] F. N. Frenkiel and P. S. Klebanoff, *Higher order correlations in a turbulent field*, Phys. Fluids A, 10 (1967), pp. 507–520.

[8] U. Frisch, *Turbulence. The Legacy of A. N. Kolmogorov*, Cambridge University Press, Cambridge, UK, 1995.

[9] W. Heisenberg, *On the theory of statistical and isotropic turbulence*, Proc. Roy. Soc. London A, 195 (1948), pp. 402–406

[10] O. Hinze, *Turbulence*, 2nd ed., McGraw–Hill, New York, 1975.

[11] S. Kruse, *A Generalized Parametrix Method, Smoothness of Random Fields and Applications to Parabolic Stochastic Partial Differential Equations*, Ph.D. thesis, Universität Mannheim, Mannheim, Germany, 2001.

[12] B. E. Launder and B. I. Sharma, *Application of the energy dissipation model of turbulence to the calculation of flow near a spinning disc*, Lett. Heat Mass Transfer, 1 (1974), pp. 131–138.

[13] N. Marheineke, *Turbulent Fibers—On the Motion of Long, Flexible Fibers in Turbulent Flows*, Ph.D. thesis, Technische Universität Kaiserslautern, Kaiserlautern, Germany, 2005.

[14] J. A. Olsen and R. J. Kerekes, *The motion of fibres in turbulent flow*, J. Fluid Mech., 337 (1998), pp. 47–64.

[15] P. Sagaut, *Large Eddy Simulation for Incompressible Flows. An Introduction*, 2nd ed., Springer-Verlag, New York, 2004.
[16] H. Schlichting, *Grenzschicht-Theorie*, Verlag G. Braun, Karlsruhe, Germany, 1982.
[17] G. I. Taylor, *The spectrum of turbulence*, Proc. Roy. Soc. London A, 164 (1938), pp. 476–490.
[18] D. C. Wilcox, *Turbulence Modeling for CFD*, 2nd ed., DCW Industries, La Cañada, CA, 1998.
[19] P. K. Yeung and Y. Zhou, *On the Universality of the Kolmogorov Constant in Numerical Simulations of Turbulence*, ICASE Report 97-64, NASA CR-97-206251, 1997.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1727–1748

THE LINEAR LIMIT OF THE DIPOLE PROBLEM FOR THE THIN FILM EQUATION*

MARK BOWEN† AND THOMAS P. WITELSKI‡

Abstract. We investigate self-similar solutions of the dipole problem for the one-dimensional thin film equation on the half-line $\{x \geq 0\}$. We study compactly supported solutions of the linear moving boundary problem and show how they relate to solutions of the nonlinear problem. The similarity solutions are generally of the second kind, given by the solution of a nonlinear eigenvalue problem, although there are some notable cases where first-kind solutions also arise. We examine the conserved quantities connected to these first-kind solutions. Difficulties associated with the lack of a maximum principle and the non–self-adjointness of the fundamental linear problem are also considered. Seeking similarity solutions that include sign changes yields a surprisingly rich set of (coexisting) stable solutions for the intermediate asymptotics of this problem. Our results include analysis of limiting cases and comparisons with numerical computations.

Key words. thin-film equations, similarity solutions, dipole problem

AMS subject classifications. 35K65, 35B40, 35C05, 76A20

DOI. 10.1137/050637832

1. Introduction.

The dipole problem for the one-dimensional "absolute-valued" thin-film equation

$$\frac{\partial h}{\partial t} = -\frac{\partial}{\partial x}\left(|h|^n \frac{\partial^3 h}{\partial x^3}\right), \tag{1.1}$$

where n is a real constant, is defined as an initial-boundary value problem on the half-line $\{x \geq 0\}$, starting from bounded compactly supported initial data

$$h(x,0) = h_0(x) \qquad \text{for } 0 \leq x \leq \ell_0, \tag{1.2}$$

with $h(x,0) \equiv 0$ for $x \geq \ell_0$. At the origin, we impose the boundary conditions [11, 21, 38]

$$h = h_x = 0 \qquad \text{at } x = 0. \tag{1.3}$$

For $n > 0$, it is well known that compactly supported solutions of (1.1) exist with fronts that have a finite speed of propagation (see, for example, [6, 7, 8, 9, 10]). Meanwhile, for $n = 0$, where (1.1) is linear, solutions starting from compact initial data instantaneously gain support over the entire domain for $t > 0$. In [12], Bernis, Hulshof, and Quiros compared the $n \to 0$ limit with the $n = 0$ (linear) behavior in the Cauchy problem for (1.1). In a similar spirit, we will study related issues for the more complicated problem of second-kind similarity solutions for the dipole problem (1.1)–(1.3).

*Received by the editors August 9, 2005; accepted for publication (in revised form) March 28, 2006; published electronically July 17, 2006.
http://www.siam.org/journals/siap/66-5/63783.html

†Graduate School of Mathematical Sciences, University of Tokyo, 3-8-1 Komaba, Meguro, Tokyo 153-8914, Japan, and School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK (bowen@ms.u-tokyo.ac.jp). This author was supported by EPSRC and JSPS Postdoctoral Fellowships.

‡Department of Mathematics, Duke University, Durham, NC 27708-0320 (witelski@math.duke.edu). This author was supported by NSF grants DMS-0074049, DMS-0244498, and DMS-0239125.

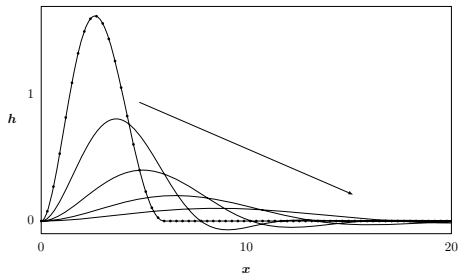


FIG. 1.1. *Numerical simulation of the dynamics of the dipole initial-boundary value problem for* (1.1) *with* $n = 0.1$. *Starting from nonnegative compact initial data (dotted curve), the dynamics approach a self-similar mixed draining/spreading evolution with the appearance of sign changes.*

To this end, we consider solutions of the moving boundary problem on $0 \le x \le \ell(t)$ with boundary conditions [32]

$$h = h_x = |h|^n h_{xxx} = 0 \qquad \text{as } x \to \ell^-(t), \tag{1.4}$$

where $\ell(0) \equiv \ell_0$. From formal considerations, we can write the equation for interface motion as

$$\frac{d\ell}{dt} = \lim_{x \to \ell^-(t)} |h|^{n-1} h_{xxx}. \tag{1.5}$$

A sample numerical simulation of the dynamics for (1.1)–(1.3) with small positive n is shown in Figure 1.1. We will present evidence that, for appropriate initial data, the dipole similarity solutions accurately predict the dynamics of (1.1).

The dipole problem for (1.1) was first referred to by Barenblatt [1] as an extension of the analogous problem for the porous medium equation [3, 4, 28, 29],

$$\partial_t h = \partial_x \left(|h|^m \partial_x h \right); \tag{1.6}$$

a comparison between dipole similarity solutions of (1.6) and experimental results on gravity currents was undertaken in [33]. Barenblatt noted that the properties of the thin-film equation (1.1) were likely to yield an intricate set of nontrivial similarity solutions. An important difference between (1.6) and (1.1) is that the thin-film equation does not have a maximum principle. That is, solutions starting from nonnegative initial data do not necessarily stay nonnegative [16, 18, 19]; indeed, observe Figure 1.1.

Equation (1.1) with $n > 0$ arises in the study of surface tension–driven thin films [9, 15, 16, 17, 18, 23, 34, 35, 36]. For these problems, $h(x,t)$ is representative of the thickness of the fluid film, and from physical considerations, the solutions must be nonnegative, $h(x,t) \ge 0$. More recently, similar higher-order nonlinear degenerate parabolic equations have been applied to image processing problems [25, 26]. In this paper, we are primarily interested in the mathematical structure of solutions to (1.1) for small n, and we will admit solutions that change sign (consequently, the usual h^n in the thin-film equation is written as $|h|^n$ in (1.1)). Removing the positivity requirement means that $x = \ell(t)$ in (1.5) is not necessarily the first point in the solution of (1.1) at which $h = 0$, in contrast with the study in [12]. This yields a much richer set of solutions, including the traditional nonnegative solution and new classes of solutions with sign changes.

In section 2 we introduce a similarity solution of (1.1) and clarify our use of the terminology "first-kind" and "second-kind." The latter class of solution will be most prevalent herein, reducing the thin-film equation (1.1) to a fourth-order non–self-adjoint eigenvalue problem that will be the main focus of the paper. Section 3 considers the fundamental linear problem for (1.1) in terms of the known conserved quantities of (1.1)–(1.4). We focus on the linear case in detail in section 4. Classes of solutions with compact support or with support on the half-line are obtained and shown to be consistent with the conserved quantity results of section 3. In section 5, we return to the fully nonlinear problem and consider the limit $n \to 0$. This small n study links the solutions of the nonlinear problem to the linear results found in section 4. In section 6, numerical simulations of (1.1) are compared with the predictions given by the dipole solutions. We conclude, in section 7, with a discussion of the results obtained in the previous sections.

2. Similarity solutions. We seek self-similar solutions of the thin film equation (1.1) taking the form

$$h = \tau^{-\alpha} H(\eta), \qquad \eta = x/\tau^{\beta}, \qquad \tau = t + t_0. \tag{2.1}$$

Similarity solutions of this general form solve many different classes of initial and boundary value problems for (1.1): source-type solutions of the Cauchy problem, draining solutions of the Dirichlet problem, and Boltzmann solutions of the dam-break problem [13, 20, 22, 32, 38]. Moreover, these special solutions often act as large-time attractors for solutions starting from much wider classes of initial conditions; we also expect this to be true for the dipole problem.

Substituting (2.1) into (1.1) and separating temporal and spatial variables forces the relation

$$n\alpha + 4\beta = 1, \tag{2.2}$$

with the similarity function $H(\eta)$ satisfying a boundary value problem on $0 \le \eta \le L$ for the nonlinear ordinary differential equation,

$$\alpha H + \tfrac{1}{4}(1 - n\alpha)\eta H' - \left(|H|^n H'''\right)' = 0, \tag{2.3a}$$

$$\text{with} \quad H = H' = 0 \qquad \text{at } \eta = 0, \tag{2.3b}$$

$$\text{and} \quad H = H' = |H|^n H''' = 0 \qquad \text{as } \eta \to L^-. \tag{2.3c}$$

We will refer to (2.3) as problem (P). Here, L is a constant corresponding to the free boundary (1.4) of the form $\ell(t) = L\tau^{\beta}$. Problem (P) has been previously investigated by application of dynamical systems techniques [21, 27].

If a second independent relation between α and β were known, then these scaling constants could be determined explicitly. For the Cauchy problem, the conserved quantity $\int h\,dx$ determines $\alpha = \beta$, and hence $\alpha = \beta = 1/(n+4)$ [13, 32]. For the Dirichlet problem, boundary conditions set $\beta = 0$, and hence $\alpha = 1/n$ [22, 38], while for dam-break, appropriate boundary conditions set $\alpha = 0$ and $\beta = 1/4$ [20]. All of these scenarios are classified as similarity solutions of the *first kind*; see Barenblatt [1]. For the dipole problem, there is in general no such direct second relation between α and β. In (2.3a), β has been eliminated via $\beta = (1 - n\alpha)/4$, but α is left undetermined.

Consequently problem (P) is a challenging nonlinear eigenvalue problem for the scaling constant α, and the similarity solution (2.1) is said to be of the *second kind* [1]. There are, however, two exceptional cases (which we will come to in sections 3 and 6.1).

We note that problem (P) is invariant under the rescaling $H(\eta) \to \lambda^{4/n}\hat{H}(\hat{\eta})$, $\eta \to \lambda\hat{\eta}$ for any $\lambda > 0$ when $n \neq 0$. In addition, for the $n = 0$ linear problem, H and η can be rescaled independently. Therefore, to uniquely specify a solution of (P), an additional arbitrary normalization condition must be imposed. For this, we choose to set the average of the square of the similarity solution to be unity, i.e.,

$$\frac{1}{L}\int_0^L H^2\,d\eta = 1; \tag{2.3d}$$

this is a simple condition on the L^2 norm of the solution, which will be convenient in our numerical scheme for solving the boundary value problem. A different, local, normalization condition was used in [21], which was more convenient for their shooting method calculations.

For self-similar solutions (2.1), the quantity $\int h\,dx$ scales as $O(\tau^{\beta-\alpha})$. It was proved rigorously in [11] that a solution of problem (P) exists only for $n < 2$ and necessarily has $\alpha > \beta$. We therefore restrict attention to solutions where $\frac{d}{dt}\int h\,dx < 0$, and hence from (2.2), $\alpha > 1/(n+4)$. Previous analysis [21] has also indicated that there exists a critical value of $n = n_c$, such that $-4 < n_c < 0$, below which solutions of the form (2.1) do not exist; we return to this issue in a later section.

Rather than trying to construct solutions to (P) directly, we split it into two subproblems:

$$(\mathrm{P}_A) = \{(\mathrm{P}) \text{ with (2.3c) replaced by } \quad H = H' = 0 \quad \text{at} \quad \eta = L\} \tag{2.4}$$

and

$$(\mathrm{P}_B) = \{(\mathrm{P}) \text{ with (2.3c) replaced by } \quad H = |H|^n H''' = 0 \quad \text{at} \quad \eta = L\}\,. \tag{2.5}$$

For a given value of L, both (P_A) and (P_B) define fourth-order boundary value problems, with α playing the role of an eigenvalue. Consequently, we expect nontrivial solutions to the subproblems to exist only for specific values of α at each L: there will be continuous families of solutions of these problems defined by curves in the (L, α) parameter plane; see Figure 4.6(left). Intersections of the solution curves for problems (P_A) and (P_B) then identify solutions of the original problem (P), that is, $\mathrm{P} = \mathrm{P}_A \cap \mathrm{P}_B$ (see Figure 4.6(right)). We will develop this in sections 4–6.

3. Conserved quantities. As described above, similarity solutions are said to be of the first kind if the scaling parameters α, β can be obtained directly by combining (2.2) with another relation. Additional scaling relations can often be obtained from equations for conserved quantities, such as the conservation of $\int h\,dx$, or of an energy functional. For the dipole problem, conserved quantities are known for only two specific values of n, namely the first moment, $\int_0^\infty xh\,dx$, for $n = 1$ [11, 21, 31] (which will be discussed further in section 6.1), and $\int_0^\infty gh\,dx$ for $n = 0$, where we specify possible functions $g(x,t)$ below. There is a subtle difference between these two conserved integrals. The former conserved quantity provides a unique specification of $\alpha = 1/3$ for all solutions, whereas, as will become evident in what follows, the latter (through the choice of g) provides a countably infinite set of viable first-kind similarity solutions.

Consider the linear problem, (1.1) with $n = 0$. From (2.2), $\beta = 1/4$, but α is undetermined in (2.1). Suppose that α can be determined by finding a function $g(x,t)$ such the integral of the product gh is conserved [31]. From integration by parts (with appropriate assumptions on the decay of h as $x \to \infty$), we find that

$$\frac{d}{dt}\left(\int_0^\infty gh\,dx\right) = \int_0^\infty (g_t - g_{xxxx})h\,dx - [gh_{xxx} - g_x h_{xx} - g_{xx}h_x - g_{xxx}h]\Big|_0^\infty = 0. \tag{3.1}$$

In other words, g should satisfy the backward fourth-order linear diffusion equation, $g_t - g_{xxxx} = 0$, and the boundary terms should vanish, i.e., $g(0,t) = g_x(0,t) = 0$ [21]. There are, in fact, an infinite number of suitable polynomial solutions for $g(x,t)$ given by

$$g(x,t) = \sum_{m=0}^{k} \frac{x^{4m+2}t^{k-m}}{(4m+2)!(k-m)!} \quad \text{or} \quad \sum_{m=0}^{k} \frac{x^{4m+3}t^{k-m}}{(4m+3)!(k-m)!}, \tag{3.2}$$

for $k = 0, 1, 2, 3, \ldots$. Each choice of $g(x,t)$ provided by (3.2) therefore specifies a conserved quantity for $n = 0$ via (3.2), which can then be used to identify corresponding similarity solutions. Using the relationship $x = \eta t^{1/4}$ allows (3.2) to be written as separable functions of η and t:

$$g_{2k}(x,t) = t^{k+2/4} \sum_{m=0}^{k} \frac{\eta^{4m+2}}{(4m+2)!(k-m)!},$$
$$g_{2k+1}(x,t) = t^{k+3/4} \sum_{m=0}^{k} \frac{\eta^{4m+3}}{(4m+3)!(k-m)!}, \tag{3.3}$$

where the index of g corresponds to the parity of its terms. Writing the integral in (3.2) in terms of self-similar variables and requiring it to be time-independent then yields two sequences of critical exponents for (2.1),

$$\bar{\alpha}_{2k} = k + \tfrac{3}{4}, \qquad \bar{\alpha}_{2k+1} = k + 1, \tag{3.4}$$

arising from the corresponding choices for g in (3.3). We have therefore constructed a countably infinite set of first-kind similarity solutions for $n = 0$. We note that (3.3) can then be written in the form

$$g_j(x,t) = t^{\bar{\alpha}_j - 1/4}\bar{G}_j(\eta) \quad \text{for} \quad j = 0, 1, 2, 3\ldots, \tag{3.5}$$

and the value of the conserved integral $\int_0^\infty g_j(x,t)h(x,t)\,dx = \int_0^\infty \bar{G}_j(\eta)\bar{H}(\eta)\,d\eta$, for similarity solutions (2.1).

We also note the existence of transformations connecting these critical $\bar{\alpha}$ values and the corresponding $\bar{H}(\eta)$ similarity solutions. Since, for $n = 0$, equation (1.1) is linear, if $h(x,t)$ is a solution, then its derivatives are also solutions (subject to the boundary conditions). Taking a ∂_x-derivative of the similarity solution (2.1) yields another solution of the form $h(x,t) = t^{-[\bar{\alpha}+1/4]}\bar{H}'(\eta)$, and hence the transformation

$$\{\bar{\alpha}_{2k}, \bar{H}(\eta)\} \to \{\bar{\alpha}_{2k} + \tfrac{1}{4}, \bar{H}'(\eta)\}. \tag{3.6}$$

Similarly, taking the ∂_t-derivative of (2.1) yields the transformation

$$\{\bar{\alpha}_j, \bar{H}(\eta)\} \to \{\bar{\alpha}_j + 1, -\bar{\alpha}_j \bar{H}(\eta) - \tfrac{1}{4}\eta \bar{H}'(\eta)\}. \tag{3.7}$$

Note that (3.7) can be applied to any $\bar{\alpha}$ in (3.4), $\bar{\alpha}_j \to \bar{\alpha}_{j+2}$, but (3.6) only maps even to odd modes, $\bar{\alpha}_{2k} \to \bar{\alpha}_{2k+1}$. This is due to the constraints imposed by the boundary conditions at $x = 0$ given in (1.3). These two transformations can be used to obtain the solutions for all k in terms of the first solution, with $\bar{\alpha}_0 = 3/4$.

4. Similarity solutions of the $n = 0$ linear problem. We consider subproblem (P_A) with $n = 0$, that is, the linear problem,

$$\alpha H + \tfrac{1}{4}\eta H' - H'''' = 0, \qquad 0 \le \eta \le L, \tag{4.1a}$$

$$H(0) = H'(0) = 0, \qquad H(L) = H'(L) = 0. \tag{4.1b}$$

The structure of the set of solutions for this boundary value problem is nontrivial; for any fixed value of the support, L, this is a fourth-order non–self-adjoint eigenvalue problem for the compactly-supported second-kind similarity solutions $\{\alpha, H(\eta)\}$.

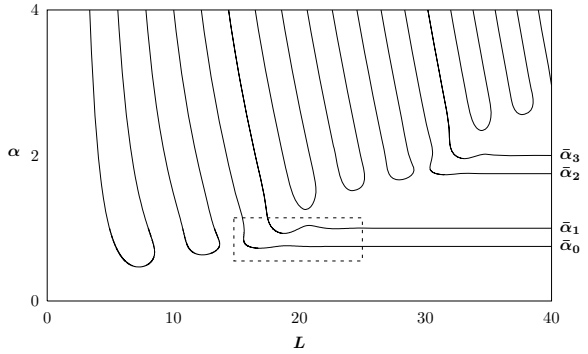


FIG. 4.1. *Branches of solution curves in the (L, α) parameter plane for the $n = 0$ (P_A) subproblem,* (4.1). *Details of the structure in the boxed region will be discussed later (see Figure* 4.5*).*

The results from a computational study for the set of nontrivial solutions of (P_A) are shown in Figure 4.1. The arrangement of the solution branches exhibits interesting regular patterns in different regions of the (L, α) parameter space. These patterns correspond to different qualitative forms of the $H(\eta)$ solution profiles (see Figure 4.3). In this section, we perform an asymptotic analysis of (4.1) in the two distinct limits, $L \to \infty$ and $\alpha \to \infty$, in an effort to understand the structure of Figure 4.1. In the process, we determine the solutions defined on the half-line that correspond to the critical $\bar{\alpha}$ values (3.4). Interestingly, these first-kind solutions are limiting cases of sequences of compactly supported second-kind solutions with increasing support (and increasing numbers of sign changes).

4.1. Limiting behavior for $L \to \infty$. The analysis of this limit problem is composed of two stages. First, we obtain analytic solutions of (4.1) on the half-line $\{\eta \ge 0\}$ in terms of an integral representation and employ steepest descents to construct the leading-order asymptotic approximation to $H(\eta)$. This approach yields the critical values (3.4) from an argument independent of section 3. Second, we consider the leading-order behavior of solutions for large but finite L; this will explain the "wavy" structure of the solution curves and the "bends" shown in Figure 4.1.

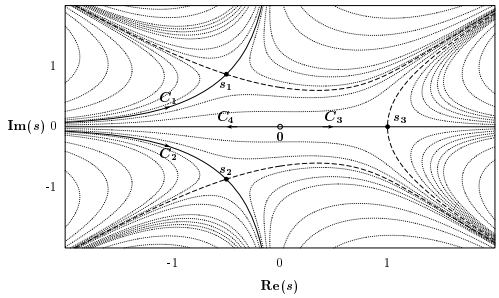


FIG. 4.2. *The steepest descent contours defining the solutions* (4.4) *in the complex plane.* s_1, s_2, s_3 *are the saddle points of* $\phi(s) = \eta s - s^4$.

4.1.1. Solutions on the half-line. We can express the solutions of (4.1) in terms of a generalized Laplace transform,

$$H(\eta) = \int_C Y(s) e^{\eta s}\, ds, \tag{4.2}$$

with (4.1a) then taking the form

$$\int_C \left[\alpha Y - \frac{1}{4}\frac{d}{ds}(sY) - s^4 Y\right] e^{\eta s}\, ds = -sY(s)e^{\eta s}\Big|_C. \tag{4.3}$$

Choosing the contour C appropriately so that the boundary terms vanish yields a first-order equation for $Y(s)$ with solution $Y(s) = s^{4\alpha-1}e^{-s^4}$. Therefore, we define four linearly independent solutions of (4.1a) as

$$\hat{H}_k(\eta) = \int_{C_k} s^{4\alpha-1} e^{-s^4} e^{\eta s}\, ds, \tag{4.4}$$

for $k = 1, 2, 3, 4$. The boundary terms in (4.3) vanish for $|s| \to \infty$ in sectors centered along the real and imaginary axes. The boundary terms in (4.3) also vanish at the origin, $s = 0$. The four contours can therefore be defined as shown in Figure 4.2. Deforming contours $C_{1,2}$ onto the real and imaginary axes, we can express these integrals more explicitly as the complex conjugate solutions

$$\hat{H}_{1,2}(\eta) = e^{\pm i\pi(4\alpha-1)} \int_0^\infty t^{4\alpha-1} e^{-t^4} e^{-\eta t}\, dt + e^{\pm i2\pi\alpha} \int_0^\infty t^{4\alpha-1} e^{-t^4} e^{\pm i\eta t}\, dt, \tag{4.5}$$

which can also be written in terms of generalized hypergeometric functions. Using steepest descents [5], we can obtain the leading-order asymptotic behaviors of these solutions as $\eta \to \infty$ from the contributions of the three saddle points of the phase function $\phi(s) = \eta s - s^4$ (that is, $\phi'(s_k) = 0$, with $s_{1,2} = (\eta/4)^{1/3} e^{\pm i2\pi/3}$ and $s_3 = (\eta/4)^{1/3}$). Solutions $\hat{H}_{1,2}(\eta)$ are exponentially decaying as $\eta \to \infty$,

$$\hat{H}_{1,2}(\eta) \sim \sqrt{\frac{\pi}{6}} \left(\frac{\eta}{4}\right)^{2(2\alpha-1)/3} e^{\pm i\pi(8\alpha-1)/3} e^{-3(\eta/4)^{4/3}[1\mp i\sqrt{3}]/2}. \tag{4.6}$$

The third solution $\hat{H}_3(\eta)$ is exponentially growing as $\eta \to \infty$ and cannot contribute to the bounded solutions of (4.1a) on $0 \le \eta < \infty$,

$$\hat{H}_3(\eta) = \int_0^\infty t^{4\alpha-1} e^{-t^4} e^{\eta t}\, dt \sim \sqrt{\frac{\pi}{6}} \left(\frac{\eta}{4}\right)^{2(2\alpha-1)/3} e^{3(\eta/4)^{4/3}}. \tag{4.7}$$

The final linearly independent solution is given by a contour from the origin to infinity along the negative real axis,

$$\hat{H}_4(\eta) = e^{i\pi(4\alpha-1)} \int_0^\infty t^{4\alpha-1} e^{-t^4} e^{-\eta t}\, dt. \tag{4.8}$$

Laplace's method applied to this integral yields the leading-order algebraic decay behavior as $\eta \to \infty$,

$$\hat{H}_4(\eta) \sim e^{-(4\alpha-1)[1-i\pi]} \sqrt{\frac{2\pi}{4\alpha-1}} \left(\frac{4\alpha-1}{\eta}\right)^{4\alpha}. \tag{4.9}$$

A local analysis of the solutions of (4.1a) at $\eta = 0$ shows that there are four regular, linearly independent solutions with power series in η^4, starting with terms $1, \eta, \eta^2$, and η^3, respectively. Noting that (4.9) is singular as $\eta \to 0$ suggests that WKB analysis would yield that the origin is a turning point of (4.1a) [5]. Because the asymptotic formulae for $\hat{H}_k(\eta)$ are not uniformly valid down to $\eta = 0$, we must apply the boundary conditions (4.1b) directly to the linear combination of the integral forms for the solutions, (4.5) and (4.8). Specifically, imposing the conditions that $H(0) = H'(0) = 0$ and that $H''(0)$ is real and positive yields

$$H(\eta) = \left(2\mathrm{Re}(e^{-i\pi(2\alpha+3/4)}\hat{H}_1(\eta)) + B(\alpha)\hat{H}_4(\eta)\right), \tag{4.10}$$

with

$$B(\alpha) = e^{-i4\pi\alpha}\left(2\cos(\pi[2\alpha+1/4]) - \sqrt{2}\right). \tag{4.11}$$

Since (4.9) decays algebraically as $\eta \to \infty$, while (4.6) decays exponentially, if $B(\alpha) \neq 0$, then $H(\eta) \sim B(\alpha)\hat{H}_4(\eta)$. However, since $|\hat{H}_4(\eta)|$ given by (4.9) is strictly positive and monotone decreasing, then only the trivial solution would satisfy the interface boundary conditions (4.1b). Hence we conclude that $B(\alpha) = 0$, corresponding to the infinite sequence of roots $\bar{\alpha}$,

$$\bar{\alpha}_{2p} = p + \tfrac{3}{4}, \qquad \bar{\alpha}_{2p+1} = p + 1, \qquad \text{for } p = 0, 1, 2, 3, \ldots. \tag{4.12}$$

Using (4.6) in (4.10), the leading-order behavior of solutions of (4.1) on the half-line for $\eta \to \infty$ is then given by

$$\bar{H}(\eta) \sim \sqrt{\frac{2\pi}{3}} \left(\frac{\eta}{4}\right)^{\frac{4}{3}(\bar{\alpha}-\frac{1}{2})} e^{-\frac{3}{2}(\eta/4)^{4/3}} \cos\left(\tfrac{3\sqrt{3}}{2}\left(\tfrac{\eta}{4}\right)^{4/3} + \tfrac{\pi}{3}[2\bar{\alpha} + \tfrac{11}{4}]\right). \tag{4.13}$$

Similarly, expanding (4.10) for $\eta \to 0$ yields the local behavior of the solution near the origin,

$$\bar{H}(\eta) \sim \frac{1 + \cos(2\pi\bar{\alpha}) - \sin(2\pi\bar{\alpha})}{2\sqrt{2}} \sum_{k=0}^{\infty} \left(\frac{\Gamma(k+\frac{5}{4})}{\Gamma(4k+3)}\eta^{4k+2} - \frac{\Gamma(k+\frac{3}{2})}{\Gamma(4k+4)}\eta^{4k+3}\right). \tag{4.14}$$

Indeed, (4.12) coincides exactly with (3.4). The form of (4.13) is suggestive of the strongly damped oscillatory solutions shown in Figure 4.3 (right). In Figure 4.4 we show the excellent agreement between (4.13) for $\bar{\alpha}_0 = 3/4$ and a numerically computed solution on the first branch of solutions at large L. Next, we improve upon this result by determining the oscillatory structure of the branches of solutions at large L as they approach the singular values, $\alpha(L) \to \bar{\alpha}$.

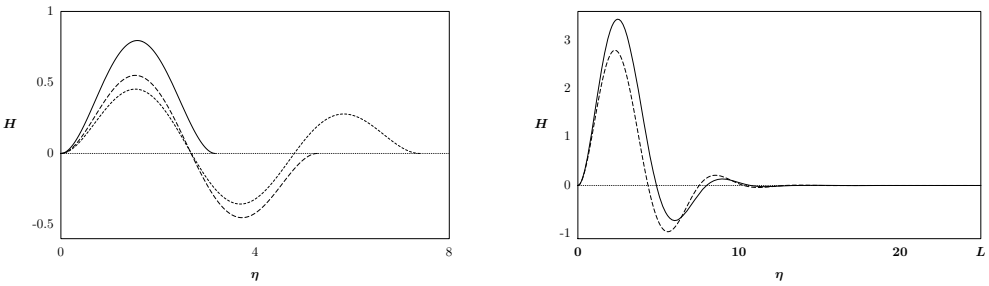


FIG. 4.3. *Examples of numerically computed solutions $H(\eta)$ of* (4.1)*: (left) weakly decaying oscillatory solutions from the first three solution branches (different L's) for fixed $\alpha = 5$, and (right) strongly damped solutions from the first two branches for fixed $L = 25$ with different values of α.*

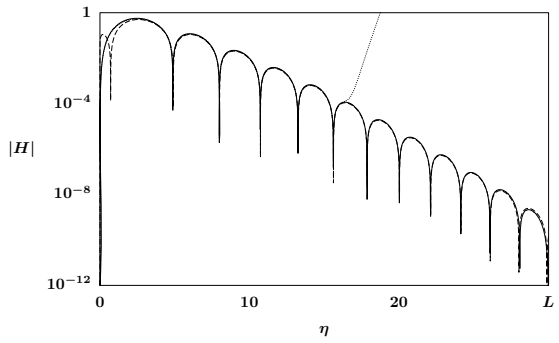


FIG. 4.4. *Log-linear plot comparing the asymptotic solution $\bar{H}(\eta)$ of the half-line problem (dashed curve for* (4.13)*, dotted curve using the first twenty terms of* (4.14)*) for $\bar{\alpha}_0 = 3/4$ with a numerically computed solution for $L = 30$ (solid curve) from the first branch of solutions of* (P_A).

4.1.2. Behavior for $L \to \infty$. Having obtained the $\bar{\alpha}$ scaling exponents for the half-line problem, we now consider how these values are approached in the limit $L \to \infty$. In what follows, we shall use $\bar{\alpha}$ to denote the values taken from (4.12); where necessary, additional subscripts will be used to distinguish special cases. We will obtain the asymptotic behavior of $\alpha(L) \sim \bar{\alpha} + \delta(L)$, with $\delta \to 0$ as $L \to \infty$.

The general solution of (4.1a) can be written as

$$H(\eta) = c_1\hat{H}_1(\eta) + c_2\hat{H}_2(\eta) + c_3\hat{H}_3(\eta) + c_4\hat{H}_4(\eta). \tag{4.15}$$

On the half-line, the contribution from $\hat{H}_3$ was omitted ($c_3 = 0$), as it was inconsistent with the required far-field behavior as $\eta \to \infty$. However, for finite L, the contribution of $\hat{H}_3$ is bounded and and must be retained in (4.15). We begin by imposing $H(0) = H'(0) = 0$ from (4.1b) on (4.15). This yields c_1, c_2, and c_4 in terms of c_3:

$$\begin{aligned} c_1 &= \tfrac{1}{4}(-1-i)\omega + ic_3\omega, \qquad c_2 = \tfrac{1}{4}(-1+i)\omega^{-1} - ic_3\omega^{-1}, \\ c_4 &= \tfrac{1}{4}(1+i)\omega - \tfrac{1}{2}\omega^2 + \tfrac{1}{4}(1-i)\omega^3 + c_3(-i\omega + \omega^2 + i\omega^3), \end{aligned} \tag{4.16}$$

where $\omega = e^{-i2\pi\alpha}$ (setting $c_3 = 0$ recovers (4.10), (4.11)). The remaining boundary conditions from (4.1b), namely $H(L) = H'(L) = 0$, are then imposed on (4.15) to yield a system of equations relating c_3, δ, L. These equations can be simplified using

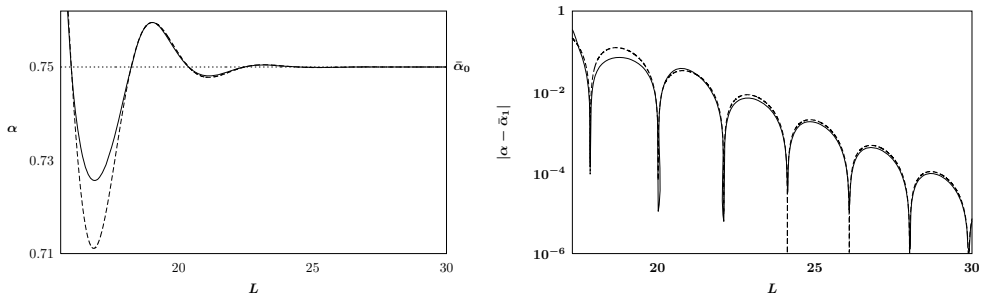


FIG. 4.5. *Details from Figure* 4.1*: comparison between the numerical results (solid curves) and the asymptotic results* $\alpha(L) = \bar{\alpha} + \delta(L)$*,* (4.18)*, (dashed curve) for large* L*. (left) for* $\bar{\alpha}_0 = 3/4$*, and (right)* $|\alpha - \bar{\alpha}|$ *compared with* $|\delta(L)|$ *for* $\bar{\alpha}_1 = 1$*.*

the asymptotics of the $\hat{H}_k(\eta)$ from section 4.1.1 and the knowledge that $c_3 \to 0$ and $\delta \to 0$ in the limit $L \to \infty$. We postulate that the relevant leading-order balance in the $H'(L) = 0$ equation is between the $\hat{H}_3'$ and $\{\hat{H}_1', \hat{H}_2'\}$ terms. These terms dominate as the process of differentiation introduces into them an additional factor of $(L/4)^{1/3}$. Working through the algebra yields the leading-order asymptotic expression for $L \to \infty$,

$$c_3 \sim -\tfrac{1}{\sqrt{2}} \exp\left(-\tfrac{9}{2}[L/4]^{4/3}\right) \sin\left(\tfrac{\pi}{3}\left(\tfrac{11}{4} - 2\bar{\alpha}\right) - \tfrac{3\sqrt{3}}{2}[L/4]^{4/3}\right). \tag{4.17}$$

To obtain a corresponding expression for $\delta(L)$ we consider the $H(L) = 0$ equation. As $\hat{H}_4(\eta)$ is inherently linked to the requirement that $B(\bar{\alpha}) = 0$ from (4.11) (when L is infinite), we expect this term to contribute in the analysis of the limit $L \to \infty$. This is indeed the case, and $\hat{H}_4(\eta)$ balances with all of the terms from (4.15). We also make the a priori assumption that c_3, as given by (4.17), decays more rapidly than δ as $L \to \infty$. Expanding complex exponentials as Taylor series, $e^{i\rho\alpha} \sim e^{i\rho\bar{\alpha}}(1 + i\rho\delta)$, and imposing $B(\bar{\alpha}) = 0$ removes the leading-order contribution from $\hat{H}_4(\eta)$, leaving the higher-order terms to balance with the remaining leading-order terms; this becomes obvious once c_3 is substituted in from (4.17), as this shows that $\hat{H}_3(\eta)$ and $\{\hat{H}_1, \hat{H}_2\}$ have the same decay rate as $L \to \infty$. The result of this analysis is an expression that can be simplified (exploiting the asymptotic ordering $c_3 \ll \delta \ll 1$ as $L \to \infty$) to give

$$\delta \sim \pm \frac{[L/\sqrt{2}]^{\frac{2}{3}(8\bar{\alpha}-1)}}{\pi\sqrt{2}\,(4\bar{\alpha}-1)^{4\bar{\alpha}-\frac{1}{2}}} \exp\left(4\bar{\alpha} - 1 - \tfrac{3}{2}[L/4]^{4/3}\right) \sin\left(\tfrac{\pi}{3}\left(\tfrac{9}{4} - 2\bar{\alpha}\right) - \tfrac{3\sqrt{3}}{2}[L/4]^{4/3}\right). \tag{4.18}$$

The positive sign is taken for $\bar{\alpha}_{\text{odd}}$, while the negative is used for $\bar{\alpha}_{\text{even}}$. It is clear that (4.18) decays more slowly than (4.17) as $L \to \infty$, thereby justifying our original assumption on the ordering $c_3 \ll \delta$; (4.18) contains an algebraically growing factor absent from (4.17) as well as having a slower exponential decay rate. Figure 4.5 shows a comparison between (4.18) and the computed solution branches from Figure 4.1.

4.2. Limiting behavior for $\alpha \to \infty$. Turning to the limit of large α, to balance the αH term in (4.1a) as $\alpha \to \infty$, the solution must have large gradients, with $d/d\eta = O(\alpha^{1/4})$. We employ a change of variables to write the solution as

$$H(\eta) = A(z), \qquad \eta = Lz, \tag{4.19}$$

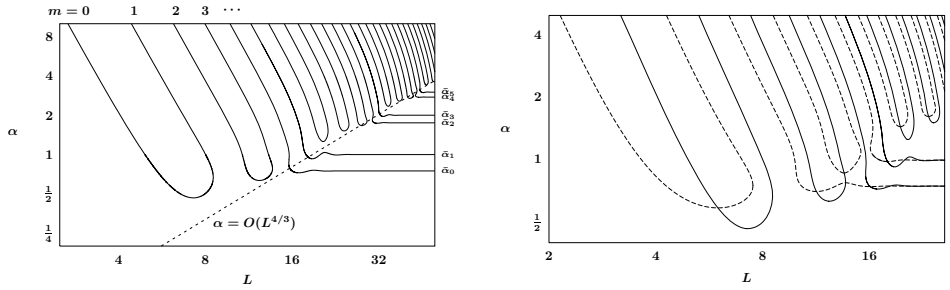


FIG. 4.6. *Log-log plots of the solution branches in the* (L, α) *parameter plane: (left) solutions of* (P_A) *with asymptotic regimes for* $L \to \infty$ *and* $\alpha \to \infty$ *connected at "bends" by* (4.26)*, and (right) solution branches for* (P_A) (4.1) *(solid curves) and* (P_B) (4.27) *(dashed curves).*

with $0 \le z \le 1$. The limit $\alpha \to \infty$ suggests that $L \sim \gamma/\alpha^{1/4}$, where γ is a positive constant. Equation (4.1a) then takes the form

$$A - \frac{1}{\gamma^4} A'''' = \frac{1}{4\alpha} z A'. \tag{4.20}$$

Writing $A(z)$ as a regular perturbation series, $A = A_0(z) + \alpha^{-1} A_1(z) + O(\alpha^{-2})$, yields the leading-order problem

$$A_0'''' - \gamma^4 A_0 = 0, \qquad A_0(0) = A_0'(0) = 0, \qquad A_0(1) = A_0'(1) = 0. \tag{4.21}$$

This is a fourth-order linear self-adjoint eigenvalue problem. Two linearly independent solutions of (4.21) that satisfy the boundary conditions at $z = 0$ are

$$A_a(z) = \cosh(\gamma z) - \cos(\gamma z), \qquad A_b(z) = e^{\gamma z} - \sin(\gamma z) - \cos(\gamma z). \tag{4.22}$$

Using these, we can write the solution satisfying the boundary condition $A_0(1) = 0$ as

$$A_0(z) = C[A_a(z)/A_a(1) - A_b(z)/A_b(1)], \tag{4.23}$$

where C is a normalization constant. Finally, imposing the remaining boundary condition, $A_0'(1) = 0$, yields the condition $\cos(\gamma)\cosh(\gamma) = 1$. This transcendental equation has an infinite number of positive solutions, and for large γ the values approach the asymptotic form for $m \to \infty$,

$$\underline{\gamma}_m \sim [m + \tfrac{3}{2}]\pi, \tag{4.24}$$

where m is a nonnegative integer ($m = 0, 1, 2, \ldots$). Consequently, for any fixed value of m, we obtain an estimate for the solutions of (P_A) as

$$H(\eta) \sim A_0(\eta/L), \qquad \underline{\alpha}_m \sim \left([m + \tfrac{3}{2}]\pi/L\right)^4, \qquad L \to 0. \tag{4.25}$$

The index m gives the number of sign changes of the nearly periodic solutions (see the $m = 0, 1, 2$ solutions in Figure 4.3(left)). As seen in (4.20), the decay of the oscillations is a weak effect in the limit $\alpha \to \infty$ as it enters at next order. In this regime, the index m parametrizes the solution branches in the (L, α) plane, as shown in Figure 4.6(left). In contrast, for $L \to \infty$ we found very different limiting behavior, $\alpha \to \bar{\alpha}_q$, (4.12).

The boundary between these two asymptotic regimes can be estimated by determining where (4.18) yields $O(1)$ corrections to α as $L \to \infty$. The dominant balance of the algebraic powers in (4.18) as $\bar{\alpha} \to \infty$ yields the estimate

$$\alpha = O(L^{4/3}), \qquad L \to \infty, \tag{4.26}$$

for the location of the overlap region where (4.12) and (4.25) must match together; see Figure 4.6(left).

4.3. The boundary value problem (P_B). Now consider the boundary value problem (P_B) for $n = 0$, where the $H'(L) = 0$ condition in (4.1) is replaced by $H'''(L) = 0$:

$$\alpha H + \tfrac{1}{4}\eta H' - H'''' = 0, \qquad 0 \le \eta \le L, \tag{4.27a}$$

$$H(0) = H'(0) = 0, \qquad H(L) = H'''(L) = 0. \tag{4.27b}$$

As in the case of (4.1), we expect this problem to produce continuous families of solutions on curves in the (L, α) parameter plane. The analysis of the solutions of (4.27) closely follows the results given above for (P_A), and so we omit most of the details. For the limit $\alpha \to \infty$, the solutions are given by

$$H(\eta) \sim A_0(\eta/L), \qquad \underline{\alpha}_m \sim ([m+1]\pi/L)^4, \qquad L \to 0, \tag{4.28}$$

for $m = 0, 1, 2, \ldots$. Note that these solution branches for (P_B) alternate in the plane with (4.25) for $\alpha \to \infty$, as shown in Figure 4.6(right). Since the results of the analysis given in section 4.1.1 are independent of the details of the boundary conditions at $\eta = L$, they also apply to the solutions of (P_B). That is, for $L \to \infty$, the solutions of (4.27) also approach $\bar{\alpha}$ given by (4.12), but the details of the oscillatory corrections (4.18) will be somewhat different (see Figure 4.6(right)).

4.4. The structure of the set of dipole similarity solutions. We now combine what has been learned about the structures of subproblems (P_A), (P_B) to obtain the similarity solutions of (P). The dipole solutions are given by the isolated intersection points of the solution branches of the two subproblems, as shown in Figure 4.6(right) and Figure 4.7.

We note that there are no intersections for $L \to 0$ and large α. This is because the branches given by (4.25) and (4.28) are uniformly spaced in this case. Recalling that (4.25) and (4.28) were obtained in the limit where (4.1a) reduced to a formally self-adjoint equation, we can attribute the more complicated structure of the solution branches for moderate values of α, L to the influence of non–self-adjoint effects. In particular, at finite α, these branches are observed either (i) to connect to an adjacent branch, corresponding to solutions with one more (or one less) sign change, to form a *loop*[1] with L bounded from above, or (ii) to extend to $L \to \infty$ by connecting through a bend to one of the branches approaching the first-kind half-line solutions specified by (4.12). In the latter case, from (4.13), the sequence of compactly supported solutions with m sign-changes approaches the half-line solutions, with the size of the support behaving like $L = O(m^{3/4})$ as $m \to \infty$.

[1]With the number of sign changes becoming degenerate, as one zero occurs at $\eta = L$ at the minimum of the loop, where $d\alpha/dL = 0$.

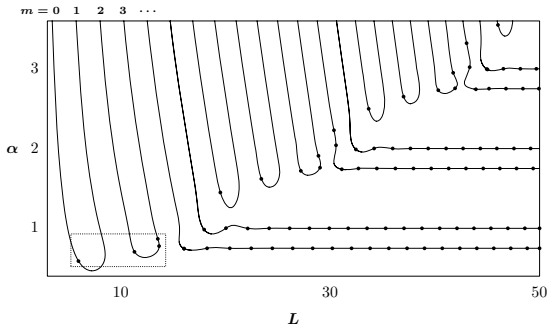


FIG. 4.7. *Dipole similarity solutions (solid dots) shown in the* (L, α) *parameter plane as determined by points on the branches of the linear subproblem* (P_A) *(curves)* (4.1) *where the flux at the interface,* $H'''(L)$*, vanishes. The four solutions enclosed in the boxed region will be considered further in section* 5.3.

In the loop case, we can show that there must be at least one dipole solution on each loop. Consider a loop for (P_A) with solutions given by (4.25) for $L \to 0$. Depending on m, the flux at the interface for small L is either positive or negative along a branch, with

$$H'''(L) \sim \tfrac{1}{2}(-1)^m e^{-(m+\frac{3}{2})\pi}, \qquad L \to 0. \tag{4.29}$$

Since the flux is continuous on a (P_A) loop and the sign of the flux on consecutive branches is opposite, there must be at least one zero of the flux on each loop, and the result follows. Figure 4.7 shows the dipole solutions marked as solid dots on the (P_A) solution branches. Recalling the interpretation of the index m as the number of sign changes of $H(\eta)$, this figure indicates one solution for $m = 0$ (the unique nonnegative solution), no solution for $m = 1$, one solution for $m = 2$, two solutions for $m = 3$, and so on.[2] In contrast to this partially ordered set of solutions, for self-adjoint problems, we would expect exactly one solution for each value of m. The nonexistence or nonuniqueness of solutions with m sign changes is an interesting result of the non–self-adjoint structure of this problem.

5. Limiting behavior for $n \to 0$. Having completely analyzed the linear problem, we build upon it to describe the nonlinear problem for small n. Similar to the "slightly nonlinear" limit used in [22], we express the nonlinear mobility coefficient in (2.3a) as

$$|H|^n = e^{n \ln |H|} \sim 1 + n \ln |H| + O(n^2), \tag{5.1}$$

this expansion being valid everywhere that $|H| \gg e^{-1/n}$, that is, everywhere except exponentially narrow boundary layers at zeros of the solution. Careful analysis is required to fully resolve the local structure there [37], but this will not be crucial in the current problem. Substituting (5.1) into (4.1a) yields the problem

$$\alpha H + \tfrac{1}{4}\eta H' - H'''' = n\left[\tfrac{1}{4}\alpha\eta H' + (\ln|H|H''')'\right] + O(n^2). \tag{5.2}$$

We consider two limits of this problem to describe the influence of weak nonlinearity on the solutions for $n \to 0$.

[2] Observe that there is one solution for $m = 4$, two for $m = 5$, and three for $m = 6$. We will see that the numbers of solutions also depends on n.

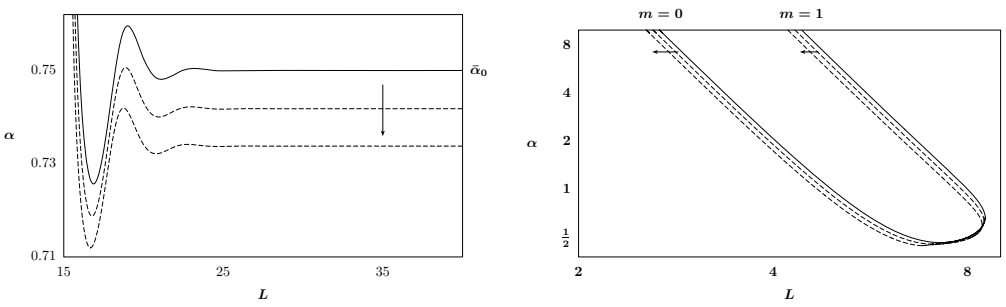


FIG. 5.1. *Shifts of the solution curves in the* (L, α) *plane for* $n \to 0$*: (left) the shift to the* $\bar{\alpha}$ *values for* $n \to 0$*, accurately predicted by* (5.7)*; (right) the first solution loop for* $n = 0$ *(solid curve) and the corresponding curve for* $n = 0.1, 0.2$ *(dashed curves). The shift of the curve for large* α *is accurately predicted by* (5.12).

5.1. Weakly nonlinear solutions on the half-line. Building on the results from section 4.1.1, we seek solutions for the nonlinear problem on the half-line in the form of a regular expansion,

$$H(\eta) = H_0(\eta) + nH_1(\eta) + O(n^2), \qquad \alpha = \alpha_0 + n\alpha_1 + O(n^2), \tag{5.3}$$

where the leading-order solution is $H_0 = \bar{H}(\eta)$, $\alpha_0 = \bar{\alpha}$, given by (4.13) and (4.12). Then, at next order, (5.2) yields

$$\alpha_0 H_1 + \tfrac{1}{4}\eta H_1' - H_1'''' = \tfrac{1}{4}\alpha_0 \eta H_0' + (\ln|H_0| H_0''')' - \alpha_1 H_0. \tag{5.4}$$

We will focus on determining the eigenvalue correction, α_1. To this end, we must enforce the Fredholm alternative on the right-hand side of (5.4) to yield a solvability condition. That is, the inner product of the right-hand side with the solution of the homogeneous adjoint problem must vanish. The linear adjoint problem is

$$\alpha_0 \bar{G} - \tfrac{1}{4}(\eta \bar{G}\,)' - \bar{G}'''' = 0, \qquad \bar{G}(0) = \bar{G}'(0) = 0. \tag{5.5}$$

Note that we have already made use of the solutions of the adjoint problem in section 3 to compute the critical $\bar{\alpha}$ values (3.4). From (3.3), the solutions are the polynomials

$$\{(\bar{G}(\eta); \bar{\alpha})\} = \{(\eta^2; \tfrac{3}{4}), (\eta^3; 1), (\eta^2 + \tfrac{1}{360}\eta^6; \tfrac{7}{4}), \ldots\}. \tag{5.6}$$

Hence, for a given half-line solution, the correction to the $n = 0$ value of $\bar{\alpha}$ is

$$\alpha_1 = -\left(\tfrac{1}{4}\bar{\alpha}\int_0^\infty (\eta\bar{G}\,)'\bar{H}\,d\eta + \int_0^\infty \bar{G}' \ln|\bar{H}|\bar{H}'''\,d\eta\right) \Big/ \int_0^\infty \bar{G}\bar{H}\,d\eta. \tag{5.7}$$

Figure 5.1(left) illustrates the predicted uniform downward shift to the $\bar{\alpha}_0 = \frac{3}{4}$ branch for $n = 0.01, 0.02$ with $\alpha_1 \approx -0.82$.

5.2. Weakly nonlinear solutions for $L \to 0$. Recalling the results of section 4.2, we consider this case in terms of the limit $\alpha \to \infty$ with $L = \gamma/\alpha^{1/4}$ and

$$H(\eta) = A(z), \qquad \eta = \gamma\alpha^{-1/4} z, \tag{5.8}$$

with $0 \le z \le 1$. We then expand the solution of (5.2) as a perturbation series for $n \to 0$,

$$A(z) = A_0(z) + nA_1(z) + O(n^2), \qquad \gamma = \gamma_0 + n\gamma_1 + O(n^2). \tag{5.9}$$

For $\alpha \to \infty$, at leading order we recover (4.21) with solution (4.23) and $\gamma_0 = \underline{\gamma}$, (4.24). Similarly, at next order in n,

$$A_1'''' - \gamma_0^4 A_1 = 4\gamma_1\gamma_0^3 A_0 - \tfrac{1}{4}\gamma_0^4 z A_0' - (\ln|A_0|A_0''')', \tag{5.10}$$

where we have simplified the right-hand side using (4.21).

This problem is self-adjoint, and the Fredholm alternative provides a solvability condition to determine γ_1 from the integral of the product of the right-hand side of (5.10) with $A_0(z)$,

$$\gamma_1 = -\frac{\gamma_0}{32} - \frac{1}{4\gamma_0^3}\left(\int_0^1 A_0' \ln|A_0|A_0'''\,dz \Big/ \int_0^1 A_0^2\,dz\right), \tag{5.11}$$

where we have used $\int_0^1 zAA'\,dz = -\frac{1}{2}\int_0^1 A^2\,dz$. Consequently, the asymptotic form of the eigenvalue curves for the joint limit $L \to 0$ and $n \to 0$ is given by

$$\underline{\alpha} \sim \left([\gamma_0 + n\gamma_1]/L\right)^4. \tag{5.12}$$

This analytical prediction agrees very closely with the numerically computed curves for $m = 0$ ($\gamma_1 \approx -1.04$) and $m = 1$ ($\gamma_1 \approx -1.81$), as shown in Figure 5.1(right). We note that $\gamma_1 < 0$ for all m, and hence all of the curves translate to the left as n increases.

5.3. Computing the solutions of the nonlinear problem. Having established the properties of the similarity solutions for $n = 0$ in section 4 and observing the continuous dependence of the solutions on n for $n \to 0$, we now seek to cast these results in a form that will be convenient for the numerical study of the nonlinear problem. Rescaling the similarity variable so that the solution is defined on a fixed domain, $0 \le z \le 1$,

$$H(\eta) = \mathcal{H}(z), \qquad \eta = Lz, \tag{5.13}$$

the boundary value problem (P_A) then takes the form

$$L^4[\alpha\mathcal{H} + \tfrac{1}{4}(1 - n\alpha)z\mathcal{H}'] - (|\mathcal{H}|^n\mathcal{H}''')' = 0, \tag{5.14a}$$

$$\mathcal{H}(0) = \mathcal{H}'(0) = 0, \qquad \mathcal{H}(1) = \mathcal{H}'(1) = 0, \qquad \int_0^1 \mathcal{H}^2\,dz = 1. \tag{5.14b}$$

Here, the computational domain, $0 \le z \le 1$, is fixed, while L and α both appear in (5.14a) as eigenvalues. (The complete problem, (5.14) with $|\mathcal{H}|^n\mathcal{H}'''(1) = 0$, could be referred to as a nonlinear *double-eigenvalue problem* [24].) We solve this system numerically using Newton–Raphson relaxation for a finite-difference discretization of the problem. Continuation methods [30] are used to trace the solution branches of (5.14) and to obtain the dependence of one parameter on the other, say $\mathcal{H}(z)$ and L in terms of α (and, inherently, n). The dipole similarity solutions are then found at (L, α) values where the no-flux condition at $z = 1$ is also satisfied. A plot of the decay exponent α as a function of n for the first few dipole solutions is shown in Figure 5.2. As expected from (5.3) and (5.12), the figure shows that the $n = 0$ solutions can be smoothly continued to nonzero n over finite ranges. Note the strong influence of n as the two branches of solutions with three zeroes coalesce in a saddle-node-type bifurcation at a negative value of $n \approx -0.056$.

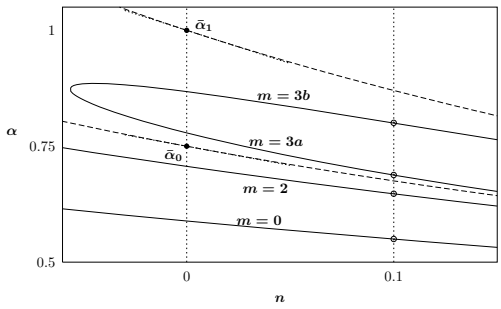


FIG. 5.2. *Numerically computed values of α as a function of n for the first four compactly supported dipole solutions (highlighted in Figure 4.7 for $n = 0$) (solid curves), along with $\bar{\alpha}(n)$ for the halfline solutions (dashed curves), as predicted by $\alpha \sim \bar{\alpha} + n\alpha_1$ (5.7) (dotted lines). Dipole solutions are denoted by their numbers of zeros, m, with two solutions for $m = 3$ labeled 3a and 3b.*

There are various approaches to handling the degeneracy of (1.1) when performing numerical computations of solutions with sign-changes. The numerical scheme employed in this paper involved the introduction of a regularization to smooth the absolute-value function,

$$|\mathcal{H}|^n \to (\mathcal{H}^2 + \epsilon^2)^{n/2}, \tag{5.15}$$

for a small regularization parameter, $\epsilon \ll 1$. Additionally, local analysis for $z \to 0$ and $z \to 1$ carried out in [11, 21] was used to appropriately implement the boundary conditions (5.14b) for different ranges of n.

6. Numerical simulations of the thin-film evolution equation (1.1). Up to this point, we have analyzed the dipole similarity solutions as solutions of the boundary value problem (2.3). We now take the opportunity to examine the role of these solutions in terms of the dynamics of the nonlinear evolution equation (1.1).

Making use of the regularization (5.15) with $\epsilon = 10^{-4}$, we consider general numerical solutions of the thin-film equation (1.1) via the regularized evolution equation

$$\partial_t h = -\partial_x([h^2 + \epsilon^2]^{n/2} h_{xxx}) \tag{6.1}$$

on a finite domain, $0 \le x \le 40$, with boundary conditions $h = h_x = 0$ at both boundaries. Note that this regularization differs from the nonnegativity-preserving form suggested by Bernis and Friedman [10], frequently used in numerical studies of (1.1) [18, 19], as we wish to explore what happens when the solutions are allowed to change sign.

We illustrate different aspects of the dipole solutions within the dynamics of the thin film equation in the following two numerical simulations, first for $n = 1$ and then for n small ($n = 0.1$).

6.1. Dynamics for $n = 1$. We consider (6.1) with $n = 1$ and nonnegative compactly supported initial data, $h(x, 0) = h_0(x) \equiv (1 - [x - 5]^2)^2$ for $4 \le x \le 6$, and $h(x, 0) \equiv 0$ elsewhere (see Figure 6.1(upper left)). For initial value problems for the thin-film equation with $n > 0$, regions of support are known to vary in time with finite speed of propagation. Hence, until the support has grown to include the boundaries $x = 0$ and $x = 40$, those respective boundary conditions will not influence the evolution of the solution. Indeed, this leads to several stages of *intermediate dynamics* [2], as follow:

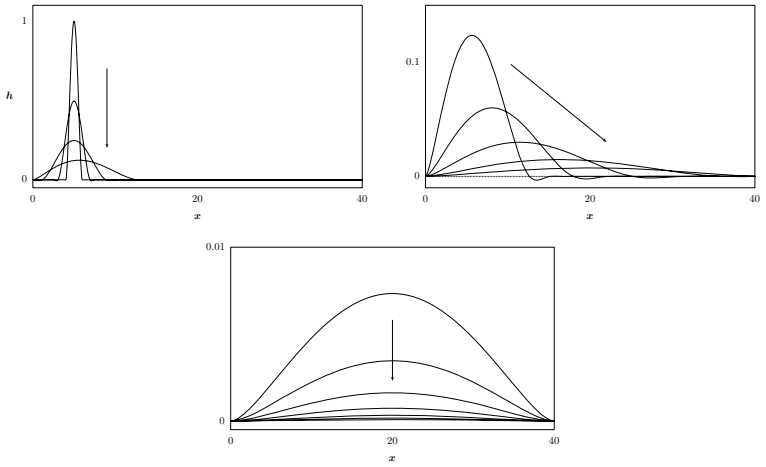


FIG. 6.1. *Profiles at various times from a numerical simulation of* (1.1) *with* $n = 1$*, illustrating the three stages of self-similar intermediate dynamics: (upper left) pure spreading of the Cauchy-type solution, (upper right) mixed spreading/draining in the dipole problem, (bottom) pure draining on the bounded domain.*

1. *Source-type evolution for the Cauchy problem.* For short times, before either boundary is within the region of support, the solution is effectively expanding within a homogeneous region of unlimited extent, and hence will take the form of the source-type similarity solution of the Cauchy problem [9, 13, 14]. This is a first-kind similarity solution with $\alpha = \beta = 1/5$ for $n = 1$ and $h(x,t) = \tau^{-\alpha} H(\eta)$, with

$$H(\eta) = \left(1 - \eta^2\right)^2 \quad \text{for } |\eta| \le 1, \qquad \eta = (x - x_0)/\tau^{1/5}. \tag{6.2}$$

   This solution conserves both $\int h\,dx$ and $\int xh\,dx$. Note that for convenience our initial data was selected to be exactly of this form, with $x_0 = 5$ and the scaled time variable $\tau = 120t + 1$. The similarity solution is stable and acts as an attractor for more general initial data [14]. This solution expands in a symmetric manner with respect to x_0, and we refer to this dynamical regime as being *pure-spreading*. This regime ends when the support extends to the $x = 0$ near-boundary (see Figure 6.1(upper left)).
2. *Mixed "spreading/draining" behavior for the dipole problem.* Once the $x = 0$ boundary has been reached, the boundary conditions dictate that $\int h\,dx$ is no longer conserved. Meanwhile, the right-interface for the region of support continues to expand into the domain; this suggests a transition to the dipole problem, which can also be described as *mixed spreading/draining*. A simple calculation for the first moment for $n > 0$ shows that

$$\frac{d}{dt}\left(\int_0^\infty xh\,dx\right) = \frac{1}{2}n(n-1)\int_0^\infty |h|^{n-2} h_x^3\,dx, \tag{6.3}$$

   subject to contributions from boundary terms vanishing. So, for $n = 1$ the first moment is a conserved quantity (see [11]). Consequently, we can determine $\alpha = 1/3$, $\beta = 1/6$ for this special case, and the dipole similarity solution is of the first kind. This regime ends when the support extends to the $x = 40$ far-boundary (see Figure 6.1(upper right)).

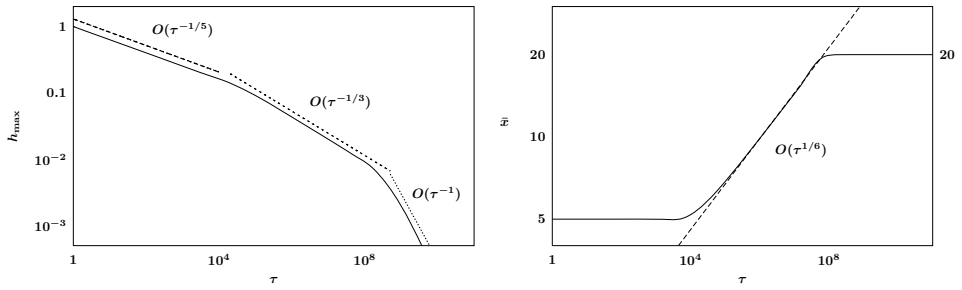


FIG. 6.2. *Graphs illustrating changes in the self-similar regimes for $n = 1$: (left) the solid curve is $h_{\max}(t)$ from the numerical computation, while the dashed lines have slopes $-1/5, -1/3$, and -1 corresponding to analytic predictions for self-similar pure spreading, mixed spreading/draining, and pure draining respectively; (right) similarly, the computed and predicted values of $\bar{x}$. The appropriately scaled/shifted time-variable here is $\tau = 120t + 1$.*

3. *"Pure draining" behavior on the finite domain.* Once the region of support has extended to the entire domain, the behavior of the solution will be controlled by the boundary conditions at both edges. The solution approaches a separable form

$$h(x,t) = \tau^{-1} H(x - 20), \tag{6.4}$$

and hence $\alpha = 1$ and $\beta = 0$, as studied in [22, 38]; the symmetry of (6.4) is apparent in Figure 6.1(bottom).

We note that the transitions between stages that we are outlining also occur for the analogous porous medium problem, where the origin of the dipole solution is usually justified by other means (see, for instance, [2]); the equivalent analysis for the one-dimensional heat equation can be found in [1].

Some comments on the influence of the regularization in (6.1) are necessary. The solutions of this modified PDE can be hoped to agree with results from the thin-film equation everywhere that $|h| \gg \epsilon$. The regularization is nonnegligible for $|h| \lesssim \epsilon$, where (6.1) effectively becomes a rescaled version of the $n = 0$ linear problem. This can make accurate statements about the edge of the support and details of the solutions at sign changes obtained from the numerical simulations open to question. Indeed, the solutions in Figure 6.1(upper left) and (upper right) do become negative (see also Figure 1.1), and interestingly the dipole solution appears to have only a single sign change ($m = 1$) (a solution not present for the $n \to 0$ limit). Furthermore, for longer times (not shown), when the solution everywhere satisfies $|h| \lesssim \epsilon$, the evolution will be dominated by the regularization to yield a linear "draining" behavior (see section 5.1 of [22]). To circumvent these difficulties, we use two measures that are not sensitive to the regularization to establish the agreement of the simulations with the analytic predictions given above: (i) the maximum of the solution, $h_{\max}(t) = \max_x |h(x,t)|$, and (ii) the quantity $\bar{x} = \int xh\,dx / \int h\,dx$ (see Figure 6.2). The slope of $h_{\max}(t)$ on a log-log graph gives the decay exponent α; the position of the maximum, $x_{\max}(t)$, would similarly yield the spreading exponent β.

While the special property of the $n = 1$ case, provided by the conservation law (6.3), gives an analytic prediction for the value of α, this means that all similarity solutions for $n = 1$ of the form (2.1) are first-kind and will have $\alpha = 1/3$. Hence plots of $h_{\max}(t)$ cannot be used to distinguish between different similarity solutions that may coexist. Our numerical calculations for (2.3) suggest that at least two solutions

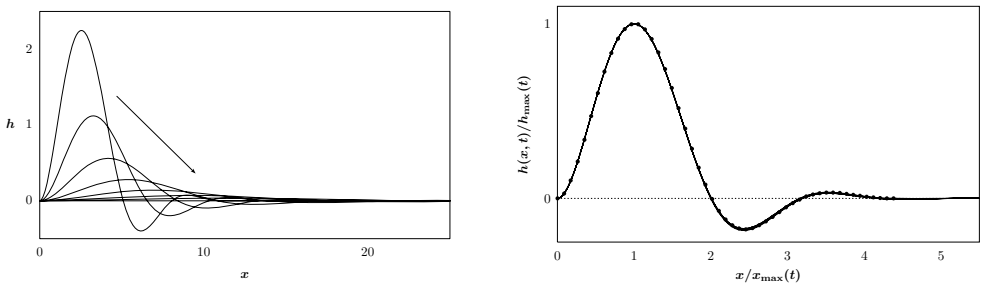


FIG. 6.3. *(left) Solution profiles for* (6.1) *with* $n = 0.1$ *starting from the* $m = 2$ *dipole profile* $H(x)$*; (right) the same profiles rescaled by* $x_{\max}(t)$*,* $h_{\max}(t)$ *(seven almost indistinguishable solid curves) and compared with the compactly supported similarity solution (dots).*

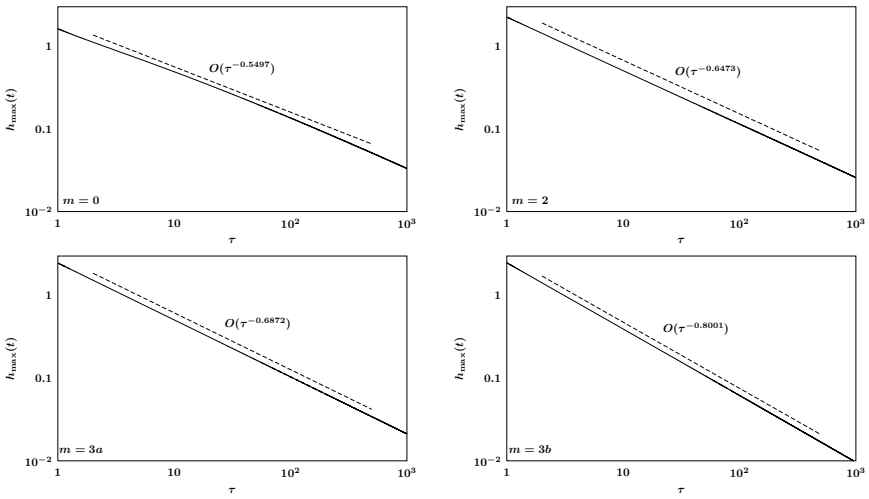


FIG. 6.4. *Log-log plots of* $h_{\max}(t)$ *versus* $\tau = 1 + t$ *for numerical solutions of* (6.1) *with* $n = 0.1$ *and initial conditions given by self-similar profiles* $H(x)$*. Comparisons are shown with the expected decay rates (dashed lines) for the respective dipole similarity solutions:* $m = 0\,(\alpha \approx 0.5497)$*,* $m = 2\,(\alpha \approx 0.6473)$ *(see Figure* 6.3*),* $m = 3a\,(\alpha \approx 0.6872)$*,* $m = 3b\,(\alpha \approx 0.8001)$*.*

($m = 0$ (positive), and $m = 1$ (single sign change)), and maybe more, exist at $n = 1$; the stability of these solutions is still an open problem.

We now turn to numerical simulations for $n \neq 1$, where $h_{\max}(t)$ can be used to provide numerical evidence for the coexistence of multiple stable similarity solutions.

6.2. Coexistence of multiple stable dipole solutions. We consider numerical simulations of (6.1) with $n = 0.1$ on a sufficiently large domain. For this value of n we have predictions for the dipole similarity solutions (see Figure 5.2). We use the compactly supported numerical solutions of (2.3) as the initial conditions for the simulations of (1.1), $h_0(x) = H(x)$ for $0 \le x \le L$, and $h_0 \equiv 0$ for $x > L$, i.e., at $t = 0$, $x = \eta$ with $\tau = t + 1$. We carry out four such computations for evolutions starting from the similarity profile for $m = 0$ (the unique nonnegative solution), $m = 2$ (the unique two-sign-change solution), and the two similarity profiles with three sign changes (called $m = 3a$ and $m = 3b$ in Figure 5.2). Figure 6.3 shows the evolution starting from the $m = 2$ profile—the regularization appears to have negligible influence on the collapse of the rescaled profiles onto the similarity profile. In fact, Figure 6.4 shows that in all

four of the simulations, the evolution follows the predicted self-similar scalings found from the dipole similarity solutions. The conclusion suggested by these simulations is that all of these similarity solutions are observable in the solution dynamics of (1.1) and are stable with respect to the perturbations introduced by the numerics and the regularization.

The sense of their "stability" needs to be qualified slightly. The compactly supported solutions of the similarity solution problem (2.3) do not precisely satisfy (6.1) on the entire domain due to lack of higher order smoothness at the moving interface. In fact, the dynamics will include very slow-scale evolution towards one of the analytic similarity solutions $\bar{H}(\eta)$ defined on the entire half-line for very long times. This drift is slightly visible in the plot for $m = 0$ in Figure 6.4 but has negligible influence for the higher-order solutions, where the discontinuities at the moving interface are slightly less pronounced. This behavior is generic for $n = 0$ and n small with the regularization (6.1).

7. Discussion. The (L, α) pair for a dipole solution provides information on both the decay rate of solutions (via (2.1)) and the speed of the interface, i.e., $\ell'(t) \sim \beta L \tau^{\beta-1}$. The (L, α) parameter plane for $n = 0$, illustrated in Figure 4.7, indicates the complexity associated with finding second-kind similarity solutions for (1.1). For a given value of n, multiple second-kind similarity solutions can be constructed, and there is no complete natural ordering for the solutions in terms of the number of sign-changes that they possess. For example, when $n = 0$, somewhat surprisingly, no dipole solutions with a single sign-change exist, while it is possible to find two solutions that change sign three times. This is consistent with the absence of a theory for higher-order problems analogous to the Sturm–Louiville theory, which yields a well-ordered set of solutions for second-order problems.

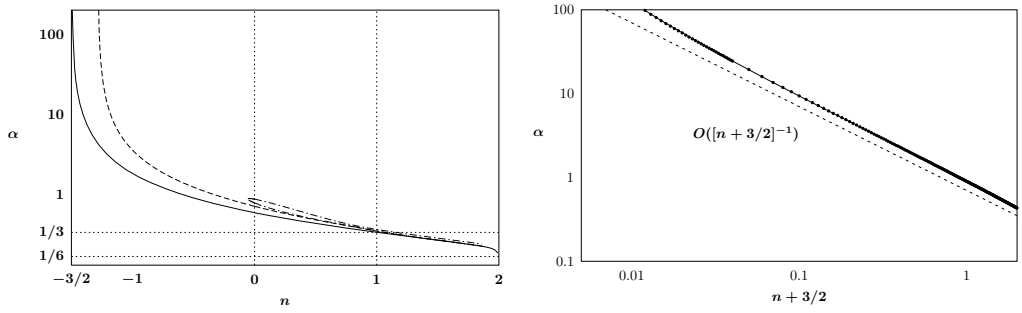


FIG. 7.1. *(left) The decay exponent α for the first few dipole similarity solutions, calculated from* (5.14) *over a wide range of n values. The positive ($m = 0$) solution exists on $-3/2 < n < 2$, while the solutions with sign-changes exist on subsets of this interval. (right) A log-log plot of numerically calculated values for the $m = 0$ solution (solid curve with dots), illustrating the approach to $\alpha \sim O([n + 3/2]^{-1}) \to \infty$ (dashed curve) as $n \to -3/2^+$.*

Figure 7.1 shows the n-dependence of the decay exponent for the dipole solutions over a wider range of n. One of the most important features of the figure is that as $n \to -3/2^+$ we lose existence of not only the sign-change solutions, but also the nonnegative dipole solution. Previous studies of mass-conserving similarity solutions and separable solutions, of which (6.2) and (6.4) are respectively representative, have shown that the former is valid down to $n = -4$, while the latter is applicable as long as $n > 0$. As the boundary conditions on the dipole solutions effectively combine parts of both problems, it is expected that $-4 < n_c < 0$, where n_c indicates the value

of n below which dipole solutions of the form (2.1) no longer exist; the work [11] has shown that, for large enough negative n, finite time extinction dipole solutions can be expected. The numerical simulations are extremely compelling in suggesting that $n_c = -3/2$ (see Figure 7.1(right)), although we do not have any analytical verification of this hypothesis. The effect of the global conservation law for $n = 1$, (6.3), is also clearly seen in Figure 7.1. The self-similar dipole solutions, by definition, satisfy (6.3) and so have α-exponents which should all approach $\alpha = 1/3$ as $n \to 1$. The small discrepancies are a consequence of errors introduced into the simulations by the regularization (6.1).

We have shown, when $n = 0$, that there are critical values of α (see (3.4)) for which first-kind similarity solutions exist featuring an infinite number of sign-changes. These solutions necessarily have $L = \infty$, and the results (4.17) and (4.18) yield that for each first-kind similarity solution there exists a compactly supported, second-kind solution where $|\alpha - \bar{\alpha}| < \delta$ for arbitrary δ; the size of the support $L \to \infty$ as $\delta \to 0$. This can lead to difficulties distinguishing sensitive details of the dynamics of solutions.

Acknowledgments. The authors wish to thank Professor J. R. King for stimulating discussions.

REFERENCES

[1] G. I. Barenblatt, *Scaling, Self-Similarity, and Intermediate Asymptotics*, Cambridge University Press, Cambridge, UK, 1996.
[2] G. I. Barenblatt, *Scaling*, Cambridge University Press, Cambridge, UK, 2003.
[3] G. I. Barenblatt and J. L. Vazquez, *A new free boundary problem for unsteady flows in porous media*, European J. Appl. Math., 9 (1998), pp. 37–54.
[4] G. I. Barenblatt and Ya. B. Zeldovich, *On dipole solutions in problems of nonstationary filtration of gas under polytropic regime*, Prikl. Mat. Meh., 21 (1957), pp. 718–720 (in Russian).
[5] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*. I, Springer-Verlag, New York, 1999.
[6] E. Beretta, M. Bertsch, and R. Dal Passo, *Nonnegative solutions of a fourth-order nonlinear degenerate parabolic equation*, Arch. Ration. Mech. Anal., 129 (1995), pp. 175–200.
[7] F. Bernis, *Finite speed of propagation and continuity of the interface for thin viscous flows*, Adv. Differential Equations, 1 (1996), pp. 337–368.
[8] F. Bernis, *Finite speed of propagation for thin viscous flows when* $2 \le n < 3$, C. R. Acad. Sci. Paris Sér. I, 322 (1996), pp. 1169–1174.
[9] F. Bernis, *The free boundary of thin viscous flows*, ZAMM Z. Angew. Math. Mech., 76 (1996), pp. 369–372.
[10] F. Bernis and A. Friedman, *Higher order nonlinear degenerate parabolic equations*, J. Differential Equations, 83 (1990), pp. 179–206.
[11] F. Bernis, J. Hulshof, and J. R. King, *Dipoles and similarity solutions of the thin film equation in the half-line*, Nonlinearity, 13 (2000), pp. 413–439.
[12] F. Bernis, J. Hulshof, and F. Quirós, *The "linear" limit of thin film flows as an obstacle-type free boundary*, SIAM J. Appl. Math., 61 (2000), pp. 1062–1079.
[13] F. Bernis, L. A. Peletier, and S. M. Williams, *Source type solutions of a fourth order nonlinear degenerate parabolic equation*, Nonlinear Anal., 18 (1992), pp. 217–234.
[14] A. J. Bernoff and T. P. Witelski, *Linear stability of source-type similarity solutions of the thin film equation*, Appl. Math. Lett., 15 (2002), pp. 599–606.
[15] A. L. Bertozzi, *Symmetric singularity formation in lubrication-type equations for interface motion*, SIAM J. Appl. Math., 56 (1996), pp. 681–714.
[16] A. L. Bertozzi and M. Pugh, *The lubrication approximation for thin viscous films: Regularity and long-time behaviour of weak solutions*, Comm. Pure. Appl. Math., 49 (1996), pp. 85–123.
[17] A. L. Bertozzi, *Lubrication approximations for surface tension driven interfaces: Some open problems*, ZAMM Z. Angew. Math. Mech., 76 (1996), pp. 373–376.
[18] A. L. Bertozzi, *The mathematics of moving contact lines in thin liquid films*, Notices Amer. Math. Soc., 45 (1998), pp. 689–697.

[19] A. L. Bertozzi and M. C. Pugh, *Long-wave instabilities and saturation in thin film equations*, Comm. Pure Appl. Math., 51 (1998), pp. 625–661.
[20] M. Bowen, *High Order Diffusion*, Ph.D. thesis, School of Mathematical Sciences, Division of Theoretical Mechanics, University of Nottingham, Nottingham, UK, 1998.
[21] M. Bowen, J. Hulshof, and J. R. King, *Anomalous exponents and dipole solutions for the thin film equation*, SIAM J. Appl. Math, 62 (2001), pp. 149–179.
[22] M. Bowen and J. R. King, *Asymptotic behaviour of the thin film equation in bounded domains*, European J. Appl. Math., 12 (2001), pp. 135–157.
[23] P. Constantin, T. F. Dupont, R. E. Goldstein, L. P. Kadanoff, M. J. Shelley, and S. Zhou, *Droplet breakup in a model of the Hele-Shaw cell*, Phys. Rev. E, 47 (1993), pp. 4169–4181.
[24] L. Fox, L. Hayes, and D. F. Mayers, *The double eigenvalue problem*, in Topics in Numerical Analysis (Proceedings of the Royal Irish Academy Conference, University College, Dublin, 1972), Academic Press, London, 1973, pp. 93–112.
[25] J. B. Greer and A. L. Bertozzi, *H^1 solutions of a class of fourth order nonlinear equations for image processing*, Discrete Contin. Dynam. Systems, 10 (2004), pp. 349–366.
[26] J. B. Greer and A. L. Bertozzi, *Traveling wave solutions of fourth order PDEs for image processing*, SIAM J. Math. Anal., 36 (2004), pp. 38–68.
[27] J. Hulshof, *A local analysis of similarity solutions of the thin film equation*, in Nonlinear Analysis and Applications (Warsaw, 1994), Gakuto Internat. Ser. Math. Sci. Appl. 7, Gakkōtosho, Tokyo, 1996, pp. 179–192.
[28] J. Hulshof, J. R. King, and M. Bowen, *Intermediate asymptotics of the porous medium equation with sign changes*, Adv. Differential Equations, 6 (2001), pp. 1115–1152.
[29] J. Hulshof and J. L. Vázquez, *The dipole solution for the porous medium equation in several space dimensions*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 20 (1993), pp. 193–217.
[30] H. B. Keller, *Numerical Methods for Two-Point Boundary Value Problems*, Dover Publications, New York, 1992.
[31] J. R. King, *Integral results for nonlinear diffusion equations*, J. Engrg. Math., 25 (1991), pp. 191–205.
[32] J. R. King and M. Bowen, *Moving boundary problems and non-uniqueness for the thin film equation*, European J. Appl. Math., 12 (2001), pp. 321–356.
[33] S. E. King and A. W. Woods, *Dipole solutions for viscous gravity currents: Theory and experiments*, J. Fluid Mech., 483 (2003), pp. 91–109.
[34] A. A. Lacey, *The motion with slip of a thin viscous droplet over a solid surface*, Stud. Appl. Math., 76 (1982), pp. 217–230.
[35] T. G. Myers, *Thin films with high surface tension*, SIAM Rev., 40 (1998), pp. 441–462.
[36] H. Ockendon and J. R. Ockendon, *Viscous Flow*, Cambridge University Press, Cambridge, UK, 1995.
[37] N. F. Smyth and J. M. Hill, *High-order nonlinear diffusion*, IMA J. Appl. Math., 40 (1988), pp. 73–86.
[38] G. J. B. van den Berg, M. Bowen, J. R. King, and M. M. A. El-Sheikh, *The self-similar solution for draining in the thin film equation*, European J. Appl. Math., 15 (2004), pp. 329–346.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1749–1775

EXACT ARTIFICIAL BOUNDARY CONDITIONS FOR CONTINUUM AND DISCRETE ELASTICITY*

SUNMI LEE†, RUSSEL E. CAFLISCH‡, AND YOUNG-JU LEE†

Abstract. For the continuum and discrete elastic equations, we derive exact artificial boundary conditions (ABCs), often referred to as transparent boundary conditions, that can be applied at a planar interface below which there are no forces. Solution of the elasticity equations can then be performed using this interface as an artificial boundary, often with greatly reduced computational effort, but without loss of accuracy. A general solvability requirement is presented for the existence of an artificial boundary operator for discrete systems (such as discrete elasticity) on an unbounded (semi-infinite) domain. The solvability requirement is validated by introducing a sum-of-exponentials ansatz for the solution below the artificial boundary. We also derive a new expression for the total energy for the system, involving only the region above the artificial boundary. Numerical examples are provided to confirm and illustrate the accuracy and effectiveness of the results.

Key words. elasticity, discrete elasticity, artificial boundary conditions, transparent boundary conditions, atomistic strain

AMS subject classifications. Primary, 65N55; Secondary, 74B05, 70C20

DOI. 10.1137/050644252

1. Introduction.

Many of the boundary value problems arising in applied mathematics are formulated on unbounded domains. It is in general a nontrivial task to solve such problems numerically [6], since the numerical solution naturally requires boundary conditions at a finite depth in the body.

The main motivation of the present work comes from the numerical simulation of strain fields in semi-infinite domains. For the strain equations, the use of a physical boundary condition, such as the zero displacement field at a certain depth, has been a common practice [21]. On the other hand, due to the long range of elastic interactions, the zero boundary condition must be imposed at considerable depth in order to accurately compute the strain field [4], which entails large computational cost.

The purpose of this paper is to derive exact *artificial boundary conditions* (ABCs) such that the solution on the (bounded) computational domain coincides with the exact solution on the unbounded domain. Such exact artificial boundary conditions are oftentimes referred to as *transparent boundary conditions* (TBCs) [6].

There have been various works on ABCs for a wide range of problems. For example, certain ABCs for the Poisson and Helmholtz equations on infinite domains are investigated in [1] using domain decomposition and Fourier techniques. For general elliptic problems, approximate ABCs and error estimates are performed within the finite element framework in [3]. Boundary element methods for homogeneous elasto-

*Received by the editors November 4, 2005; accepted for publication (in revised form) April 20, 2006; published electronically July 31, 2006. This research was supported in part by the MARCO Center on Functional Engineered NanoArchitectonics (FENA) and by the NSF through grant DMS-0402276.

http://www.siam.org/journals/siap/66-5/64425.html

†Department of Mathematics, University of California at Los Angeles, 520 Portola Plaza, Los Angeles, CA 90095 (mathever@gmail.com, yjlee@math.ucla.edu, http://www.math.ucla.edu/∼yjlee). The first author was partially supported by the National Institute for Mathematical Sciences, Korea.

‡Department of Mathematics and Department of Materials Science and Engineering, University of California at Los Angeles, 520 Portola Plaza, Los Angeles, CA 90095 (caflisch@math.ucla.edu, http://www.math.ucla.edu/∼caflisch).

static and elastodynamic cases, linear elastostatic problems, time dependent heat and wave equations, and electromagnetic scattering problems are also treated in an exact manner using the Dirichlet to Neumann boundary condition in [2, 7, 8, 9].

For the elasticity problem, several local and nonlocal artificial boundary conditions are provided in terms of the finite element formulation in [12, 13, 14]. For a discrete elastic strain model for an epitaxial thin film, ABCs were derived recently by Russo and Smereka [20] using a formulation that is somewhat different from our model.

In the present work, we perform an analysis for the equations of both continuum and discrete elastic models. The discrete elastic equations correspond to an atomistic strain model introduced in the recent work by Schindler et al. [21]. Although full details are provided only for a discrete strain model, a general solvability requirement is formulated, which results in the well-posedness or the solvability of the system in an infinite domain. This work is a discrete analogue of the work by Hagstrom and Keller [11]. The solvability requirement is then validated by analyzing the solution on the exterior domain using a sum-of-exponentials ansatz. This framework, on the one hand, leads us to derive the abstract ABC operator in the form of a Schur complement operator and, on the other hand, guides the construction of the explicit ABC operator for actual implementations. Thanks to the ABC operator, the force balance equation that needs to be solved in the infinite domain can be posed as a reduced equation on the bounded domain, whose solution has been shown to coincide with the exact solution on the full (unbounded) domain. In addition, a new formula is derived for the total elastic energy of the system, involving only the solution above the artificial boundary. The latter is particularly important for practical applications such as thin epitaxial film growth simulations.

The rest of the paper is structured as follows. In section 2, we introduce some preliminaries and notation to ease the presentation. The ABCs, total energy formula, and variational principle for continuum elasticity are derived in section 3. In section 4, we briefly review the discrete elastic strain model and introduce the general solvability requirement, present an abstract form of the ABC operator, and derive explicit ABCs for a specific discrete strain model. The total energy formula and the variational principle for the discrete strain model are also presented. Several illustrative numerical results are provided in section 5. Conclusions are discussed in section 6. Some details are saved for the appendix.

2. Preliminaries. Suppose that the domain Ω is a half-infinite body, e.g., $\Omega = \{(x, y, z) \in \mathbb{R}^3 : z < h(x, y)\}$ for $h : \mathbb{R}^2 \mapsto \mathbb{R}$ being a bounded function. See Figure 2.1 for a schematic description. The interface Γ_2 on which the artificial boundary will be imposed is illustrated in Figure 2.1. For both the continuum and discrete problems, the domain Ω is divided into a finite part Ω_1 and a semi-infinite part (an exterior domain) $\Omega_2 = \Omega \backslash \Omega_1$. The requirement on the choice of Ω_2 is that its boundary Γ_2 is planar and normal to the depth variable and that there are no external forces in Ω_2. For the boundary condition for both continuum and discrete elasticity equations, we assume that the periodic conditions are imposed in x- and y-directions (lateral directions) and that the Neumann condition (i.e., the variational principle with no constraint at the boundary) is imposed on the top layer Γ_1 unless explicitly stated otherwise. Use of the Neumann condition is only for simplicity and to ensure that the problem is well-posed; it does not influence the resulting ABCs.

We use boldface lower case letters for vectors in $\mathbb{R}^d$ with $d = 2$ or 3 and boldface capital letters for symmetric tensors or square matrices. The differential operator ∂_k denotes the partial derivative with respect to the kth coordinate variable, i.e., $\partial/\partial x_k$,

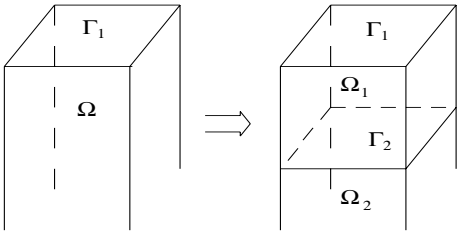


FIG. 2.1. *The domain decomposition: An artificial boundary* Γ_2 *(the horizontal plane) divides* Ω *into* Ω_1 *and* Ω_2. Γ_1 *is the top boundary (surface) of* Ω.

and the operator $\nabla\cdot$ is the standard *divergence* operator defined through

$$\nabla\cdot = (\partial/\partial x, \partial/\partial y)\cdot \qquad \text{for } d = 2,$$
$$\nabla\cdot = (\partial/\partial x, \partial/\partial y, \partial/\partial z)\cdot \qquad \text{for } d = 3.$$

The notation ∇ denotes the usual *gradient* operator for $d = 2$ and $d = 3$ given, respectively, as

$$\nabla = \begin{pmatrix} \partial/\partial x \\ \partial/\partial y \end{pmatrix}, \quad \nabla = \begin{pmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{pmatrix},$$

and Δ is the *Laplace* operator $\nabla\cdot\nabla$.

For two vectors $\mathbf{u}$ and $\mathbf{v}$, $\mathbf{u}\cdot\mathbf{v}$ is the dot product; for a vector $\mathbf{v} = (v_k)_{k=1,\dots,d}$ and a tensor $\mathbf{N} = (N_{kl})_{k,l=1,\dots,d}$, $\mathbf{v}\cdot\mathbf{N} = \sum_{k=1}^{d} v_k N_{k\ell}$. The magnitude of a vector $\mathbf{u}$ will be denoted by $|\mathbf{u}| = (\mathbf{u}\cdot\mathbf{u})^{1/2}$.

Although the letters i, j, k are used for indices, we shall also use $\imath$ to denote the imaginary unit $\sqrt{-1}$, and the complex conjugate of a complex number v shall be denoted by $\overline{v}$. Also, for the matrix $\mathbf{N}$, $\mathbf{N}^H$ and $\mathbf{N}^T$ denote the complex conjugate transpose and the real transpose of $\mathbf{N}$, respectively. Finally, we shall use χ to denote the usual characteristic function that is defined as

$$\chi(x) = \begin{cases} 1 & \text{for } x \in \Omega_1, \\ 0 & \text{for } x \notin \Omega_1. \end{cases} \tag{2.1}$$

Some other notation will be introduced in each section as necessary.

3. The ABCs for continuum elasticity. In this section, we review the continuum elastic equations from an energetic viewpoint. We then derive the artificial boundary (or ABC) operator $\mathcal{A}$, as well as a new expression for the total energy and a formulation of the force balance equations depending on only the displacement on and above the interface Γ_2 on which the artificial boundary condition is given.

3.1. Continuum elasticity. Continuum elasticity is formulated in terms of a displacement field $\mathbf{u} = \mathbf{u}(\mathbf{x}) = \mathbf{y}(\mathbf{x}) - \mathbf{x}$ between the equilibrium position $\mathbf{x}$ of a material point and the elastically deformed position $\mathbf{y}(\mathbf{x})$ of that point. The strain tensor $\mathbf{S}$ has components defined as $S_{k\ell} = (\partial_k u_\ell + \partial_\ell u_k)/2$ in which u_k are the components of $\mathbf{u}$.

The derivation of the linear elasticity equations can be made via a variational principle for the total energy $\mathcal{E}$ in a domain Ω, namely,

$$\delta\mathcal{E} = 0. \tag{3.1}$$

The total elastic energy $\mathcal{E}$ for the linear elasticity is given as follows:

$$\mathcal{E} = \int_\Omega E d\mathbf{x}, \tag{3.2}$$

where the integrand is the energy density

$$E = \frac{1}{2}\sum_{k,\ell} S_{k\ell} T_{k\ell} - \mathbf{u}\cdot\mathbf{f}\,\chi, \tag{3.3}$$

$\mathbf{f} = (f_k)$ is a body force, and $\mathbf{T} = (T_{k\ell})$ is the stress tensor defined, for an isotropic material, as

$$T_{k\ell} = \lambda\delta_{k\ell}\sum_i S_{ii} + 2\tau S_{k\ell}. \tag{3.4}$$

The parameters λ and τ are the Lamé constants. In the absence of external force on the boundary Γ_1, (3.1) reduces to the classical Navier equations of linear elasticity, i.e.,

$$\begin{aligned} -\nabla\cdot\mathbf{T} &= \mathbf{f}\chi \quad \text{in } \Omega, \\ \mathbf{n}\cdot\mathbf{T} &= 0 \quad \text{on } \Gamma_1, \end{aligned} \tag{3.5}$$

where $\mathbf{n}$ is the outer unit normal vector.

For linear elasticity with cubic symmetry, the elastic energy density E is the following:

$$E = \frac{C_{11}}{2}\sum_i S_{ii}^2 + 2C_{44}\sum_{k\neq\ell} S_{k\ell}^2 + C_{12}\sum_{k\neq\ell} S_{kk}S_{\ell\ell}, \tag{3.6}$$

where C_{11}, C_{44}, and C_{12} are the cubic elastic moduli, i.e., the Voigt constants. The linear elasticity equations with cubic symmetry are

$$\begin{aligned} -C_{11}\partial_k\partial_k u_k - C_{44}\sum_{l\neq k}\partial_l\partial_l u_k \\ -(C_{12}+C_{44})\sum_{l\neq k}\partial_k\partial_l u_l = f_k\chi \quad \text{in } \Omega \end{aligned} \tag{3.7}$$

for $k = 1,\dots,d$. Note that the isotropic linear elasticity equations (3.5) can be recovered from (3.7) by choosing the following Voigt constants:

$$(C_{11}, C_{44}, C_{12}) = (\lambda+2\tau, \tau, \lambda). \tag{3.8}$$

For the study of the ABCs for continuum elasticity, we restrict our attention to the isotropic linear elasticity, namely, (3.7) with the Voigt constants given in (3.8), for simplicity. It is easily generalized to the anisotropic case.

3.2. Two dimensional case. In this section, we construct the artificial boundary operator $\mathcal{A}$ for the two dimensional case. The main idea is to analytically solve the force balance equation (3.7) on the exterior domain Ω_2 by introducing a sum-of-exponentials ansatz, which must be modified to include algebraic terms.

We assume that the solution is periodic in the x-direction with 2π periodicity and that the interface Γ_2 is a line, i.e., $\Gamma_2 = \{(x,y) \in \mathbb{R}^2 : y = 0\}$. We first look for a modal solution $\mathbf{u}(x,y)$ for $y < 0$ as

$$\begin{aligned}\mathbf{u}(x,y) &= \widehat{\mathbf{u}}(\mu,y)\,\mathrm{e}^{\imath\mu x} \\ &= \widehat{\mathbf{u}}(\mu)\,\mathrm{e}^{\beta y}\,\mathrm{e}^{\imath\mu x} = \begin{pmatrix} \hat{u}(\mu) \\ \hat{v}(\mu) \end{pmatrix}\mathrm{e}^{\beta y}\,\mathrm{e}^{\imath\mu x}.\end{aligned} \tag{3.9}$$

Since $\mathbf{u}$ in (3.9) is the solution to (3.7), for each μ, $\widehat{\mathbf{u}}(\mu)$ should satisfy the following linear system:

$$\mathbf{M}(\mu,\beta)\widehat{\mathbf{u}}(\mu) = 0,$$

where $\widehat{\mathbf{u}}(\mu) = (\widehat{u}(\mu), \widehat{v}(\mu))^T$ and

$$\mathbf{M}(\mu,\beta) = \begin{pmatrix} -(\lambda+2\tau)\mu^2 + \tau\beta^2 & \imath\mu(\lambda+\tau)\beta \\ \imath\mu(\lambda+\tau)\beta & -\tau\mu^2 + (\lambda+2\tau)\beta^2 \end{pmatrix}.$$

A nontrivial solution can be attained only if

$$\det \mathbf{M}(\mu,\beta) = \tau(\lambda+2\tau)(\beta^2-\mu^2)^2 = 0, \tag{3.10}$$

which implies that $\beta = \pm|\mu|$. Since the solution $\mathbf{u}$ should decay as $y \to -\infty$, then $\beta = |\mu|$ is the proper choice. Note that for $\mu = 0$, the only solution is $\beta = 0$, which corresponds to a trivial solution, the constant displacement field.

We now compute the zero eigenvector for $\mathbf{M}(\mu,|\mu|)$. It is easy to see that the matrix $\mathbf{M}(\mu,|\mu|)$ has a zero eigenvector given by $\mathbf{q}_1 = (\imath, \mu/|\mu|)^T$ and a generalized eigenvector $\mathbf{q}_2 = (0, -c/\mu)^T$ satisfying $\mathbf{M}(\mu)\mathbf{q}_1 = 0$ and $\mathbf{M}(\mu)\mathbf{q}_2 = -(\lambda+3\tau)|\mu|\mathbf{q}_1$ with $c = (\lambda+3\tau)/(\lambda+\tau)$, from which we obtain the general solution to the equation (3.7) as follows:

$$\widehat{\mathbf{u}}(\mu,y) = ((a_\mu + b_\mu y)\mathbf{q}_1 + b_\mu \mathbf{q}_2)\,\mathrm{e}^{\imath\mu x + |\mu| y}, \tag{3.11}$$

where

$$a_\mu = -\widehat{u}(\mu,0)\imath \quad \text{and} \quad b_\mu = -c^{-1}(\mu\widehat{v_0}(\mu,0) + \imath|\mu|\widehat{u}(\mu,0)). \tag{3.12}$$

From this, we obtain the following simple but important lemma.

LEMMA 3.1. *A solution to* (3.7) *on the domain* Ω_2 *with a given boundary value* $\mathbf{u}_0(x)$ *on* Γ_2 *is given by the following:*

$$\mathbf{u}(x,y) = \frac{1}{2\pi}\int_0^{2\pi} \mathbf{G}(x-x',y)\mathbf{u}_0(x')dx', \tag{3.13}$$

where $\mathbf{G}$ *is defined, using* $c = (\lambda+3\tau)/(\lambda+\tau)$, *as*

$$\begin{aligned}\mathbf{G}(x-x',y) &= \sum_{\mu=-\infty}^{\infty} \mathbf{G}_\mu(x-x',y), \\ \mathbf{G}_\mu(x-x',y) &= \begin{pmatrix} 1+\frac{|\mu|}{c}y & -\frac{\mu}{c}\imath y \\ -\frac{\mu}{c}\imath y & 1-\frac{|\mu|}{c}y \end{pmatrix}\mathrm{e}^{|\mu|y}\,\mathrm{e}^{\imath\mu(x-x')}.\end{aligned}$$

This analytic expression for the solution $\mathbf{u}$ on the domain Ω_2 is used to derive the ABC operator.

3.3. The ABC operator for the two dimensional case. In this section, using Lemma 3.1, we construct the ABC operator. First, consider the expression of the solution $\mathbf{u}$ in the exterior domain Ω_2 given in (3.13). By taking the derivative of $\mathbf{u}$ with respect to y, one finds that

$$\partial_y(\mathbf{u}(x,y)) = \frac{1}{2\pi}\int_0^{2\pi} \partial_y(\mathbf{G}_\mu(x-x',y))\mathbf{u}_0(x')\,dx'. \tag{3.14}$$

Note that the normal component of the stress tensor $\mathbf{n}\cdot\mathbf{T}$ is given by

$$\mathbf{n}\cdot\mathbf{T} = \begin{pmatrix} \tau(\partial_y u + \partial_x v) \\ (\lambda+2\tau)\partial_y v + \lambda\partial_x u \end{pmatrix}, \tag{3.15}$$

and observe that it can be written in terms of $\mathbf{u}$ on the interface Γ_2 as follows:

$$\mathbf{n}\cdot\mathbf{T} = \sum_{\mu=-\infty}^{\infty} \frac{1}{2\pi}\int_0^{2\pi} \mathbf{A}_\mu \mathbf{u}_0(x')\,dx', \tag{3.16}$$

where

$$\mathbf{A}_\mu = \frac{2}{\lambda+3\tau}\begin{pmatrix} \tau(\lambda+2\tau)|\mu| & \tau^2\imath\mu \\ -\tau^2\imath\mu & \tau(\lambda+2\tau)|\mu| \end{pmatrix} \mathrm{e}^{\imath\mu(x-x')}. \tag{3.17}$$

Define the artificial boundary operator $\mathcal{A}$ by the following:

$$\mathcal{A}\mathbf{u}_0(x) = \sum_{\mu=-\infty}^{\infty} \frac{1}{2\pi}\int_0^{2\pi} \mathbf{A}_\mu \mathbf{u}_0(x')\,dx'. \tag{3.18}$$

It is interesting to note that the operator $\mathcal{A}$ is real and symmetric since $\mathbf{A}_\mu(x-x') = \mathbf{A}_\mu^H(x'-x)$.

3.4. The ABC operator for the three dimensional case. We now extend the previous analysis to the three dimensional case by constructing the solution of the homogeneous linear elasticity problem in a semi-infinite domain, Ω_2. Assume that Γ_2 is the plane $z=0$. As in the two dimensional case, assume that in the lateral direction, the solution is periodic with 2π periodicity for both variables, x and y. The following result is the analogue to Lemma 3.1.

LEMMA 3.2. *A solution to* (3.7) *with given boundary data* $\mathbf{u}_0(x,y)$ *on the interface* Γ_2 *is given by the following:*

$$\mathbf{u}(x,y,z) = \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi} \mathbf{G}(x-x',y-y',z)\mathbf{u}_0(x',y')\,dx'dy', \tag{3.19}$$

where $\mathbf{G}$ *is defined, using* $c=(\lambda+3\tau)/(\lambda+\tau)$ *and* $d=|(\mu,\nu)|$, *as*

$$\begin{aligned}&\mathbf{G}(x-x',y-y',z)\\ &= \sum_{\mu,\nu=-\infty}^{\infty}\begin{pmatrix} 1+\frac{\mu^2}{cd}z & \frac{\mu\nu}{cd}z & -\frac{\mu}{c}\imath z \\ \frac{\mu\nu}{cd}z & 1+\frac{\nu^2}{cd}z & -\frac{\nu}{c}\imath z \\ -\frac{\mu}{c}\imath z & -\frac{\nu}{c}\imath z & 1-\frac{d}{c}z \end{pmatrix}\mathrm{e}^{dz}\mathrm{e}^{\imath(\mu,\nu)\cdot(x-x',y-y')}.\end{aligned}$$

For the definition of the artificial boundary operator $\mathcal{A}$, note that the normal component of the stress tensor $\mathbf{T}$ is

$$\mathbf{n}\cdot\mathbf{T}=\begin{pmatrix}\mu(\partial_z u+\partial_x w)\\ \mu(\partial_z v+\partial_y w)\\ (\lambda+2\tau)\partial_z w+\lambda(\partial_x u+\partial_y v)\end{pmatrix}, \tag{3.20}$$

where $\mathbf{n}$ is the outer unit normal vector to the interface Γ_2. It is easy to see that

$$\mathbf{n}\cdot\mathbf{T}=\sum_{\mu,\nu=-\infty}^{\infty}\frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi}\mathbf{A}_{\mu,\nu}\mathbf{u}_0(x',y')\,dx'dy', \tag{3.21}$$

where

$$\mathbf{A}_{\mu,\nu}=\begin{pmatrix}\tau\left(\frac{\mu^2}{cd}+d\right) & \tau\frac{\mu\nu}{cd} & \frac{2\tau^2}{\lambda+3\tau}\imath\mu\\ \tau\frac{\mu\nu}{cd} & \tau\left(\frac{\nu^2}{cd}+d\right) & \frac{2\tau^2}{\lambda+3\tau}\imath\nu\\ -\frac{2\tau^2}{\lambda+3\tau}\imath\mu & -\frac{2\tau^2}{\lambda+3\tau}\imath\nu & (\lambda+2\tau)\left(-\frac{d}{c}+d\right)\end{pmatrix}\mathrm{e}^{\imath(\mu,\nu)\cdot(x-x',y-y')}.$$

Define the artificial boundary operator $\mathcal{A}$ as follows:

$$\mathcal{A}\mathbf{u}_0(x,y)=\sum_{\mu,\nu=-\infty}^{\infty}\frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi}\mathbf{A}_{\mu,\nu}\mathbf{u}_0(x',y')\,dx'dy'. \tag{3.22}$$

Similarly to the two dimensional case, the operator $\mathcal{A}$ is symmetric.

3.5. The total energy and force balance equation. In this section, we find an alternative total energy formula for (3.2) and also a force balance equation for (3.5) that involve only the domain Ω_1 and Γ_2, using the ABC operator constructed in the previous sections. For convenience, denote $\mathbf{u}_0$ to be the displacement field of $\mathbf{u}$ at Γ_2.

Write the total elastic energy in Ω in terms of the total energy $\mathcal{E}_1$ in Ω_1 and the total energy $\mathcal{E}_2$ in Ω_2 as follows:

$$\begin{aligned}\mathcal{E}_{total}&=\frac{1}{2}\int_\Omega\mathbf{S}:\mathbf{T}\,d\mathbf{x}-\int_\Omega\mathbf{u}\cdot\mathbf{f}\chi\,d\mathbf{x}\\ &=\frac{1}{2}\left\{\int_{\Omega_1}\mathbf{S}:\mathbf{T}\,d\mathbf{x}-\int_{\Omega_1}\mathbf{u}\cdot\mathbf{f}\,d\mathbf{x}\right\}+\frac{1}{2}\int_{\Omega_2}\mathbf{S}:\mathbf{T}\,d\mathbf{x}\\ &=\mathcal{E}_1+\mathcal{E}_2.\end{aligned}$$

Let $\mathcal{L}$ denote the linear elasticity operator:

$$\mathcal{L}\mathbf{u}=\tau\Delta\mathbf{u}+(\lambda+\tau)\nabla(\nabla\cdot\mathbf{u}).$$

Note that $\mathcal{E}_2$ can be written in terms of the boundary data $\mathbf{u}_0(\mathbf{x})$ on the interface Γ_2 as follows:

$$\begin{aligned}\mathcal{E}_2&=\frac{1}{2}\int_{\Omega_2}\mathbf{S}:\mathbf{T}\,d\mathbf{x}\\ &=-\frac{1}{2}\int_{\Omega_2}\mathbf{u}\cdot\mathcal{L}\mathbf{u}\,d\mathbf{x}+\frac{1}{2}\int_{\Gamma_2}\mathbf{u}_0\cdot(\mathbf{n}\cdot\mathbf{T})\,d\Gamma\\ &=\frac{1}{2}\int_{\Gamma_2}\mathbf{u}_0\cdot\mathcal{A}\mathbf{u}_0\,d\Gamma,\end{aligned}$$

where we use the fact that $\mathcal{L}\mathbf{u} = 0$ in the domain Ω_2 and the definition (3.22) of the artificial boundary operator $\mathcal{A}$.

Consequently, the total energy $\mathcal{E}_{total}$ in the domain Ω is

$$\begin{aligned}\mathcal{E}_{total} &= \mathcal{E}_1 + \mathcal{E}_2 \\ &= \frac{1}{2}\int_{\Omega_1} \mathbf{S} : \mathbf{T}\, d\mathbf{x} - \int_{\Omega_1} \mathbf{u}\cdot\mathbf{f} + \frac{1}{2}\int_{\Gamma_2} \mathbf{u}_0 \cdot \mathcal{A}\mathbf{u}_0\, d\Gamma.\end{aligned} \tag{3.23}$$

This is the new formula for the total energy (3.2) that involves only the domain Ω_1 and Γ_2. Now apply integration by parts to the first term in (3.23) and obtain

$$\begin{aligned}\mathcal{E}_{total} = &-\frac{1}{2}\int_{\Omega_1} \mathbf{u}\cdot\mathcal{L}\mathbf{u}\, d\mathbf{x} - \int_{\Omega_1} \mathbf{u}\cdot\mathbf{f}\, d\mathbf{x} \\ &+\frac{1}{2}\int_{\Gamma_2} \mathbf{u}_0\cdot\mathcal{A}\mathbf{u}_0 - \mathbf{u}_0\cdot(\mathbf{n}\cdot\mathbf{T})\, d\Gamma.\end{aligned} \tag{3.24}$$

Application of the variational principle for the new expression of the total energy (3.24) results in the following force balance equations, which use the ABC operator $\mathcal{A}$ in the ABC on Γ_2:

$$\begin{aligned}-\mathcal{L}\mathbf{u} &= \mathbf{f} \qquad \text{in } \Omega_1, \\ \mathbf{n}\cdot\mathbf{T} &= \mathcal{A}\mathbf{u}_0 \quad \text{on } \Gamma_2.\end{aligned}$$

4. The ABCs for discrete elasticity. In this section, we study the analogue of the ABCs for discrete elasticity. In particular, we discuss the solvability (well-posedness) of the discrete strain model in an unbounded or semi-infinite domain.

It is not trivial to show directly the well-posedness of the discrete strain model in an infinite domain. As discussed in Hagstrom and Keller [11], the well-posedness can be derived from a so-called solvability requirement, which is a solvability condition for the exterior domain problem for which the force term is zero. Generally, the validation of this solvability requirement is done by introducing a sum-of-exponentials ansatz for the solution below the artificial boundary. It is difficult, however, to validate this condition fully in an analytic manner [11] except for simple problems such as the Laplace equation. Numerical validation is partially used, since an analytic validation could not be made fully for the current problem of interest.

The importance of the framework developed in this section is that it identifies how the solvability requirement can be used to show well-posedness of the discrete equations posed on the unbounded domain, and also clarifies why an appropriate use of the ABC operator leads to the exact boundary condition. To the best of our knowledge, it is the first attempt to formulate a general discussion on the solvability of discrete systems in an infinite domain in terms of solvability requirements. Furthermore, this formal discussion leads to an understanding of the ABC operator as a Schur complement operator and reveals various properties of the resulting reduced system on the finite domain. These properties of the reduced system are important when one attempts to develop an appropriate solver for the reduced system (see the concluding remark in section 6).

Throughout this section, we assume that the lattice of the discrete strain model is connected [19]. We begin this section by briefly reviewing the discrete elastic model introduced in [21].

4.1. Discrete elasticity. To describe the strain energy at each atom, $\mathbf{i} = (i, j, k)$, introduce the translation operators, $T_k^{\pm}$, and the discrete difference operators, $D_k^{\pm}$, D_k^0, defined as follows:

$$\begin{aligned} T_k^{\pm} f(\mathbf{i}) &= f(\mathbf{i} \pm \mathbf{e_k}), \\ D_k^{+} f(\mathbf{i}) &= \frac{(T_k^{+} - 1) f(\mathbf{i})}{h}, \\ D_k^{-} f(\mathbf{i}) &= \frac{(1 - T_k^{-}) f(\mathbf{i})}{h}, \\ D_k^{0} f(\mathbf{i}) &= \frac{(T_k^{+} - T_k^{-}) f(\mathbf{i})}{2h}, \end{aligned}$$

where h is the lattice constant and $\mathbf{e}_k$ is the vector in the kth direction for $k = 1, 2, 3$ with $\|\mathbf{e}_k\| = h$. Throughout this paper, we assume the lattice constant $h = 1$ for simplicity. We use i for the depth-like index, with $-\infty < i \leq n$. Here n is the maximum height of the material. An ABC is sought at $i = 0$, assuming that there is no force for $i < 0$.

Let $\mathbf{u}(\mathbf{i}) = (u_k(\mathbf{i}))_{k=1,\dots,d}$ be the displacement at the discrete point $\mathbf{i}$ relative to an equilibrium lattice. The discrete strain components defined below ((4.1) and (4.2)) can be used to describe the discrete elastic energy. For $k, \ell = 1, 2, 3$ and $p, q = \pm$,

$$S_{k\ell}^{\pm}(\mathbf{u}(\mathbf{i})) = D_{\ell}^{\pm} u_k(\mathbf{i}), \tag{4.1}$$

$$S_{k\ell}^{pq}(\mathbf{u}(\mathbf{i})) = \frac{1}{2}(D_{\ell}^{q} u_k(\mathbf{i}) + D_k^{p} u_{\ell}(\mathbf{i})). \tag{4.2}$$

The discrete energy density at a point $\mathbf{i}$ is then given by

$$E(\mathbf{i})(\mathbf{u}, \mathbf{u}) = \sum_{k,p} \alpha_k^p (S_{kk}^p(\mathbf{u}))^2 + \sum_{k \neq \ell, p, q} \left\{ 2\beta_{k\ell}^{pq} (S_{k\ell}^{pq}(\mathbf{u}))^2 + \gamma_{k\ell}^{pq} S_{kk}^p(\mathbf{u}) S_{\ell\ell}^q(\mathbf{u}) \right\}.$$

The subsequent discussion uses three constant displacement fields, denoted by $\mathbf{1}_k$ for $k = 1, 2, 3$, for a constant displacement in the kth component. For convenience, denote $\mathbf{1}$ for any constant vector. With some abuse of notation, it is used to denote a constant vector formed by taking the linear combinations of $\mathbf{1}_k$ and $\mathbf{1}_\ell$ with $k \neq \ell$.

The elastic constants should be chosen to ensure positivity of the (total) energy density, as discussed, for example, in [17]. A sufficient condition for the positivity is

$$\min_{k,p} \alpha_k^p \geq \max_{pq} \gamma^{pq} + c \tag{4.3}$$

for some positive constant $c > 0$. One consequence of positivity is that rigid body motions are the only local displacements that entail no internal energy.

A discrete version of the elastic energy density E at a lattice point $\mathbf{i} = (i, j, k)$ is then given as follows:

$$\mathcal{E}_{total} = \mathcal{E}_{total}(\mathsf{U}, \mathsf{U}) = \widetilde{\mathcal{E}}(\mathsf{U}, \mathsf{U}) - (\mathsf{F}, \mathsf{U}), \tag{4.4}$$

where

$$\widetilde{\mathcal{E}}(\mathsf{U}, \mathsf{U}) = \sum_{\mathbf{i}} E(\mathbf{i})(\mathbf{u}, \mathbf{u}), \tag{4.5}$$

$$\begin{aligned} \mathsf{U} &= (U_n, \dots, U_1, U_0, U_{-1}, \dots)^T, \\ \mathsf{F} &= (F_n, \dots, F_1, F_0, F_{-1}, \dots)^T, \end{aligned} \tag{4.6}$$

where U_i and F_i are the vectors of size N consisting of displacement components $\mathbf{u}$ and force components $\mathbf{f}$ at depth i. The total energy formula (4.4) is modified in section 5.3 to include effects of lattice mismatch. Under traction-free (i.e., Neumann) boundary conditions on the surface Γ_1, the external force vector F must be orthogonal to any constant vector field. As shown in (5.6) in section 5.3, this is also true for the effective force due to lattice mismatch in a thin film. Now, due to the boundary condition, the periodic condition in the lateral direction, and Neumann condition on the surface Γ_1, and from the assumption that the lattice is connected, it follows that

$$\widetilde{\mathcal{E}}(\mathsf{U},\mathsf{U}) = 0 \quad \Longleftrightarrow \quad \mathsf{U} = \mathbf{1}; \tag{4.7}$$

see also Martinsson and Babuska [19] for further discussion on connectivity.

As described in detail in section 4.5, the total energy $\mathcal{E}_{total}$ has the following alternative form:

$$\mathcal{E}_{total} = \mathcal{E}_{total}(\mathsf{U},\mathsf{U}) = \frac{1}{2}(\mathbf{H}\mathsf{U},\mathsf{U}) - (\mathsf{F},\mathsf{U}), \tag{4.8}$$

where

$$\mathbf{H} = \begin{pmatrix} \cdots & \cdots & 0 & 0 & 0 & 0 & \cdots \\ \cdots & A_{i+1i+1} & A_{i+1i} & 0 & \ddots & 0 & \vdots \\ \vdots & A_{ii+1} & A_{ii} & A_{ii-1} & 0 & \ddots & \vdots \\ \vdots & 0 & A_{i-1i} & A_{i-1i-1} & A_{i-1i-2} & 0 & \vdots \\ \vdots & \cdots & 0 & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \ddots & \ddots & \vdots \end{pmatrix}. \tag{4.9}$$

The discrete strain equations are derived from the following optimization problem:

$$\min \mathcal{E}_{total} = \min \left(\frac{1}{2}(\mathbf{H}\mathsf{U},\mathsf{U}) - (\mathsf{F},\mathsf{U}) \right). \tag{4.10}$$

Note that the off-diagonal block matrices satisfy $A_{i+1i} = A_{ii+1}^T$ for all $i \le n$. Furthermore, since the material is homogeneous below the artificial boundary, $A_{ii+1} = A_{-10}$ and $A_{ii} = A_{00}$ are independent of i for all $i < 0$. Both A_{00} and A_{-10} are invertible. In particular, the proof that A_{-10} is invertible is included in the appendix.

Denote

$$\mathsf{U} = \begin{pmatrix} \mathsf{U}^+ \\ U_0 \\ \mathsf{U}^- \end{pmatrix} \quad \text{and} \quad \mathsf{F} = \begin{pmatrix} \mathsf{F}^+ \\ F_0 \\ 0 \end{pmatrix}, \tag{4.11}$$

in which U^- and U^+ are vectors consisting of all U_i for $i < 0$ and $i > 0$, respectively. The vector F^+ of forces is defined similarly. Correspondingly, write $\mathbf{H}$ as follows:

$$\mathbf{H} = \begin{pmatrix} \mathsf{A}_{II} & \mathsf{A}_{I0}^T & 0 \\ \mathsf{A}_{I0} & A_{00} & \mathsf{B}^T \\ 0 & \mathsf{B} & \mathsf{M} \end{pmatrix}, \tag{4.12}$$

where A_{II} acts on U^+, A_{00} acts on U_0, and M acts on U^-.

An analysis in section 4.2 shows that under an appropriate solvability condition, the optimization problem (4.10) leads to the force balance equation

$$\mathbf{H}\mathsf{U} = \mathsf{F}. \tag{4.13}$$

Moreover, the analysis shows that (4.13) and the optimization problem (4.10) are well-posed.

Since the displacement $\mathbf{u}$ decays as $i \to -\infty$, one might expect that the space ℓ^2 would be the appropriate admissible solution space for the optimization problem (4.10). Coercivity of the operator $\mathbf{H}$ fails, however, for the space ℓ^2, so that it is difficult to show the solvability of the problem (4.10) directly. The solvability requirement of the next section remedies this lack of coercivity.

4.2. The solvability requirement and the general form of the ABC operator. In the region $i < 0$, i.e., below the artificial boundary, the solution of the problem (4.10) satisfies

$$\begin{aligned} A_{-10}U_0 + A_{00}U_{-1} + A_{-10}^T U_{-2} &= 0, \\ A_{-10}U_{-1} + A_{00}U_{-2} + A_{-10}^T U_{-3} &= 0, \\ A_{-10}U_{-2} + A_{00}U_{-3} + A_{-10}^T U_{-4} &= 0, \\ &\vdots\ . \end{aligned} \tag{4.14}$$

The solvability condition is phrased in terms of solutions for (4.14) that are decaying or constant.

Condition 4.1. *There exists an invertible matrix* C *such that for any* $U_0 \in \mathbb{R}^{\mathrm{N}}$, *the vector* $(U_0, U_{-1}, U_{-2}, \dots)$ *with*

$$U_i = \mathsf{C}^i U_0 \quad \forall i \le 0 \tag{4.15}$$

(where C^0 *is the identity matrix) satisfies* (4.14). *In addition,*

$$\mathsf{C}^i U_0 = U_0 \quad \forall i \le 0, \quad \forall\, U_0 \in \operatorname{span}\{\mathbf{1}_k : k = 1, 2, 3\} \quad \textit{and} \tag{4.16}$$

$$\mathsf{C}^i U_0 \to 0 \quad \textit{as} \quad i \to -\infty \quad \forall\, U_0 \in \operatorname{span}\{\mathbf{1}_k : k = 1, 2, 3\}^{\perp}. \tag{4.17}$$

Note that the constant displacement field is a trivial solution to (4.14) since it is the discretization of the differential operator $\mathcal{L}$, which is reflected in the statement (4.16). The second statement (4.17) says that if U_0 is orthogonal to all constant fields, then the solution decays to 0 at infinity.

Condition 4.1, which is validated in section 4.3, has a number of important consequences, as described in the following subsections.

4.2.1. On the general ABC operator $\mathcal{A}$. The general form of the ABC operator, under Condition 4.1, is described in this subsection.

Define the following two special vector spaces:

$$\Theta = \left\{ \mathsf{V} = (V_{-1}, V_{-2}, \dots) : \inf_{\xi \in \mathbb{R}} \|\mathsf{V} + \xi \mathbf{1}\|_{\ell^2} < \infty \quad \Psi(V_i) = 0 \quad \forall i < -1 \right\},$$

where

$$\Psi(V_i) = A_{-10}V_{i+1} + A_{00}V_i + A_{-10}^T V_{i-1}, \tag{4.18}$$

and

$$\Theta^* = \{\mathsf{G} = (G_{-1}, 0, \dots, 0, \dots) : G_{-1} \in \mathbb{R}^{\mathrm{N}}\}. \tag{4.19}$$

It is clear that both spaces Θ and Θ^* are finite dimensional. In particular, due to the constraints (4.18), the space Θ is completely determined by the first two vectors V_{-1} and V_{-2}. Due to Condition 4.1, the dimension of the space Θ is at least N; in fact, as shown below, its dimension is exactly N. By defining $\|\mathsf{V}\|_\Theta = \sum_{k=-1,-2} \|V_k\|_{\ell^2}$ as a norm on Θ, the space Θ is a Banach space, as is Θ^*. The following lemma is simple but important for the subsequent discussion (the proof can be found in the appendix).

LEMMA 4.1. *Under Condition* 4.1*, the matrix* M*, given as*

$$\mathsf{M} = \begin{pmatrix} A_{00} & A_{-10}^T & 0 & 0 & 0 & \cdots \\ A_{-10} & A_{00} & A_{-10}^T & 0 & 0 & \vdots \\ 0 & A_{-10} & A_{00} & A_{-10}^T & 0 & \vdots \\ \vdots & 0 & A_{-10} & A_{00} & \cdots & \vdots \\ \vdots & 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \cdots & \cdots & \ddots & \ddots & \vdots \end{pmatrix}, \tag{4.20}$$

is an isomorphic mapping from Θ *to* Θ^**.*

Since M is isomorphic, the following equation is solvable:

$$\mathsf{M}\mathsf{U}^- = \mathsf{G}, \tag{4.21}$$

where

$$\mathsf{U}^- = (U_{-1}, U_{-2}, U_{-3}, \dots)^T$$
$$\text{and} \quad \mathsf{G} = (-A_{-10}U_0, 0, 0, \dots)^T.$$

In particular, $\mathsf{U}^- = \mathsf{M}^{-1}\mathsf{G}$. Multiplying both sides of this equation by $\mathsf{B} = (A_{-10}^T, 0, \dots, 0)$ yields the relation

$$A_{-10}^T U_{-1} = -\mathsf{B}\mathsf{M}^{-1}\mathsf{B}^T U_0. \tag{4.22}$$

The general form of the ABC operator $\mathcal{A}$ is defined by

$$\mathcal{A} = \mathsf{B}\mathsf{M}^{-1}\mathsf{B}^T. \tag{4.23}$$

Note that the operator $\mathcal{A}$ relates U_{i-1} and U_i for $i \le 0$. Since U^- belongs to the space Θ, U_i should decay as $i \to -\infty$, unless U_0 has a nonzero component that is a constant vector.

4.2.2. The total energy formula for the system above the artificial boundary. This section introduces the new energy formula that is a by-product of the ABC operator.

Since $A_{-10}U_{i+1} + A_{00}U_i + A_{-10}^T U_{i-1} = 0$ and $F_i = 0$ for $i < 0$, the total energy

$\mathcal{E}_{total}$ from (4.8) can be written as follows:

$$\begin{aligned}\mathcal{E}_{total} &= \sum_{i\geq 0} \frac{1}{2}\left(U_i, (A_{ii+1}U_{i+1} + A_{ii}U_i + A_{ii-1}U_{i-1})\right) - (U_i, F_i)\\ &= \frac{1}{2}\left(U_0, \left(A_{01}U_1 + A_{00}U_0 + A_{-10}^T U_{-1}\right)\right) - (U_0, F_0)\\ &\quad + \sum_{i>0} \frac{1}{2}\left(U_i, (A_{ii+1}U_{i+1} + A_{ii}U_i + A_{ii-1}U_{i-1})\right) - (U_i, F_i)\,.\end{aligned}$$

This formula, however, depends on the displacement field U_{-1} below the artificial boundary. To remove this dependence and obtain an energy formula (and a reduced force balance equation) that involves displacement fields only above the artificial boundary, use the operator $\mathcal{A}$ to obtain the following alternative formula:

$$\begin{aligned}\mathcal{E}_{total} &= \frac{1}{2}\left(U_0, (A_{01}U_1 + (A_{00} - \mathcal{A})U_0)\right) - (U_0, F_0)\\ &\quad + \sum_{i>0} \frac{1}{2}\left(U_i, (A_{ii-1}U_{i-1} + A_{ii}U_i + A_{ii+1}U_{i+1})\right) - (U_i, F_i)\,.\end{aligned} \tag{4.24}$$

Note that the energy formula given in (4.24) depends only on the displacement fields U_0 and U^+ above the artificial boundary, but it includes the energy in the strain field below the artificial boundary. In addition, optimization of this formula for the energy yields the reduced equation on the upper domain with the ABC using the operator $\mathcal{A}$, as shown in the next subsection.

4.2.3. The force balance equation. Define the following admissible solution space for the optimization problem (4.10):

$$\mathbf{V} = \left\{\mathsf{V} = (V_n, \dots, V_0, V_{-1}, \dots) : \inf_{\xi\in\mathbb{R}} \|\mathsf{V} + \xi\mathbf{1}\|_{\ell^2} < \infty, \quad \Psi(V_i) = 0 \quad \forall i < 0\right\}.$$

Thanks to Condition 4.1, the force balance equation that results from minimizing the total energy in its reduced form (4.24) is

$$\widehat{\mathbf{H}}\begin{pmatrix}\mathsf{U}^+\\ U_0\end{pmatrix} = \begin{pmatrix}\mathsf{A}_{II} & \mathsf{A}_{I0}^T\\ \mathsf{A}_{I0} & A_{00} - \mathcal{A}\end{pmatrix}\begin{pmatrix}\mathsf{U}^+\\ U_0\end{pmatrix} = \begin{pmatrix}\mathsf{F}^+\\ F_0\end{pmatrix}. \tag{4.25}$$

The reduced form (4.25) of the force balance equation, as well as its properties, is the main result of this work. Note that (4.25) involves the *Schur complement* of the matrix A_{00} in the original force balance equation (4.13).

The properties of the matrices $\mathcal{A}$ and $\widehat{\mathbf{H}}$ are summarized in the following lemma, whose proof is provided in the appendix.

LEMMA 4.2. *The matrix $\mathcal{A}$ is symmetric and positive definite, the matrix $\widehat{\mathbf{H}}$ is symmetric and nonnegative definite, and the null space of $\widehat{\mathbf{H}}$ consists of the constant displacement fields* $\operatorname{span}\{\mathbf{1}_k : k = 1, 2, 3\}$.

The analysis in this section is performed for the Neumann boundary condition at the top boundary Γ_1, by which we mean that the variational principle (4.10) involves no constraint on the solution at Γ_1. In this case, it is most important to note that (4.25) is solvable since (F^+, F_0) belongs to the range of $\widehat{\mathbf{H}}$; namely, (F^+, F_0) is orthogonal to the constant vector fields, which is exactly the null space of $\widehat{\mathbf{H}}$ as noted in Lemma 4.2. In addition, the solution to (4.25) is determined up to a constant

vector. However, the additional contribution of the constant vector does not affect the total energy evaluation since the total energy is invariant with respect to the constant displacement. Furthermore, use of the Neumann condition is only to simplify the analysis. It does not affect the ABC operator $\mathcal{A}$, which can be used for any choice of boundary conditions on the top.

In passing to the next section, we summarize the most important properties of the ABC operator $\mathcal{A}$, which guide its construction.

P1 The operator $\mathcal{A}$ is a symmetric and positive definite matrix mapping $\mathbb{R}^N$ to $\mathbb{R}^N$.

P2 The relation between U_{-1} and U_0 is that $U_{-1} = -(A_{-10}^T)^{-1}\mathcal{A}U_0 = \mathsf{C}U_0$.

4.3. Validation of the solvability requirement, Condition 4.1. In this section, Condition 4.1 is derived by introducing a sum-of-exponentials ansatz. Much of the derivation, including the most crucial steps, is analytic, but some steps are based on numerical evidence. In related work on the Laplace equation, Hagstrom and Keller [11] performed a completely analytic validation of the analogue of Condition 4.1.

The following presentation is mostly based on the thesis of Lee [18] and is similar to the work by Russo and Smereka [20], which used the palindromic eigenvalue problem [15, 16]. Although these works did not state a general solvability condition like Condition 4.1, their analysis is equivalent to a validation of this condition. Throughout this section, denote $\mathcal{F}$ and $\mathcal{F}^{-1}$ to be the discrete forward and backward Fourier transforms, respectively.

4.3.1. Two dimensional case. The force balance equations at a point $(x_m, y_i) = (m, i)$ are

$$\begin{aligned} -(\mathcal{L}\mathbf{u})_1 &= -C_{11}D_x^+D_x^-u - C_{44}D_y^+D_y^-u - (C_{12}+C_{44})D_y^0D_x^0v = 0, \\ -(\mathcal{L}\mathbf{u})_2 &= -C_{44}D_x^+D_x^-v - C_{11}D_y^+D_y^-v - (C_{12}+C_{44})D_x^0D_y^0u = 0. \end{aligned} \tag{4.26}$$

Since the solution is periodic in the x-direction, we introduce the following ansatz:

$$\begin{aligned} \mathbf{u}(m,i) &= \frac{1}{N_x}\sum_{\mu=0}^{N_x-1}\widehat{\mathbf{u}}(\mu,i)\,\mathrm{e}^{2\pi\imath\mu m/N_x} \\ &= \frac{1}{N_x}\sum_{\mu=0}^{N_x-1}\widehat{\mathbf{u}}(\mu)\gamma^i\,\mathrm{e}^{2\pi\imath\mu m/N_x}, \end{aligned} \tag{4.27}$$

where N_x is such that $\mathbf{u}(m,i) = \mathbf{u}(N_x+m,i)$ for all m.

From (4.27), the force balance equations (4.26) become

$$P(\mu,\gamma)\widehat{\mathbf{u}}(\mu,i) = \left(\gamma^2\widehat{A}_{-10}(\mu) + \gamma\widehat{A}_{00}(\mu) + \widehat{A}_{-10}^H(\mu)\right)\widehat{\mathbf{u}}(\mu,i) = 0 \tag{4.28}$$

for $\mu = 0, 1, \ldots, N_x - 1$, where

$$\widehat{A}_{-10}(\mu) = \begin{pmatrix} -C_{44} & -\imath\frac{C_{12}+C_{44}}{2}\sin(2\pi\mu/N_x) \\ -\imath\frac{C_{12}+C_{44}}{2}\sin(2\pi\mu/N_x) & -C_{11} \end{pmatrix},$$

$$\widehat{A}_{00}(\mu) = \begin{pmatrix} 2C_{44}+2C_{11}(1-\cos(2\pi\mu/N_x)) & 0 \\ 0 & 2C_{11}+2C_{44}(1-\cos(2\pi\mu/N_x)) \end{pmatrix},$$

and $\widehat{A}_{-10}^H$ is the complex transpose of the matrix $\widehat{A}_{-10}$.

Nontrivial solutions for this system require that

$$\det P(\mu,\gamma)=0. \tag{4.29}$$

This is the well-known palindromic eigenvalue problem [15, 16, 20]. Note that for $\mu=0$, which corresponds to the constant vector in the Fourier expansion of the solution ansatz (4.27), the only solution to (4.29) is $\gamma=1$, which corresponds to the constant solution to (4.14).

For $\mu\neq 0$, (4.29) has four solutions that occur in pairs $(\gamma_k,\overline{\gamma}_k{}^{-1})$ for $k=1,2$, since

$$\det(P(\mu,\gamma))=0 \quad\Longleftrightarrow\quad \overline{\det(P(\mu,\gamma))}=0 \tag{4.30}$$

and

$$\overline{P(\mu,\gamma)}=\overline{\gamma}^2 P(\mu,\overline{\gamma}^{\,-1}).$$

We then pick a pair of solutions (γ_1,γ_2) with $|\gamma_k|>1$ for $k=1,2$, which are the relevant choices since the corresponding solution is decaying as $i\to-\infty$ for $\mu\neq 0$, and we also pick two linearly independent eigenvectors $\mathbf{q}_1(\mu)$ and $\mathbf{q}_2(\mu)$ that correspond to γ_1 and γ_2, respectively [10, 20]; i.e.,

$$P(\mu,\gamma_1)\mathbf{q}_1(\mu)=P(\mu,\gamma_2)\mathbf{q}_2(\mu)=0. \tag{4.31}$$

It is possible that $|\gamma_k|=1$ or that $\gamma_1=\gamma_2$ and there is a generalized eigenvector, but these possibilities have not been seen numerically. Indeed, the occurrence of a generalized eigenvector in the continuous case (cf. section 3.2) does not seem to have consequences for the discrete case.

We then arrive at the general solution for $\widehat{\mathbf{u}}(\mu,i)$ given as follows:

$$\widehat{\mathbf{u}}(\mu,i)=\mathbf{q}_1(\mu)\gamma_1^i+\mathbf{q}_2(\mu)\gamma_2^i. \tag{4.32}$$

For the zero mode $\mu=0$, two linearly independent vectors $\mathbf{q}_k(0)$ are $\mathbf{q}_1=(1,0)^T$ and $\mathbf{q}_2=(0,1)^T$. Note that omitting this mode would make 0 an eigenvalue for the operator $\mathcal{A}$, but that $\mathcal{A}$ should be positive definite as indicated in property **P1** in subsection 4.2.3.

4.3.2. Three dimensional case. As in the two dimensional case, consider the force balance equations at a point $(x_m,y_n,z_i)=(m,n,i)$:

$$\begin{aligned}
-(\mathcal{L}\mathbf{u})_1 &= -C_{11}D_x^+D_x^-u-C_{44}(D_y^+D_y^-u+D_z^+D_z^-u)\\
&\quad -(C_{12}+C_{44})(D_y^0D_x^0v+D_z^0D_x^0w)\\
&= 0,\\
-(\mathcal{L}\mathbf{u})_2 &= -C_{11}D_y^+D_y^-v-C_{44}(D_x^+D_x^-v+D_z^+D_z^-v)\\
&\quad -(C_{12}+C_{44})(D_y^0D_x^0u+D_z^0D_y^0w)\\
&= 0,\\
-(\mathcal{L}\mathbf{u})_3 &= -C_{11}D_z^+D_z^-w-C_{44}(D_x^+D_x^-w+D_y^+D_y^-w)\\
&\quad -(C_{12}+C_{44})(D_z^0D_x^0u+D_z^0D_y^0v)\\
&= 0.
\end{aligned} \tag{4.33}$$

Introduce the solution ansatz as follows:

$$\mathbf{u}(m,n,i) = \frac{1}{N_x N_y}\sum_{\mu=0}^{N_x-1}\sum_{\nu=0}^{N_y-1}\widehat{\mathbf{u}}(\mu,\nu,i)\,\mathrm{e}^{(2\pi\imath\mu m)/N_x+(2\pi\imath\nu n)/N_y} = \frac{1}{N_x N_y}\sum_{\mu=0}^{N_x-1}\sum_{\nu=0}^{N_y-1}\widehat{\mathbf{u}}(\mu,\nu)\gamma^i\,\mathrm{e}^{(2\pi\imath\mu m)/N_x+(2\pi\imath\nu n)/N_y}, \tag{4.34}$$

where N_x and N_y are the periods in x and y for $\mathbf{u}$. From ansatz (4.34), the force balance equations become

$$P(\mu,\nu,\gamma)\widehat{\mathbf{u}}(\mu,\nu,i) = \left(\gamma^2\widehat{A}_{-10}(\mu,\nu)+\gamma\widehat{A}_{00}(\mu,\nu)+\widehat{A}^H_{-10}(\mu,\nu)\right)\widehat{\mathbf{u}}(\mu,\nu,i) = 0 \tag{4.35}$$

for each $\mu = 0,1,\ldots,N_x$ and $\nu = 0,1,\ldots,N_y$, where $\widehat{A}_{-10} = \widehat{A}_{-10}(\mu,\nu)$ and $\widehat{A}_{00} = \widehat{A}_{00}(\mu,\nu)$ are given by

$$\widehat{A}_{-10} = \begin{pmatrix} -C_{44} & 0 & -s_1 \\ 0 & -C_{44} & -s_2 \\ -s_1 & -s_2 & -C_{11} \end{pmatrix},$$

$$\widehat{A}_{00} = \begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix},$$

in which

$$s_1 = \imath\frac{C_{12}+C_{44}}{2}\sin(2\pi\mu/N_x),$$
$$s_2 = \imath\frac{C_{12}+C_{44}}{2}\sin(2\pi\nu/N_y)$$

and

$$\begin{aligned} a_{11} &= 2C_{11}(1-\cos(2\pi\mu/N_x)) + 2C_{44}(1-\cos(2\pi\nu/N_y)) + 2C_{44},\\ a_{12} &= -(C_{12}+C_{44})\sin(2\pi\mu/N_x)\sin(2\pi\nu/N_y),\\ a_{21} &= a_{12},\\ a_{22} &= 2C_{44}(1-\cos(2\pi\mu/N_x)) + 2C_{11}(1-\cos(2\pi\nu/N_y)) + 2C_{44},\\ a_{33} &= 2C_{44}(1-\cos(2\pi\mu/N_x)) + 2C_{44}(1-\cos(2\pi\nu/N_y)) + 2C_{11}. \end{aligned}$$

A nontrivial solution can be found only if

$$\det P(\mu,\nu,\gamma) = 0. \tag{4.36}$$

As in the two dimensional case, for $\mu = \nu = 0$, the only solution is $\gamma = 1$, and for $(\mu,\nu) \neq (0,0)$, there are three pairs of eigenvalues, namely $(\gamma_k, \overline{\gamma}_k{}^{-1})$ with $|\gamma_k| > 1$ for $k = 1,2,3$, and corresponding eigenvectors $\mathbf{q}_k(\mu,\nu)$ that are mutually linearly independent, from which the general solution can be given as follows:

$$\widehat{\mathbf{u}}(\mu,\nu,i) = \mathbf{q}_1(\mu,\nu)\gamma_1^i + \mathbf{q}_2(\mu,\nu)\gamma_2^i + \mathbf{q}_3(\mu,\nu)\gamma_3^i. \tag{4.37}$$

Note that if the three values γ_k are distinct, then it can be seen directly that there exist three linearly independent eigenvectors $\mathbf{q}_k(\mu,\nu)$ corresponding to the three eigenvalues γ_k (see the appendix). Often in our computation, as seen in the work by Russo and Smereka [20], it happens that $\gamma_k = \gamma_\ell$ with $k \neq \ell$. When this happens, it is difficult to establish analytically the existence of linearly independent eigenvectors; this is always found to be the case, however, in the numerical computations.

4.4. On the discrete ABC operator $\mathcal{A}$ and Condition 4.1. In this section, the ABC operator $\mathcal{A}$ is constructed for the three dimensional case only, since the two dimensional construction is similar but simpler. We first construct the operator C that relates U_{i-1} and U_i by $U_{i-1} = \mathsf{C}U_i$, as indicated in **P2**. We then construct $\mathcal{A} = -(A_{-10}^T)\mathsf{C}$. Finally, we discuss the validation of Condition 4.1.

Note that the Fourier transforms $\widehat{A}_{-10}$ and $\widehat{A}_{00}$ of A_{-10} and A_{00} consist of 3×3 block matrices $\widehat{A}_{-10}(\mu,\nu)$ and $\widehat{A}_{00}(\mu,\nu)$. Since the vectors $\mathbf{q}_i(\mu,\nu)$ from (4.37) are mutually independent, define the following mutually orthonormal vectors:

$$\widetilde{\mathbf{q}}_i = c_i(\mathbf{q}_{i'} \times \mathbf{q}_{i''}),$$

in which each triple (i,i',i'') is a rearrangement of $(1,2,3)$ and the constants c_i's are chosen so that

$$\widetilde{\mathbf{q}}_i \cdot \mathbf{q}_j = \delta_{ij} \quad \text{for } i,j = 1,2,3. \tag{4.38}$$

It follows that

$$\widehat{\mathbf{u}}(\mu,\nu,k-1) = \mathbf{C}(\mu,\nu)\widehat{\mathbf{u}}(\mu,\nu,k), \tag{4.39}$$

in which

$$\mathbf{C}(\mu,\nu) = \begin{pmatrix} \widetilde{\mathbf{q}}_1^T \\ \widetilde{\mathbf{q}}_2^T \\ \widetilde{\mathbf{q}}_3^T \end{pmatrix}^{-1} \begin{pmatrix} \gamma_1^{-1}\widetilde{\mathbf{q}}_1^T \\ \gamma_2^{-1}\widetilde{\mathbf{q}}_2^T \\ \gamma_3^{-1}\widetilde{\mathbf{q}}_3^T \end{pmatrix}. \tag{4.40}$$

The matrix C is

$$\mathsf{C} = \mathcal{F}^{-1}\mathbf{C}\mathcal{F}, \tag{4.41}$$

in which

$$\mathbf{C} = \operatorname{diag}(\mathbf{C}(\mu,\nu))_{\mu=0,\ldots,N_x-1,\nu=0,\ldots,N_y-1}. \tag{4.42}$$

To construct the ABC operator $\mathcal{A}$, multiply $-\widehat{A}_{-10}^H(\mu,\nu)$ by $\mathbf{C}(\mu,\nu)$. Note that $\mathcal{A} = \mathcal{F}^{-1}\mathbf{A}\mathcal{F}$, where $\mathbf{A}$ is a diagonal block matrix consisting of the submatrices $\mathbf{A}(\mu,\nu) = -\widehat{A}_{-10}^H(\mu,\nu)\mathbf{C}(\mu,\nu)$ for $\mu = 0,\ldots,N_x-1$ and $\nu = 0,\ldots,N_y-1$, namely,

$$\mathbf{A} = \operatorname{diag}(\mathbf{A}(\mu,\nu))_{\mu=0,\ldots,N_x-1,\,\nu=0,\ldots,N_y-1}, \tag{4.43}$$

and also for both two and three dimensional cases, the operator $\mathcal{A} = \mathcal{F}^{-1}\mathbf{A}\mathcal{F}$ is symmetric and positive definite. It is quite difficult to see this directly from the Fourier analysis discussed in this section, but it follows from the variational principle based on the general form of the ABC operator as discussed in section 4.2.

Finally, Condition 4.1 can be validated from the construction of the matrix C. For any data $U_0 \in \mathbb{R}^{\mathrm{N}}$ which consists of displacement $\mathbf{u}$ on the interface $i = 0$, the vectors U_i for all $i < 0$ can be written as follows:

$$U_i = \mathcal{F}^{-1}\mathbf{C}^i\mathcal{F}U_0 = \mathsf{C}^i U_0. \tag{4.44}$$

The matrix C is invertible. It satisfies (4.17), because $|\gamma| > 1$ for $(\mu, \nu) \neq 0$, while for $(\mu, \nu) = 0$, $\gamma = 1$ and the corresponding term in (4.34) has no dependence on m and n, so that (4.16) is also satisfied. This completes the validation of Condition 4.1.

4.5. Total energy. In this section we derive alternative general energy formulas that involve a product of stress and strain. Note that in the section 4.2, the energy and the variational principle are written in terms of displacement times force. For some applications, such as a heteroepitaxial thin film, as described in section 5.3, it is much more convenient to write the energy in the form of stress times strain, as in (4.3).

The analysis of this section relies on the following "summation by parts" formulas:

$$\sum_{j\le 0}(D^+f)_j g_j = f_1 g_0 - \sum_{j\le 0} f_j (D^- g)_j, \tag{4.45}$$

$$\sum_{j\le 0}(D^-f)_j g_j = f_0 g_1 - \sum_{j\le 0} f_j (D^+ g)_j, \tag{4.46}$$

$$\sum_{j\le 0}(D^0f)_j g_j = \frac{1}{2}\left(f_1 g_0 + f_0 g_1\right) - \sum_{j\le 0} f_j (D^0 g)_j, \tag{4.47}$$

where D^+, D^-, and D^0 are the forward, backward, and centered finite difference operators, respectively. The total energy can be decomposed into two parts:

$$\mathcal{E}_{total} = \mathcal{E}_{i\ge 0} + \mathcal{E}_{i\le -1}, \tag{4.48}$$

where $\mathcal{E}_{i\ge 0} = \sum_{i\ge 0} E_i$ and $\mathcal{E}_{i\le -1} = \sum_{i\le -1} E_i$, and $i = 0$ is the layer in which the ABCs are imposed. Use (4.45)–(4.47) to derive the following relations, in two and three space dimensions, respectively:

$$\begin{aligned}\mathcal{E}_{i\le -1} = &\sum_{i_1} \alpha v_0 (D_y^+ v)_{-1} + \beta u_0 (D_y^+ u)_{-1} \\ &+ \sum_{i_1} \alpha v_{-1} (D_y^- v)_0 + \beta u_{-1} (D_y^- u)_0 \\ &+ \sum_{i_1} \big[\beta (u_0 (D_x^0 v)_{-1} + u_{-1} (D_x^0 v)_0) \\ &\qquad + \gamma \left(v_0 (D_x^0 u)_{-1} + v_{-1} (D_x^0 u)_0\right)\big] \\ &- \sum_{i_1, i\le -1} \frac{1}{2}\mathbf{u}_i \cdot (\mathcal{L}\mathbf{u}_i) - \mathbf{u}_i \cdot \mathbf{f}_i\end{aligned} \tag{4.49}$$

and

$$
\begin{aligned}
(4.50)\quad \mathcal{E}_{i\le -1} = &\sum_{i_1,i_2} \alpha w_0 (D_z^+ w)_{-1} + \beta (u_0 (D_z^+ u)_{-1} + v_0 (D_z^+ v)_{-1}) \\
&+ \sum_{i_1,i_2} \alpha w_{-1} (D_z^- w)_0 + \beta (u_{-1} (D_z^- u)_0 + v_{-1} (D_z^- v)_0) \\
&+ \sum_{i_1,i_2} \big[\beta \left(v_0 (D_y^0 w)_{-1} + v_{-1} (D_y^0 w)_0 + u_0 (D_x^0 w)_{-1} + u_{-1} (D_x^0 w)_0\right) \\
&\qquad + \gamma \left(w_0 (D_y^0 v)_{-1} + w_{-1} (D_y^0 v)_0 + w_{-1} (D_x^0 u)_0 + w_0 (D_x^0 u)_{-1}\right)\big] \\
&- \sum_{i_1,i_2,i\le -1} \frac{1}{2} \mathbf{u}_i \cdot (\mathcal{L} \mathbf{u}_i) - \mathbf{u}_i \cdot \mathbf{f}_i,
\end{aligned}
$$

in which $\mathcal{L}$ is the operator introduced in (4.26) and (4.33) and $\mathbf{f}_i$ is the force. In these formulas, the subscript refers to the depth-like index i.

Due to the assumption that $\mathbf{f}_i = 0$ for $i \le -1$, the last terms are zero in both the two and three dimensional cases. This leads to the following formulas:

$$
\begin{aligned}
(4.51)\qquad \mathcal{E}_{total} = &\mathcal{E}_{i\ge 0} + \sum_{i_1} \alpha v_0 (D_y^+ v)_{-1} + \beta u_0 (D_y^+ u)_{-1} \\
&+ \sum_{i_1} \alpha v_{-1} (D_y^- v)_0 + \beta u_{-1} (D_y^- u)_0 \\
&+ \sum_{i_1} \big[\beta (u_0 (D_x^0 v)_{-1} + u_{-1} (D_x^0 v)_0) \\
&\qquad + \gamma \left(v_0 (D_x^0 u)_{-1} + v_{-1} (D_x^0 u)_0\right)\big]
\end{aligned}
$$

in two dimensions and

$$
\begin{aligned}
(4.52)\quad \mathcal{E}_{total} = &\mathcal{E}_{i\ge 0} + \sum_{i_1,i_2} \alpha w_0 (D_z^+ w)_{-1} + \beta (u_0 (D_z^+ u)_{-1} + v_0 (D_z^+ v)_{-1}) \\
&+ \sum_{i_1,i_2} \alpha w_{-1} (D_z^- w)_0 + \beta (u_{-1} (D_z^- u)_0 + v_{-1} (D_z^- v)_0) \\
&+ \sum_{i_1,i_2} \big[\beta \left(v_0 (D_y^0 w)_{-1} + v_{-1} (D_y^0 w)_0 + u_0 (D_x^0 w)_{-1} + u_{-1} (D_x^0 w)_0\right) \\
&\qquad + \gamma \left(w_0 (D_y^0 v)_{-1} + w_{-1} (D_y^0 v)_0 + w_{-1} (D_x^0 u)_0 + w_0 (D_x^0 u)_{-1}\right)\big]
\end{aligned}
$$

in three dimensions. In both (4.51) and (4.52), we replace U_{-1} by $\mathsf{C}U_0$ whenever $\mathbf{u}_{-1}$ appears.

If the ABCs are imposed on the layer $i = 0$ where there is no force, then the total energy could be computed by the following new energy formulas that do not involve U_{-1} or the operator C:

$$
\begin{aligned}
(4.53)\qquad \mathcal{E}_{total} = &\mathcal{E}_{i>0} + \sum_{i_1} \alpha v_1 (D_y^+ v)_0 + \beta u_1 (D_y^+ u)_0 \\
&+ \sum_{i_1} \alpha v_0 (D_y^- v)_1 + \beta u_0 (D_y^- u)_1 \\
&+ \sum_{i_1} \big[\beta (u_1 (D_x^0 v)_0 + u_0 (D_x^0 v)_1) \\
&\qquad + \gamma \left(v_1 (D_x^0 u)_0 + v_0 (D_x^0 u)_1\right)\big]
\end{aligned}
$$

in two space dimensions and

$$\begin{aligned}(4.54)\quad \mathcal{E}_{total} = \mathcal{E}_{i>0} &+ \sum_{i_1,i_2} \alpha w_1(D_z^+ w)_0 + \beta(u_1(D_z^+ u)_0 + v_1(D_z^+ v)_0) \\ &+ \sum_{i_1,i_2} \alpha w_0(D_z^- w)_1 + \beta(u_0(D_z^- u)_1 + v_0(D_z^- v)_1) \\ &+ \sum_{i_1,i_2} \big[\beta\left(v_1(D_y^0 w)_0 + v_0(D_y^0 w)_1 + u_1(D_x^0 w)_0 + u_0(D_x^0 w)_1\right) \\ &\qquad + \gamma\left(w_1(D_y^0 v)_0 + w_0(D_y^0 v)_1 + w_0(D_x^0 u)_1 + w_1(D_x^0 u)_0\right)\big]\end{aligned}$$

in three space dimensions, respectively. In both (4.53) and (4.54), we replace U_{-1} by $-(A_{-10}^T)^{-1}\mathcal{A}U_0$ whenever $\mathbf{u}_{-1}$ appears.

5. Numerical results. In this section, sample computations are performed to validate and illustrate the ABCs developed in previous sections. Throughout this section, the elastic constants C_{11}, C_{12}, C_{44} are assumed to be $C_{11} = 8$, $C_{12} = 4$, and $C_{44} = 4$ unless explicitly stated otherwise.

5.1. The ABCs for continuum elasticity. This section shows the effectiveness of the ABCs for continuum elasticity equations (3.7). The Lamé constants are chosen to be $\lambda = 1$ and $\tau = 1$.

The test problem is (3.7) on $\Omega = [0, 2\pi)\times(-\infty, 0)$ with data on $\Gamma_1 = [0, 2\pi)\times\{y = 0\}$. Periodicity is assumed in the lateral direction, and there is no body force; i.e., $\mathbf{f} = 0$. The interface Γ_2 at which the artificial boundary condition is imposed is the line $[0, 2\pi) \times \{y = -1\}$.

The Dirichlet data given on Γ_1 is as follows:

$$\mathbf{u} = (u, v) = (\cos x + \sin 2x, 0),$$

for which the exact solution to (3.7) is

$$(5.1)\qquad \mathbf{u} = \left(\left(1 + \frac{y}{2}\right)\cos x\, \mathrm{e}^y + (1 + y)\sin 2x\, \mathrm{e}^{2y}, \frac{7}{2}\sin x\, \mathrm{e}^y - y\cos 2x\, \mathrm{e}^{2y}\right).$$

This exact solution is compared to the solution of (3.7) with the exact artificial boundary condition (3.16) on the interface Γ_2 and also to the solutions with the following two alternative boundary conditions:

- The zero Dirichlet boundary condition $\mathbf{u}(x, -1) = 0$.
- The Neumann boundary condition $\mathbf{n}\cdot\mathbf{T} = 0$ on $y = -1$.

Figure 5.1 shows the u-displacement field at the line $y = -0.75$ for the exact solution, the solution using ABCs, and the two alternative solutions. Although there is still error, due to discretization of the continuum equation, it is clear that the solution obtained with the exact ABCs (3.16) is in good agreement with the analytic solution (5.1). On the other hand, the solutions obtained with the other two boundary conditions are in error by about 20%–30% at the peaks.

5.2. The ABCs for discrete elasticity. In this section, we investigate the ABCs for the discrete elastic equations for both two and three space dimensions with the Dirichlet data given on the boundary Γ_1. As in the continuum case, there are no external forces, and periodic boundary conditions are imposed in the lateral directions.

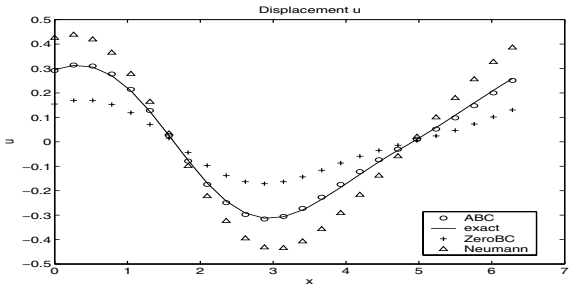


FIG. 5.1. *Test of the ABCs for the continuum solution in two space dimensions. Comparison of the u-displacement field of* $\mathbf{u} = (u, v)$ *given at* $y = -0.75$ *for the exact solution (line) and for the following boundary conditions: ABC (circle), zero-displacement (plus), and Neumann (triangle).*

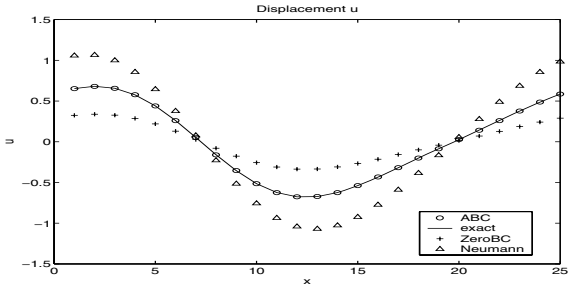


FIG. 5.2. *Test of the exact discrete ABCs in two dimensions: a comparison of the u-displacement field of* $\mathbf{u} = (u, v)$ *at* $(x, y) = (x, 3)$. *The boundary* Γ_1 *is at* $y = 1$, *and the interface* Γ_2 *is at* $y = 5$ *for the exact solution (line) and for the following boundary conditions: ABC (circle), zero-displacement (plus), and Neumann (triangle).*

More precisely, for the two dimensional case, the lattice Ω consists of $N_x = 25$ layers in the x-direction and $N_y = 5$ in the y-direction, and the prescribed Dirichlet boundary condition for $\mathbf{u}$ on $\Gamma_1 = \{y = 1\}$ is

$$\mathbf{u} = (\cos x + \sin 2x, \sin x). \tag{5.2}$$

For the three dimensional case, the lattice Ω consists of $N_x = N_y = 25$ layers in the x- and y-directions and $N_z = 4$ layers in the z-direction, and the Dirichlet data on $\Gamma_1 = \{z = 1\}$ is

$$\mathbf{u} = (\cos x + \sin 2x, \sin y, \sin x). \tag{5.3}$$

Numerical results are plotted in Figures 5.2 and 5.3. For numerical experiments, the exact ABCs and other approximate boundary conditions are imposed on $\Gamma_2 = \{y = 5\}$ for the two dimensional case and $\Gamma_2 = \{z = 4\}$ for the three dimensional case, respectively. The results show that the solution with the ABCs is much more accurate than those from the Dirichlet and Neumann boundary conditions. Indeed, the accuracy obtained with the ABCs operator is within the round-off error, i.e., $O(10^{-14})$.

5.3. Numerical simulations for thin films. In heteroepitaxial growth, a thin film of one material (e.g., Ge) is grown on top of a substrate of a second material (e.g., Si), with perfect, single crystalline structure in both materials and with the lattice

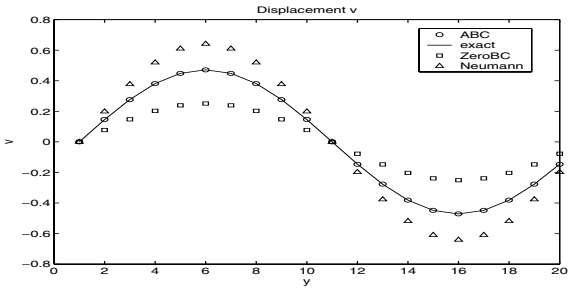


FIG. 5.3. *Test of the exact discrete ABCs in three dimensions: a comparison of the v-displacement field of* $\mathbf{u} = (u, v, w)$ *at* $(x, y, z) = (10, y, 3)$. *The boundary* Γ_1 *is at* $z = 1$, *and the interface* Γ_2 *is at* $z = 4$ *for the exact solution (line) and for the following boundary conditions: ABC (circle), zero-displacement (square), and Neumann (triangle).*

structure of the film determined by the substrate. If the lattice constants a_f and a_s for the film and substrate are different (e.g., $a_{Ge} = 1.04 \times a_{Si}$), then strain is generated in the film. This strain has important effects on the material structure, as well as on its electronic properties.

For this system, it is most convenient to define the atomic displacement relative to a single reference lattice, for example, the equilibrium lattice of the substrate, so that the displacement $\mathbf{u}$ in the film is defined relative to a nonequilibrium reference lattice. The bond displacement $\mathbf{d}^{\mathbf{k}\pm}$ is then

$$\mathbf{d}^{\mathbf{k}\pm}(\mathbf{i}) = (d_1^{k\pm}, d_2^{k\pm}, d_3^{k\pm}) = D_k^{\pm}\mathbf{u}(\mathbf{i}) - \epsilon \mathbf{e}_{\mathbf{k}}\chi, \tag{5.4}$$

in which $\epsilon = \frac{a_f - a_s}{a_s}$ is the relative lattice displacement, and χ is 0 in the substrate and 1 in the film. The resulting discrete strain equations have a force of size ϵ along the film/substrate interface, and the energy has the form

$$\mathcal{E}_{total} = \frac{1}{2}(\mathbf{H}\mathsf{U}, \mathsf{U}) - (\mathsf{F}, \mathsf{U}) + \mathcal{G}(\epsilon), \tag{5.5}$$

where

$$(\mathsf{F}, \mathsf{U}) = \sum_i \sum_{p=\pm, k=1,2,3} \epsilon\, D_k^p u_k \chi. \tag{5.6}$$

Further details are given, for example, in [5].

In this section, we compare the displacement fields $\mathbf{u}$ that are computed with the ABCs and with zero boundary conditions for a heteroepitaxial thin film. Since the forces lie on the film/substrate boundary, the artificial boundary can be taken to be any plane below this interface. Our computational domain is three dimensional with Γ_2 being of size 10×10. As in the last section, we denote NC to be the thickness of the substrate, including Γ_2. Note that on the top boundary Γ_1, the homogeneous Neumann boundary condition (no external force) is imposed.

To demonstrate the effectiveness of the ABCs, we first compute the displacement field $\mathbf{u}$ by imposing the ABCs on Γ_2 with substrate thickness NC = 1 and take it as the reference solution. We then compute two displacement fields that are generated by imposing zero boundary conditions on the bottom boundary with NC = 2 and NC = 8. For these three solutions, Figure 5.4 shows a comparison of the u component

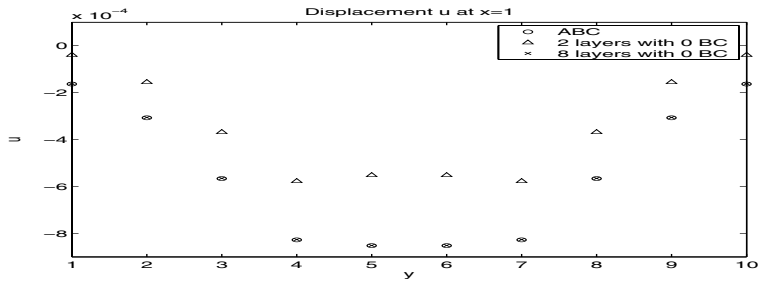


FIG. 5.4. *The comparison of u-displacement on the second layer from the top boundary* Γ_1 *with* $x = 1$. *The u-displacement computed with the ABC (circle) imposed on the first substrate layer and u-displacement computed with zero boundary condition with* NC $= 2$ *(triangle) and* NC $= 8$ *(cross).*

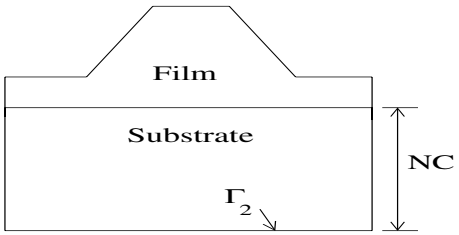


FIG. 5.5. *Schematic drawing of quantum dot geometry.*

of the displacement vector $\mathbf{u} = (u, v, w)$ on the line $x = 1$ in the second layer from the top. It is clear that the displacement field computed with the zero boundary conditions approaches the reference displacement field as the number of substrate layers increases. In addition (not shown in Figure 5.4), the results from the ABC are found to be independent (i.e., within round-off error) of the depth at which the ABC is applied.

5.4. Energy computation. This section presents results to validate the total energy formulas (4.51) and (4.52) derived in section 4.5. As in the previous section, NC denotes the number of substrate layers, including Γ_2 itself. In addition, E_A denotes the total energy computed by imposing the ABC on Γ_2, and E_Z denotes the total energy computed with the zero boundary condition on Γ_2.

For computational purposes, we take a geometry corresponding to a periodic array of quantum dots. A typical geometry is illustrated in Figure 5.5. For two space dimensions, Γ_2 is one dimensional with the material system of size $N_x = 128$ and the quantum dot of base size 64. For three space dimensions, Γ_2 is two dimensional with the material system of size $N_x = N_y = 10$ and the quantum dot of base size 8×8.

In order to validate the total energy formulas (4.51) and (4.52), by numerical computation we show first that the total energy E_A does not depend on the thickness of the substrate NC and second that the total energy E_Z obtained by imposing zero boundary conditions on Γ_2 approaches the total energy E_A as the thickness of substrates NC increases. These computational results are demonstrated in Figure 5.6, in which the thickness of the substrate NC varies from NC $= 2$ to NC $= 120$ for two space dimensions and from NC $= 2$ to NC $= 14$ for three space dimensions. The units of the total energy are 10^{12}dyne/cm^2.

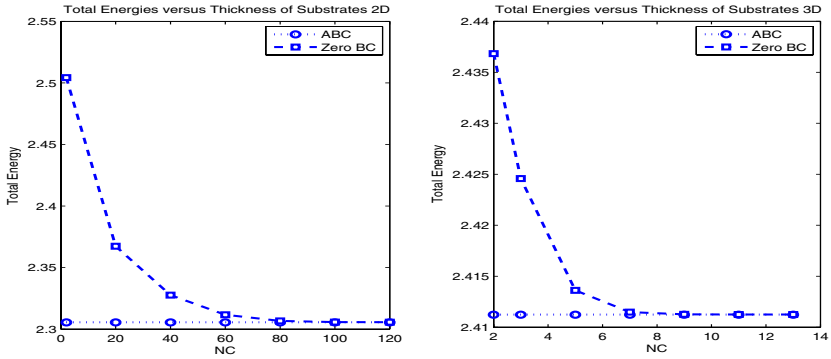


FIG. 5.6. *Total energies obtained by applying the ABC (circle) and zero boundary condition (square) as a function of the thickness of substrates for two dimensions (left) ($N_x = 128$) and three dimensions (right) ($N_x = N_y = 10$).*

6. Conclusions. In this paper, we have derived the ABCs for continuum and discrete elasticity equations. A solvability condition has been formulated and validated, under which the discrete equations in an unbounded domain can be shown to be well-posed and the reduced force balance equation can be derived. Its solution coincides with the exact solution when restricted to the bounded domain. Furthermore, a new total energy formula has been derived so that it can be computed by using only the displacement field in the region above the artificial boundary.

These results are currently being used for modeling and simulation of the growth of thin epitaxial films. By exploiting the symmetry of the resulting force balance equations in further work, we shall combine the ABCs with a multigrid method to get an accelerated simulation method for various applications.

Appendix. Several technical lemmas.

LEMMA A.1. *The matrix A_{-10} is invertible.*

Proof. Observe that

$$A_{-10}U_i = \mathcal{F}^{-1}\widehat{A}_{-10}\mathcal{F}(U_i), \tag{A.1}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ are Fourier and inverse Fourier transformations and $\widehat{A}_{-10}$ is a 3×3 (2×2 in two space dimensions) block matrix, such that for any given Fourier mode (μ,ν),

$$\widehat{A}_{-10}(\mu,\nu) = \begin{pmatrix} -C_{44} & 0 & -s_1 \\ 0 & -C_{44} & -s_2 \\ -s_1 & -s_2 & -C_{11} \end{pmatrix}, \tag{A.2}$$

where

$$s_1 = \imath\frac{(C_{12}+C_{44})}{2}\sin\left(\frac{2\pi\mu}{N_x}\right),$$
$$s_2 = \imath\frac{(C_{12}+C_{44})}{2}\sin\left(\frac{2\pi\nu}{N_y}\right).$$

The eigenvalues for $\widehat{A}_{-10}(\mu,\nu)$ can be obtained by solving the following equation:

$$\begin{aligned}\det(\widehat{A}_{-10}(\mu,\nu)-\lambda \mathrm{I}) &= -(C_{44}+\lambda)\left[(C_{44}+\lambda)(C_{11}+\lambda)-s_1^2-s_2^2\right]\\ &= -(C_{44}+\lambda)\left[\lambda^2+(C_{11}+C_{44})\lambda+C_{11}C_{44}\right.\\ &\qquad +\sin^2(2\pi\mu/N_x)(C_{12}+C_{44})^2/4\\ &\qquad \left.+\sin^2(2\pi\nu/N_y)(C_{12}+C_{44})^2/4\right].\end{aligned} \tag{A.3}$$

Hence, three eigenvalues λ_1,λ_2, and λ_3 are given as follows:

$$\begin{aligned}\lambda_1 &= -C_{44},\\ 2\lambda_2 &= -(C_{11}+C_{44})\\ &\quad +\sqrt{(C_{11}-C_{44})^2-(\sin^2(2\pi\mu/N_x)+\sin^2(2\pi\nu/N_y))(C_{12}+C_{44})^2},\\ 2\lambda_3 &= -(C_{11}+C_{44})\\ &\quad -\sqrt{(C_{11}-C_{44})^2-(\sin^2(2\pi\mu/N_x)+\sin^2(2\pi\nu/N_y))(C_{12}+C_{44})^2}.\end{aligned}$$

The eigenvalue with the smallest magnitude is λ_2 with $\sin(2\pi\nu/N_x)=\sin(2\pi\mu/N_y)=0$, in which case

$$2\lambda_2 = (-(C_{11}+C_{44})+|C_{11}-C_{44}|) = -2\min(C_{11},C_{44}). \tag{A.4}$$

It follows that no eigenvalues can be zero; hence A_{-10} is invertible. This completes the proof. ∎

LEMMA A.2. *For $\gamma_i \neq \gamma_j$, the corresponding eigenvectors $\mathbf{q}_i$ and $\mathbf{q}_j$ are linearly independent.*

Proof. Consider the linear reformulation of the palindromic eigenvalue problem (4.30) by introducing $x=\gamma y$ as follows: With $P(\mu,\nu,\gamma)=\gamma^2\widehat{A}_{-10}+\gamma\widehat{A}_{00}+\widehat{A}_{-10}^H$,

$$\begin{pmatrix} 0 & I \\ -\widehat{A}_{-10}^H & -\widehat{A}_{00}\end{pmatrix}\begin{pmatrix} y\\ x\end{pmatrix} = \gamma\begin{pmatrix} I & 0\\ 0 & \widehat{A}_{-10}\end{pmatrix}\begin{pmatrix} y\\ x\end{pmatrix}. \tag{A.5}$$

From the fact that $\widehat{A}_{-10}$ is invertible, it is obvious that the eigenvectors $\mathbf{q}_i$ and $\mathbf{q}_j$ that correspond to different eigenvalues γ_i and γ_j must be linearly independent. ∎

LEMMA A.3. *Under Condition* 4.1, *the matrix* M *given in* (4.20) *is an isomorphic mapping from* Θ *to* Θ^*.

Proof. For $\mathsf{V}^- = (V_{-1},V_{-2},\dots)^T \in \Theta$ with $V_i = \mathsf{C}^i V_0$ for $i\le 0$, as in Condition 4.1,

$$\begin{aligned}\mathsf{M}\mathsf{V}^- &= (-A_{-10}V_0,0,\dots,0,\dots)^T\\ &= \mathsf{G} = (G_{-1},0,\dots,0,\dots)^T\end{aligned} \tag{A.6}$$

if $G_{-1}=-A_{-10}V_0$. Since A_{-10} is invertible, this shows that the matrix M is onto.

To show that M is one to one, it is enough to show that $\mathsf{M}\mathsf{V}^-=0$ implies $\mathsf{V}^-=0$. Consider the energy

$$\mathcal{E}^- = \sum_{i<0} E_i \tag{A.7}$$

over the space Θ and observe that

$$(\mathsf{M}\mathsf{V}^-,\mathsf{V}^-) = \widehat{\mathcal{E}}^-, \tag{A.8}$$

in which $\widehat{\mathcal{E}}^-$ is $\mathcal{E}^-$ for $V_0 = 0$. Therefore, $\mathsf{MV}^- = 0$ implies that $\widehat{\mathcal{E}}^- = 0$. Connectivity of the lattice and $U_0 = 0$ then imply that $U_i = 0$ for all $i < 0$. This shows that the matrix M is one to one. Therefore, $\mathsf{M} : \Theta \mapsto \Theta^*$ is isomorphic. ∎

Proof of Lemma 4.2.

First, we show that $\mathcal{A}$ is symmetric. For any $\mathsf{U} = (0, 0, \dots, 0, U_0, U_{-1}, \dots, \dots)^T$ and $\mathsf{V} = (0, 0, \dots, 0, V_0, V_{-1}, \dots, \dots)^T$ that belong to the space $\mathbf{V}$, Condition 4.1 implies that

$$A_{-10}^T V_{-1} = -\mathcal{A} V_0 \quad \text{and} \quad A_{-10}^T U_{-1} = -\mathcal{A} U_0. \tag{A.9}$$

Note also that $\widetilde{\mathcal{E}}(\mathsf{U}, \mathsf{V}) = \widetilde{\mathcal{E}}(\mathsf{V}, \mathsf{U})$; i.e.,

$$\begin{aligned} \widetilde{\mathcal{E}}(\mathsf{U}, \mathsf{V}) &= \frac{1}{2}\left(U_0, \left(A_{00} V_0 + A_{-10}^T V_{-1}\right)\right) \\ &= \frac{1}{2}\left(V_0, \left(A_{00} U_0 + A_{-10}^T U_{-1}\right)\right) = \widetilde{\mathcal{E}}(\mathsf{V}, \mathsf{U}). \end{aligned} \tag{A.10}$$

Use (A.9) in (A.11) to obtain

$$(U_0, (A_{00} V_0 - \mathcal{A} V_0)) = (V_0, (A_{00} U_0 - \mathcal{A} U_0)).$$

Since A_{00} is symmetric, this implies that $(U_0, \mathcal{A} V_0) = (V_0, \mathcal{A} U_0)$ for all $U_0, V_0 \in \mathbb{R}^N$ and therefore, that $\mathcal{A}$ is symmetric. The symmetry of the operator $\mathcal{A}$ implies that the matrix $\widehat{\mathbf{H}}$ is symmetric.

Next, we show that $\mathcal{A}$ is positive definite since for $U_0 \neq 0 \in \mathbb{R}^N$,

$$\begin{aligned} (U_0, \mathcal{A} U_0) &= (U_0, \mathsf{B}\mathsf{M}^{-1}\mathsf{B}^T U_0) = (\mathsf{B}^T U_0, \mathsf{M}^{-1}\mathsf{B}^T U_0) \\ &= (\mathsf{M}\mathsf{U}^-, \mathsf{M}^{-1}\mathsf{M}\mathsf{U}^-) = (\mathsf{M}\mathsf{U}^-, \mathsf{U}^-) = \widehat{\mathcal{E}}^- > 0, \end{aligned}$$

where U^- is the unique solution of $\mathsf{M}\mathsf{U}^- = \mathsf{B}^T U_0$. Finally, we show that the matrix $\widehat{\mathbf{H}}$ is nonnegative definite. First note that $\widehat{\mathbf{H}}\mathbf{1} = 0$. Furthermore, there is no other null space for $\widehat{\mathbf{H}}$, since

$$\begin{aligned} \widehat{\mathbf{H}} \begin{pmatrix} \mathsf{U}^+ \\ U_0 \end{pmatrix} = 0 &\Longleftrightarrow (U_0, (A_{01} U_1 + (A_{00} - \mathcal{A}) U_0)) \\ &\quad + \sum_{i>0} (U_i, (A_{ii-1} U_{i-1} + A_{ii} U_i + A_{ii+1} U_{i+1})) = 0 \\ &\Longleftrightarrow \widetilde{\mathcal{E}}(\mathsf{U}, \mathsf{U}) = 0 \quad \text{with} \quad \mathsf{U} \in \mathbf{V} \\ &\Longleftrightarrow \mathsf{U} = \mathbf{1} \quad \text{by connectivity of the lattice.} \end{aligned}$$

This completes the proof of Lemma 4.2. ∎

Acknowledgment. The authors wish to thank the anonymous referees whose remarks helped us to improve our manuscript.

REFERENCES

[1] C. R. Anderson, *The Application of Domain Decomposition to the Solution of Laplace's Equation in Infinite Domains*, CAM Report 87–19, University of California at Los Angeles, Los Angeles, 1987.

[2] X. Antoine, C. Besse, and S. Descombes, *Artificial boundary conditions for one-dimensional cubic nonlinear Schrödinger equations*, SIAM J. Numer. Anal., 43 (2006), pp. 2272–2293.

[3] A. Bayliss, M. Gunzburger, and E. Turkel, *Boundary conditions for the numerical solutions of elliptic equations in exterior regions*, SIAM J. Appl. Math., 42 (1982), pp. 430–451.

[4] R. E. Caflisch, Y.-J. Lee, S. Shu, Y. Xiao, and J. Xu, *An application of multigrid methods for a discrete elastic model for epitaxial systems*, J. Comput. Phys., to appear.
[5] C. Connell, R. E. Caflisch, E. Luo, and G. D. Simms, *The elastic field of a surface step: The Marchenko–Parshin formula in the linear case*, J. Comput. Appl. Math, to appear.
[6] M. Ehrhardt, *Finite difference schemes on unbounded domains*, in Advances in the Applications of Nonstandard Finite Difference Schemes, Vol. 2, World Scientific, Hackensack, NJ, 2005, pp. 343–384.
[7] D. Givoli, *Numerical Methods for Problems in Infinite Domains*, Elsevier, Amsterdam, 1992.
[8] D. Givoli and J. B. Keller, *Nonreflecting boundary conditions for elastic waves*, Wave Motion, 12 (1990), pp. 261–279.
[9] D. Givoli, I. Patlashenko, and J. B. Keller, *High-order boundary conditions and finite elements for infinite domains*, Comput. Methods Appl. Mech. Engrg., 143 (1997), pp. 13–39.
[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1996.
[11] T. Hagstrom and H. B. Keller, *Exact boundary conditions at an artificial boundary for partial differential equations in cylinders*, SIAM J. Math. Anal., 17 (1986), pp. 322–341.
[12] H. Han and W. Bao, *Error estimates for the finite element approximation of linear elastic equations in an unbounded domain*, Math. Comp., 70 (2000), pp. 1437–1459.
[13] H. Han, W. Bao, and T. Wang, *Numerical simulation for the problem of infinite elastic foundation*, Comput. Methods Appl. Mech. Engrg., 147 (1997), pp. 369–385.
[14] H. Han and X. Wu, *The approximation of the exact boundary conditions at an artificial boundary for linear elastic equations and its application*, Math. Comp., 59 (1992), pp. 21–37.
[15] A. Hilliges, C. Mehl, and V. Mehrmann, *On the solution of palindromic eigenvalue problems*, in Proceedings of the European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS), 2004.
[16] M. E. Hochstenbach and H. A. van der Vorst, *Alternatives to the Rayleigh quotient for the quadratic eigenvalue problem*, SIAM J. Sci. Comput., 25 (2003), pp. 591–603.
[17] L. D. Landau and E. M. Lifshitz, *Theory of Elasticity*, Butterworth-Heinemann, Oxford, UK, 1986.
[18] S. Lee, *Artificial Boundary Conditions for Linear Elasticity and Atomistic Strain Models*, Ph.D. thesis, University of California at Los Angeles, Los Angeles, 2005.
[19] P. G. Martinsson and I. Babuska, *Mechanics of materials with periodic truss or frame micro-structures* I: *Korn's inequality*, in Arch. Ration. Mech. Anal., to appear.
[20] G. Russo and P. Smereka, *Computation of strained epitaxial growth in three dimensions by kinetic Monte Carlo*, J. Comput. Phys., 214 (2006), pp. 809–828.
[21] A. Schindler, M. F. Gyure, G. D. Simms, D. D. Vvendensky, R. E. Caflisch, C. Connell, and E. Luo, *Theory of strain relaxation in heteroepitaxial systems*, Phys. Rev. B, 67 (2003), no. 075316.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1776–1792

DYNAMIC BEHAVIOR OF A PACED CARDIAC FIBER*

JOHN W. CAIN†

Abstract. Consider a typical experimental protocol in which one end of a one-dimensional fiber of cardiac tissue is periodically stimulated, or paced, resulting in a train of propagating action potentials. There is evidence that a sudden change in the pacing period can initiate abnormal cardiac rhythms. In this paper, we analyze how the fiber responds to such a change in a regime without arrhythmias. In particular, given a fiber length L and a tolerance η, we estimate the number of beats $N = N(\eta, L)$ required for the fiber to achieve approximate steady-state in the sense that spatial variation in the diastolic interval (DI) is bounded by η. We track spatial DI variation using an infinite sequence of linear integral equations which we derive from a standard kinematic model of wave propagation. The integral equations can be solved in terms of generalized Laguerre polynomials. We then estimate N by applying an asymptotic estimate for generalized Laguerre polynomials. We find that, for fiber lengths characteristic of cardiac tissue, it is often the case that N effectively exhibits no dependence on L. More exactly, (i) there is a critical fiber length L^* such that, if $L < L^*$, the convergence to steady-state is slowest at the pacing site, and (ii) often, L^* is substantially larger than the diameter of the whole heart.

Key words. cardiac fiber, pacing, transient behavior, restitution, kinematic model, generalized Laguerre polynomials

AMS subject classifications. 92C50, 33C45, 92C30

DOI. 10.1137/05063845X

1. Introduction.

Cardiac cells have the property of *excitability*: when a stimulus current of sufficient strength is applied to a quiescent cell, the transmembrane voltage v undergoes a prolonged elevation, called an *action potential*, before eventually returning to its resting value. Repeatedly stimulated, or *paced*, cardiac cells exhibit sequences of action potentials. By specifying a threshold voltage $v = v_{\text{thr}}$, one may define the *action potential duration (APD)* as the amount of time in which $v > v_{\text{thr}}$ during an action potential. The recovery time during which $v < v_{\text{thr}}$ between successive action potentials is called the *diastolic interval (DI)*. As illustrated in Figure 1, we shall denote the APD following the nth stimulus by A_n and the subsequent DI by D_n.

Periodic pacing leads to one of several types of phase-locked responses depending on the underlying pacing period B. For large B, cells exhibit a 1:1 response in which every stimulus yields an identical action potential. For smaller B, one sometimes observes a period-2 response, known as *alternans*, in which APD and DI values exhibit beat-to-beat alternation [21, 23, 24, 28, 29]. If B is decreased even further, cells exhibit a 2:1 response in which only every other stimulus yields an action potential [15, 21, 35]. In what follows, we shall assume that all pacing periods are sufficiently large to ensure a 1:1 steady-state response.

In spatially extended tissue, neighboring cells are coupled electrically via gap junctions, allowing action potentials to propagate through the tissue [19, 25]. Below,

*Received by the editors August 19, 2005; accepted for publication (in revised form) May 1, 2006; published electronically July 31, 2006. The content of this article is adapted from the last chapter of the author's doctoral dissertation [3]. This work was supported by the National Science Foundation under grants DMS-9983320 and DMS-0244492 and the National Institutes of Health under grant 1R01-HL-72831.

http://www.siam.org/journals/siap/66-5/63845.html

†Department of Mathematics, Virginia Commonwealth University, Richmond, VA 23284-2014 (jwcain@vcu.edu).

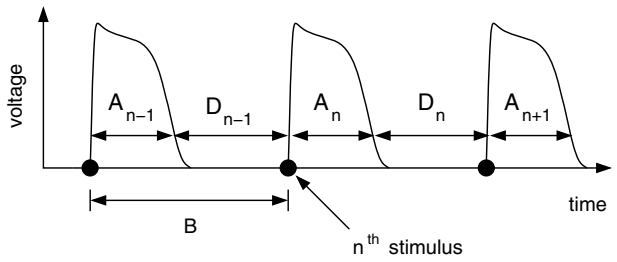


FIG. 1. *Voltage trace of several action potentials in a paced cardiac cell.*

we shall study the dynamics of a paced cardiac fiber composed of cylindrical cells joined together in an end-to-end fashion. We assume that voltage exhibits negligible radial dependence, varying only as a function of a length variable x; that is, the fiber can be treated as one-dimensional. Moreover, we shall assume that pacing is performed at one end of the fiber which we identify with $x = 0$.

Typically, propagation of action potentials in a one-dimensional fiber is modeled using a reaction-diffusion equation known as the *cable equation*, which, after nondimensionalization, takes the form

$$\frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2} + g(v, w). \tag{1}$$

Here, $v = v(x, t)$ is the transmembrane voltage and w is a vector of various dynamic variables that are used in modeling the ionic mechanism of the action potential. For a derivation of the cable equation, see the texts of Plonsey and Barr [25] and Keener and Sneyd [19]. Examples of studies in which the cable equation is used to model cardiac dynamics include [5, 6, 7, 18, 22].

Although the cable equation serves as a popular model, we remark that arrhythmias, by nature, concern the *timing* of excitation and recovery of the cells. Therefore, it is often desirable to track the progress of propagating action potentials without regard to the structure of the voltage profile. Indeed, many recent studies [4, 8, 11, 12, 16, 30, 34] have employed *kinematic* models [19, 27] of wave propagation in cardiac fibers.

In this paper, we use a kinematic model to investigate how a fiber of length L responds to a sudden change in the pacing period, say from B_{old} to B_{new}. Changing the pacing period introduces spatial variation in APD and DI. Our primary goal is to estimate the number of beats required for the fiber to "adjust" to the new pacing period, i.e., the number of beats required to reach approximate steady-state in the sense that spatial variation in DI is small. No previous studies have analyzed the transient behavior following a change in the pacing period. In the course of solving our main problem, we shall provide such an analysis. Describing the persistence of spatial DI variation under such a pacing protocol may lead to an improved understanding of the mechanisms for initiation of arrhythmias such as discordant alternans [34].

To establish notation, refer to Figure 2, which illustrates both the spatial variation in DI induced by changing the pacing period from B_{old} to $B_{\text{new}} < B_{\text{old}}$ and the aforementioned convergence to a steady-state. Here, $D_n = D_n(x)$ denotes the nth DI following the change in the pacing period—in particular, $n = 1$ corresponds to the first beat with period B_{new}. Figure 2(a) shows $D_n(x)$ for $n = 1, \ldots, 4$ and Figure 2(b) shows $D_n(x)$ for $n = 7, 8$. Note that $D_n(x)$ appears to converge pointwise to a constant

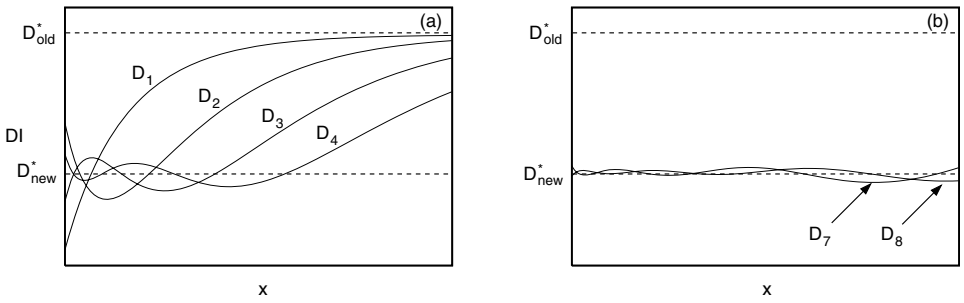


FIG. 2. *Spatial variation in DI after changing the pacing period from* B_{old} *to* $B_{\text{new}} < B_{\text{old}}$. (a) *the curves* $D_n(x)$ *for* $n = 1, \ldots, 4$; (b) *the curves* $D_n(x)$ *for* $n = 7, 8$.

D^*_{new} as $n \to \infty$.

More quantitatively, our main problem may be stated as follows. Let $\eta > 0$ be a given tolerance, and let n and D^*_{new} be as in the preceding paragraph.

Goal: Estimate the number of beats $N = N(\eta, L)$ such that $|D_n(x) - D^*_{\text{new}}| < \eta$ for all $x \in [0, L]$, $n \geq N$.

Our analysis shows that, for fibers shorter than a critical length $L = L^*$, the convergence is slowest at the pacing site $x = 0$. In other words, $N(\eta, L)$ does not depend on the fiber length L provided that $L < L^*$. Moreover, we find that $L^* \to \infty$ as the slope of the restitution curve at $DI = D^*_{\text{new}}$ tends to 1. Hence, $N(\eta, L)$ is especially unlikely to exhibit any length dependence as we approach the bifurcation to alternans.

The remainder of this paper is organized as follows. In section 2, we recall a kinematic model [19, 27] of wave propagation, which allows us to follow the progress of each action potential without tracking the complete voltage profile $v(x, t)$. From the kinematic model, we derive a recursive sequence of linear equations which can be solved to yield approximations $y_n(x)$ of $D_n(x) - D^*_{\text{new}}$, allowing us to monitor the convergence to steady-state. As explained in section 3, the functions $y_n(x)$ can be expressed in terms of generalized Laguerre polynomials. The behavior of the functions $y_n(x)$ can be approximated by recalling a large-n asymptotic estimate for the generalized Laguerre polynomials. This allows us to estimate the maximum of $|y_n(x)|$ on the interval $[0, L]$, thereby leading to an estimate of $N(\eta, L)$. The estimate of $N(\eta, L)$ is given by one of two formulas according to whether $L < L^*$ or $L > L^*$. Section 4 contains a summary and discussion of our results.

2. Derivation of the governing equations. We begin this section with a brief discussion of the restitution and dispersion curves. We then recall a kinematic model of action potentials propagating in a paced fiber. From the equations of the kinematic model, we derive a sequence of linear equations which will allow us to solve the main problem.

2.1. Restitution and dispersion curves. Cardiac cells exhibit electrical *restitution*: The steady-state APD at a given pacing period B decreases as B is shortened. Nolasco and Dahlen [23] were among the first to model restitution with a mapping

$$A_{n+1} = f(D_n) = f(B - A_n). \tag{2}$$

Guevara et al. [14] later showed that alternans can result from a period-doubling bifurcation of (2) as the pacing period B is decreased. The function f is called the *restitution function*, and its graph is called the *restitution curve*. The restitution curve is typically monotone increasing; i.e., more recovery time yields longer excitations. Many authors (see, for example, [1, 2, 15]) have fit restitution data with exponential functions of the form

$$f(DI) = APD_{\max} - ke^{-DI/\tau}, \tag{3}$$

where $APD_{\max}$, k, and τ are positive constants. We shall not specify a functional form for the restitution function but will assume that f has the same qualitative shape as (3).

Just as APD depends upon the preceding DI, the wave front velocity of an action potential in a fiber depends upon the preceding (local) DI. This dependence is often displayed graphically via the *dispersion curve*, which typically has the same qualitative shape as the restitution curve. We shall denote the functional form of the dispersion curve by $c(DI)$.

2.2. Kinematic model of wave propagation. To solve the problem of estimating $N(\eta, L)$, we need only track the progress of action potentials, not their complete structure. Hence, we shall employ a kinematic model of wave propagation [19, 27]. In doing so, we implicitly assume that recovery always occurs via a phase wave [9, 33]; i.e. the wave back of each action potential is not greatly affected by diffusion. We also adopt the following assumptions, the last two of which are specific to the pacing protocol described in the introduction:

- (A1) To a reasonable approximation, the tissue does not exhibit *memory*: As implicitly assumed in (2), A_{n+1} depends only upon D_n and is not greatly influenced by the past pacing history. Likewise, the wave front velocity of the $(n+1)$st action potential depends only upon $D_n(x)$, the preceding local DI.
- (A2) The restitution and dispersion curves are monotone increasing.
- (A3) To implement the pacing protocol outlined in the introduction, the interval between the nth and $(n+1)$st stimuli is B_{old} if $n \leq 0$ and B_{new} if $n > 0$.
- (A4) Prior to the change in the pacing period (i.e., for $n \leq 0$), the long-term pacing with period B_{old} leads to a 1:1 steady-state response in which DI is a constant,[1] say D^*_{old}. In particular, $D_0(x) \equiv D^*_{\text{old}}$.

We now recall how to use the information contained in the restitution and dispersion curves to track the wave fronts and wave backs of the action potentials. If we pace one end (say $x = 0$) of a fiber and plot $v(x,t)$ versus x and t, we obtain a surface in three-dimensional space. Taking the intersection of this surface with the plane $v = v_{\text{thr}}$, we generate a sequence of curves which we identify with the wave fronts and wave backs of the action potentials. Projecting these curves onto the xt plane yields a schematic space-time plot of the wave fronts and wave backs as illustrated in Figure 3. Here, $\phi_n(x)$ (resp., $\beta_n(x)$) denotes the time at which the nth wave front (resp., wave back) arrives at x, and

$$CL_n(x) = \phi_{n+1}(x) - \phi_n(x) \tag{4}$$

[1] Referring to (2) with $B = B_{\text{old}}$, note that D^*_{old} is the unique DI satisfying the equation $DI + f(DI) = B_{\text{old}}$.

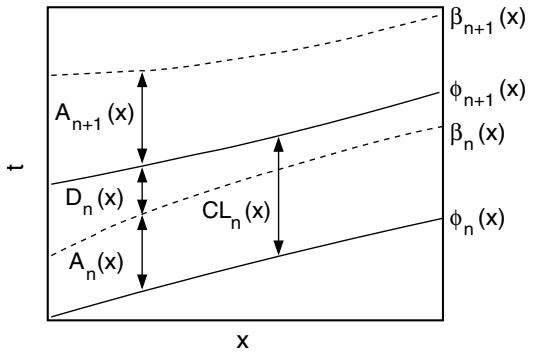


FIG. 3. *Schematic diagram of wave fronts (solid curves) and wave backs (dashed curves).*

is called the *cycle length*. Since

$$A_n(x) = \beta_n(x) - \phi_n(x) \quad \text{and} \quad D_n(x) = \phi_{n+1}(x) - \beta_n(x), \tag{5}$$

we may also express the cycle length as

$$CL_n(x) = A_n(x) + D_n(x). \tag{6}$$

From our assumption that recovery occurs via a phase wave, we may apply the restitution function locally at each x along the fiber:

$$A_n(x) = f(D_{n-1}(x)) \qquad (0 \le x \le L). \tag{7}$$

The slope of $\phi_n(x)$ is related to the speed of the nth wave front:

$$\frac{d\phi_n}{dx} = \frac{1}{c(D_{n-1}(x))} \qquad (0 < x < L). \tag{8}$$

By (6) and (7), the cycle length satisfies an algebraic condition

$$CL_n(x) = D_n(x) + f(D_{n-1}(x)), \tag{9}$$

and by (4) and (8), the cycle length also satisfies a differential equation

$$\frac{dCL_n}{dx} = \frac{1}{c(D_n(x))} - \frac{1}{c(D_{n-1}(x))}. \tag{10}$$

Combining (9) and (10), we obtain a sequence of differential equations involving only DI values:

$$\frac{d}{dx}\left[D_n(x) + f(D_{n-1}(x))\right] = G(D_n(x)) - G(D_{n-1}(x)), \tag{11}$$

where

$$G(DI) = \frac{1}{c(DI)}. \tag{12}$$

From assumption (A3) above, pacing at $x = 0$ yields the boundary condition

$$D_n(0) = \begin{cases} B_{\text{old}} - f(D_{n-1}(0)), & n \le 0, \\ B_{\text{new}} - f(D_{n-1}(0)), & n > 0, \end{cases} \tag{13}$$

while assumption (A4) yields an initial condition

$$D_0(x) \equiv D^*_{\text{old}}. \tag{14}$$

Combining (11), (13), and (14), we obtain a sequence of equations that can be solved iteratively to determine $D_n(x)$ for $n > 0$ and $0 \le x \le L$.

2.3. Derivation of the main sequence of equations. To analyze the transient behavior following the change in the pacing period, we linearize (11), (13) for $n \ge 0$. The resulting sequence of initial value problems (see (16), (17)) leads to our main sequence of equations (see (23)), which we solve exactly in the next section to obtain approximations of the functions $D_n(x)$ for $n \ge 0$.

To linearize (11), (13) for $n \ge 0$, let D^*_{new} denote the steady-state DI associated with long-term pacing with period B_{new} and let $y_n(x)$ denote our approximation of $D_n(x) - D^*_{\text{new}}$. By (14), we have

$$y_0(x) = D^*_{\text{old}} - D^*_{\text{new}}, \tag{15}$$

a constant. For $n > 0$, the linearization of (11), (13) about D^*_{new} is given by

$$\frac{d}{dx}\left[y_n(x) + \alpha y_{n-1}(x)\right] = -\lambda\left[y_n(x) - y_{n-1}(x)\right], \tag{16}$$

$$y_n(0) = -\alpha y_{n-1}(0) \qquad (n > 0), \tag{17}$$

where

$$\alpha = f'(D^*_{\text{new}}) \tag{18}$$

denotes the slope of the restitution curve evaluated at D^*_{new} and

$$-\lambda = G'(D^*_{\text{new}}). \tag{19}$$

The negative sign in (19) emphasizes that $G(DI) = 1/c(DI)$ is a monotone decreasing function, which follows from assumption (A2) in the previous subsection. We remark that

- α is dimensionless and λ has units of $(\text{length})^{-1}$;
- in the linearized dynamics, the rate of convergence to steady-state at the $x = 0$ boundary is determined by α, the Floquet multiplier [31] of the map $A_{n+1} = f(B_{\text{new}} - A_n)$.

Let us solve (16), (17) for $y_n(x)$ in terms of $y_{n-1}(x)$, resulting in a recursive sequence of equations. Rewriting (16) as

$$\frac{d}{dx}\left[y_n(x) + \alpha y_{n-1}(x)\right] = -\lambda\left[y_n(x) + \alpha y_{n-1}(x)\right] + (\alpha+1)\lambda y_{n-1}(x), \tag{20}$$

we use $e^{\lambda x}$ as an integrating factor to obtain

$$\frac{d}{dx}\left\{e^{\lambda x}\left[y_n(x) + \alpha y_{n-1}(x)\right]\right\} = (\alpha+1)\lambda e^{\lambda x} y_{n-1}(x). \tag{21}$$

Integration yields

$$e^{\lambda x}\left[y_n(x) + \alpha y_{n-1}(x)\right] = y_n(0) + \alpha y_{n-1}(0) + (\alpha+1)\lambda \int_0^x e^{\lambda s} y_{n-1}(s)\, ds. \tag{22}$$

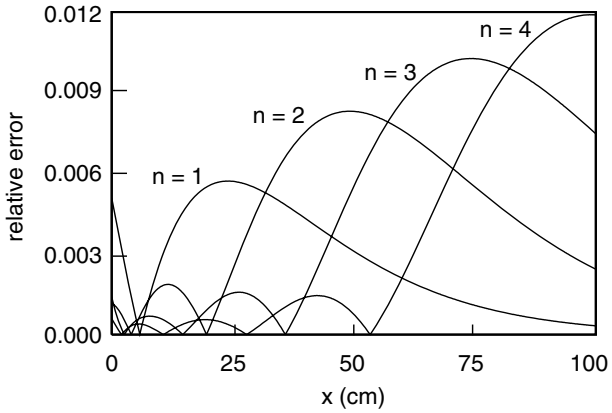


FIG. 4. *Relative error in using $y_n(x)$ to approximate $D_n(x) - D^*_{\text{new}}$ for $n = 1, \ldots, 4$.*

Finally, applying boundary condition (17) and rearranging terms, we obtain our main sequence of equations,

$$y_n(x) = -\alpha y_{n-1}(x) + (\alpha+1)\lambda \int_0^x e^{-\lambda(x-s)} y_{n-1}(s)\, ds \qquad (n \geq 1). \tag{23}$$

Numerical evidence suggests that solutions of the linearized equations (23) exhibit good quantitative agreement with solutions of the original nonlinear equations (11) and (13). Figure 4 shows the relative error $|D_n(x) - D^*_{\text{new}} - y_n(x)|/D_n(x)$ versus x for $n = 1, \ldots, 4$ after shortening the pacing period from $B_{\text{old}} = 340$ ms to $B_{\text{new}} = 320$ ms. The functions $D_n(x)$ were generated by numerical solution of (11) and (13) with f and c chosen as in (73) and (74), respectively. The functions $y_n(x)$ were generated by numerical solution of (23) with the same choices for f and c. We remark that these restitution and dispersion curves provide physiologically realistic APD values and propagation speeds for mammalian ventricular tissue [1, 13]. Note that the relative error (at least through four beats) never exceeds 0.012 even at the "un-physiological" distance of one meter from the stimulus site.

3. Estimating the rate of convergence to steady-state. Equation (23) allows us to determine y_n provided that y_{n-1} is known. In our case, $y_0(x)$ is a constant since $D_0(x) \equiv D^*_{\text{old}}$. We remark that, due to the simple form of $y_0(x)$, the recursive sequence of equations (23) can be solved exactly by successive substitutions. In doing so, it is advantageous to introduce some abstract notation (subsections 3.1 and 3.2) which helps us recognize that solutions of (23) can be expressed in terms of generalized Laguerre polynomials. Then, by applying a large-n asymptotic approximation of the Laguerre polynomials, we derive the desired estimate of $N(\eta, L)$ (see subsections 3.3 and 3.4).

3.1. Step 1: A Volterra integral operator. Motivated by (23), we define an operator $T = -\alpha I + (\alpha+1)\lambda K$ on the Banach space $(C[0, L], \|\cdot\|_\infty)$, where I denotes the identity operator and

$$(K\psi)(x) = \int_0^x e^{-\lambda(x-s)} \psi(s)\, ds. \tag{24}$$

Then clearly $y_n = Ty_{n-1} = T^n y_0$. Our goal is to estimate the rate of convergence[2] of $y_n = T^n y_0$ to 0. To do so, we exploit the fact that y_0 is a constant function; i.e.,

$$\|T^n y_0\|_\infty = |y_0| \cdot \|T^n 1\|_\infty, \tag{25}$$

where y_0 is a constant. Our main problem can now be stated as

$$\text{Given any } \eta > 0, \text{ determine } N = N(\eta, L) \text{ such that } \|T^n 1\|_\infty < \frac{\eta}{|y_0|} \text{ for all } n > N. \tag{26}$$

In the next subsection, we derive a formula for the function $(T^n 1)(x)$. Later, we will use asymptotics to learn more about the extrema of this function, using our results to estimate $\|T^n 1\|_\infty$.

3.2. Step 2: Computing powers of the operator T. Recalling that $T = -\alpha I + (\alpha+1)\lambda K$, we may apply the binomial theorem to obtain

$$(T^n\varphi)(x) = \sum_{m=0}^{n} \binom{n}{m} (-\alpha)^{n-m} (\alpha+1)^m \lambda^m (K^m\varphi)(x). \tag{27}$$

Powers of the operator K are straightforward to compute. For $m \geq 1$, we find that

$$(K^m\varphi)(x) = \int_0^x \int_0^{s_1} \cdots \int_0^{s_{m-1}} e^{-\lambda(x-s_m)} \varphi(s_m)\, ds_m\, ds_{m-1} \cdots ds_1. \tag{28}$$

Reversing the order of integration, the iterated integral (28) simplifies to a single integral

$$(K^m\varphi)(x) = \int_0^x \frac{(x-s_m)^{m-1}}{(m-1)!} e^{-\lambda(x-s_m)} \varphi(s_m)\, ds_m. \tag{29}$$

Combining (27) and (29) yields

$$(T^n\varphi)(x) = (-\alpha)^n \varphi(x) + \int_0^x \Psi_n(x-s)\, \varphi(s)\, ds \;=\; (-\alpha)^n \varphi(x) + \int_0^x \Psi_n(s)\, \varphi(x-s)\, ds, \tag{30}$$

where

$$\Psi_n(s) = e^{-\lambda s} \sum_{m=1}^{n} \binom{n}{m} (-\alpha)^{n-m} (\alpha+1)^m \lambda^m \frac{s^{m-1}}{(m-1)!}. \tag{31}$$

It follows that

$$(T^n 1)(x) \;=\; (-\alpha)^n + \int_0^x \Psi_n(s)\, ds. \tag{32}$$

The functions Ψ_n can be expressed in terms of generalized Laguerre polynomials, a well-known class of special functions which can be defined as in the following definition (see Szegő [32]).

[2]We remark that the trivial estimate $\|y_n\|_\infty = \|T^n y_0\|_\infty \leq \|T\|^n \|y_0\|_\infty$ is too weak since $\|T\|$ can exceed 1. This is especially true close to the onset of alternans (i.e., as $\alpha \to 1^-$), in which case T is a contraction only for very short fiber lengths. However, it is straightforward [10, 26] to show that the spectral radius of T is simply α. Hence, if $\alpha < 1$, then $\|T^n y_0\|_\infty$ converges to 0 as $n \to \infty$.

DEFINITION 3.1. *Let $\beta > -1$ and $n \geq 0$. Then the* generalized Laguerre polynomial $L_n^{(\beta)}(x)$ *is defined by*

$$L_n^{(\beta)}(x) = \sum_{m=0}^{n} \binom{n+\beta}{n-m} \frac{(-x)^m}{m!}. \tag{33}$$

Comparing (31) and (33) with $\beta = 1$, it is straightforward to verify that

$$\Psi_n(s) = (-\alpha)^{n-1}(\alpha+1)\lambda\, e^{-\lambda s}\, L_{n-1}^{(1)}\left(\frac{\lambda(\alpha+1)s}{\alpha}\right). \tag{34}$$

By (32) and (34), we have

$$\begin{aligned}(T^n 1)(x) &= (-\alpha)^n + (-\alpha)^{n-1}(\alpha+1)\lambda \int_0^x e^{-\lambda s}\, L_{n-1}^{(1)}\left(\frac{\lambda(\alpha+1)s}{\alpha}\right) ds \\ &= (-\alpha)^n \left(1 - \int_0^{\frac{\lambda(\alpha+1)x}{\alpha}} e^{-\frac{\alpha s}{\alpha+1}}\, L_{n-1}^{(1)}(s)\, ds\right). \end{aligned} \tag{35}$$

3.3. Step 3: Large-n asymptotic estimate of $\|T^n 1\|_\infty$. In order to estimate $\|T^n 1\|_\infty$, we must approximate the integral in (35). To do so, we will make use of an asymptotic estimate for the generalized Laguerre polynomials. However, because the estimate we will use is not uniformly valid throughout the region of integration, we will split the region of integration into two subregions.

The following asymptotic approximation as $n \to \infty$ for the generalized Laguerre polynomials appears in Szegő [32, p. 199].

THEOREM 3.2. *Let $\beta > -1$ and $n \to \infty$. Then*

$$L_n^{(\beta)}(x) = \pi^{-\frac{1}{2}}\, e^{\frac{x}{2}}\, x^{-\frac{\beta}{2}-\frac{1}{4}}\, n^{\frac{\beta}{2}-\frac{1}{4}} \left[\cos\left(\sqrt{4nx} - \frac{\beta\pi}{2} - \frac{\pi}{4}\right) + (nx)^{-\frac{1}{2}} O(1)\right]. \tag{36}$$

Moreover, given positive constants c and ω, the error term holds uniformly on the interval $cn^{-1} \leq x \leq \omega$.

Because the asymptotic approximation given by Theorem 3.2 breaks down for x small, we estimate (35) by splitting the interval of integration into two subintervals. For the boundary between the two subintervals, we use an approximation of the first root of $L_{n-1}^{(1)}$. Setting $\beta = 1$, note that the first two zeros of the approximation given by (36) occur when the argument of the cosine term is $-\pi/2$ or $\pi/2$. The $(nx)^{-1/2}$ error term influences the location of the first zero of (36) because it is not negligible for $x = O(1/n)$. Figure 5 suggests that we may approximate[3] the first root of $L_n^{(1)}(x)$ as the value of x for which the cosine term in (36) is $\pi/2$, not $-\pi/2$. That is, $x = C/n$, where

$$C = \frac{25\pi^2}{64} = 3.8553\ldots. \tag{37}$$

In what follows, we will use $C/(n-1)$ as the boundary between the two subintervals of integration.

[3] A more precise estimate of $x_{1n}^{(\beta)}$, the first root of $L_n^{(\beta)}(x)$, appears in Szegő [32]: $\lim_{n\to\infty}(n\, x_{1n}^{(\beta)}) = (j_1^{(\beta)}/2)^2 = 3.6705\ldots$, where $j_1^{(\beta)} = 3.8317\ldots$ denotes the first positive zero of the Bessel function $J_\beta(x)$.

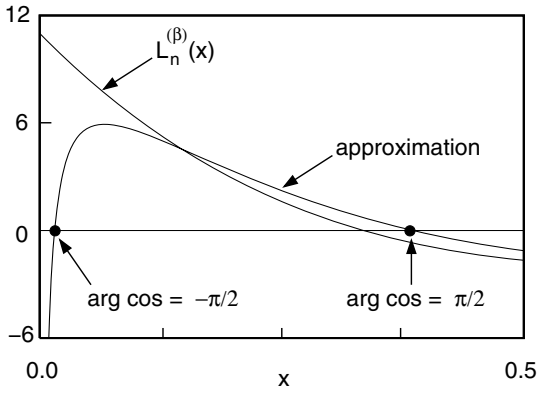


FIG. 5. *Comparison of the generalized Laguerre polynomial $L_n^{(\beta)}(x)$ with the approximation given by Theorem* 3.2 *for $n = 10$ and $\beta = 1$. As indicated in the figure, the first two roots of the approximation* (36) *occur when the argument of the cosine function is $-\pi/2$ or $\pi/2$.*

To estimate the integral in (35), we write

(38)
$$\int_0^{\frac{\lambda(\alpha+1)x}{\alpha}} e^{-\frac{\alpha s}{\alpha+1}}\, L_{n-1}^{(1)}(s)\, ds = \underbrace{\int_0^{\frac{C}{n-1}} e^{-\frac{\alpha s}{\alpha+1}}\, L_{n-1}^{(1)}(s)\, ds}_{\mathbf{I_1}} + \underbrace{\int_{\frac{C}{n-1}}^{\frac{\lambda(\alpha+1)x}{\alpha}} e^{-\frac{\alpha s}{\alpha+1}}\, L_{n-1}^{(1)}(s)\, ds.}_{\mathbf{I_2}}$$

Our expression for $(T^n 1)(x)$ now reads

$$(T^n 1)(x) = (-\alpha)^n \left(1 - \mathbf{I_1} - \mathbf{I_2}\right). \tag{39}$$

Estimating $\mathbf{I_1}$. To estimate the integral $\mathbf{I_1}$ in (38), we treat the two factors in the integrand separately. The factor $L_{n-1}^{(1)}(s)$ can be approximated by a quadratic function $q_{n-1}(s)$ by matching $L_{n-1}^{(1)}(s)$ and its derivative at $s = 0$ and using the fact that $L_{n-1}^{(1)}(C/(n-1)) \approx 0$. By algebra, we find that the polynomial

$$q_{n-1}(s) = \frac{n(n-1)^2}{C^2}\left(\frac{C}{2} - 1\right) s^2 - \frac{1}{2} n(n-1)s + n \tag{40}$$

approximates the function $L_{n-1}^{(1)}(s)$ in the interval $[0, C/(n-1)]$. Because the exponential factor in the integrand of $\mathbf{I}_1$ is $1+O(1/n)$ throughout the region of integration, we neglect this factor and compute

$$\mathbf{I_1} \approx \int_0^{\frac{C}{n-1}} q_{n-1}(s)\, ds = \frac{2C}{3} - \frac{C^2}{12} + O\left(\frac{1}{n}\right). \tag{41}$$

In what follows, we will approximate $\mathbf{I_1}$ by

$$\mathbf{I_1} \;\approx\; \frac{2C}{3} - \frac{C^2}{12} \;=\; 1.3315\ldots\,. \tag{42}$$

Estimating $\mathbf{I_2}$. Integral $\mathbf{I_2}$ in (38) can be approximated by applying Theorem 3.2 with $\beta = 1$. Setting

$$\gamma = \frac{1-\alpha}{2(1+\alpha)}, \tag{43}$$

we have

$$\mathbf{I_2} \approx \pi^{-\frac{1}{2}}(n-1)^{\frac{1}{4}} \int_{\frac{C}{n-1}}^{\frac{\lambda(\alpha+1)x}{\alpha}} s^{-\frac{3}{4}}\, e^{\gamma s}\, \cos\left(\sqrt{4(n-1)s} - \frac{3\pi}{4}\right) ds. \tag{44}$$

Upon substituting $s \mapsto \sqrt{s}$, (44) becomes

$$\mathbf{I_2} \approx 2\pi^{-\frac{1}{2}}(n-1)^{\frac{1}{4}} \int_{\sqrt{\frac{C}{n-1}}}^{\sqrt{\frac{\lambda(\alpha+1)x}{\alpha}}} s^{-\frac{1}{2}}\, e^{\gamma s^2}\, \cos\left(\sqrt{4(n-1)}\cdot s - \frac{3\pi}{4}\right) ds. \tag{45}$$

Integrating (45) by parts,

(46)

$$\mathbf{I}_2 \approx \left.\frac{s^{-\frac{1}{2}}\, e^{\gamma s^2} \sin\left(\sqrt{4(n-1)}\cdot s - \frac{3\pi}{4}\right)}{\sqrt{\pi}(n-1)^{\frac{1}{4}}}\right|_{\sqrt{\frac{C}{n-1}}}^{\sqrt{\frac{\lambda(\alpha+1)x}{\alpha}}}$$
$$- \frac{1}{\sqrt{\pi}(n-1)^{\frac{1}{4}}} \int_{\sqrt{\frac{C}{n-1}}}^{\sqrt{\frac{\lambda(\alpha+1)x}{\alpha}}} \left(-\frac{1}{2}s^{-\frac{3}{2}} + 2\gamma s^{\frac{1}{2}}\right) e^{\gamma s^2} \sin\left(\sqrt{4(n-1)}\cdot s - \frac{3\pi}{4}\right) ds.$$

At first, the two terms in (46) appear to be of the same order, namely $O(n^{-1/4})$. However, for large n the rapid oscillation of the integrand of the volume term leads to cancellation, and hence the volume term is small relative to the boundary term. (This can also be seen by integrating by parts a second time.) Thus, we approximate $\mathbf{I_2}$ by retaining only the boundary terms in (46):

$$\mathbf{I_2} \approx \frac{u(x)}{(n-1)^{\frac{1}{4}}} \sin\left(\sqrt{\frac{4\lambda(\alpha+1)(n-1)x}{\alpha}} - \frac{3\pi}{4}\right)$$
$$- \left(\frac{1}{C\pi^2}\right)^{\frac{1}{4}} e^{\frac{C(1-\alpha)}{2(1+\alpha)(n-1)}} \sin\left(\sqrt{4C} - \frac{3\pi}{4}\right), \tag{47}$$

where

$$u(x) = \left(\frac{\alpha}{\pi^2\lambda(\alpha+1)x}\right)^{\frac{1}{4}} e^{\frac{\lambda(1-\alpha)x}{2\alpha}}. \tag{48}$$

The first term in (47) is bounded by

$$\frac{u(x)}{(n-1)^{\frac{1}{4}}}, \tag{49}$$

and, for large n, the second term of (47) is approximately

$$\left(\frac{1}{C\pi^2}\right)^{\frac{1}{4}} \sin\left(\sqrt{4C} - \frac{3\pi}{4}\right) = \left(\frac{1}{C\pi^2}\right)^{\frac{1}{4}} = 0.4026\ldots. \tag{50}$$

By (39), (42), (49), and (50), we obtain the following approximate upper bound for $|(T^n 1)(x)|$:

$$|(T^n 1)(x)| \lesssim \alpha^n \left|1 - 1.3315 + 0.4026 + \frac{u(x)}{(n-1)^{\frac{1}{4}}}\right|$$
$$= \alpha^n \left(0.0711 + \frac{u(x)}{(n-1)^{\frac{1}{4}}}\right) \quad (n \to \infty). \tag{51}$$

3.4. Step 4: Estimates for $N(\eta, L)$. Referring to the definition (48) of $u(x)$, note that the $x^{-1/4}$ factor is dominant for small x while the exponential factor is dominant for large x. Hence, we expect $u(x)$ to have a single extremum—a global minimum. Letting L^* denote the x value at which the global minimum of $u(x)$ is attained (see (53) below), we are led to consider two cases when estimating $N(\eta, L)$: the case $L < L^*$ and the case $L > L^*$. In the former case, we shall demonstrate that $|(T^n 1)(x)|$ is always maximal at $x = 0$, indicating that the convergence to steady-state is slowest at the pacing site. In the latter case, $|(T^n 1)(x)|$ is maximal either at $x = 0$ or near $x = L$.

To determine L^*, we differentiate $u(x)$ with respect to x:

$$u'(x) = \left(\frac{\alpha}{\pi^2\lambda(\alpha+1)}\right)^{\frac{1}{4}} e^{\frac{\lambda(1-\alpha)x}{2\alpha}} \left(-\frac{1}{4}x^{-\frac{5}{4}} + \frac{\lambda(1-\alpha)}{2\alpha}x^{-\frac{1}{4}}\right). \tag{52}$$

The unique x value for which $u'(x)$ has a root is

$$L^* = \frac{\alpha}{2\lambda(1-\alpha)}. \tag{53}$$

With the above considerations in mind, we now derive our estimates for $N(\eta, L)$.

Case 1. $L < L^*$. From our preceding remarks, we see that the function $(T^n 1)(x)$ exhibits damped oscillatory behavior for $x < L^*$. Hence, $|(T^n 1)(x)|$ is maximal either at $x = 0$ or at the first local extremum of $(T^n 1)(x)$. In fact, we shall see that $|(T^n 1)(x)|$ is always maximal at $x = 0$.

PROPOSITION 3.3. *Let $x^{(1)}_{1(n-1)}$ denote the first root of $L^{(1)}_{n-1}(x)$. Then the first local extremum of $(T^n 1)(x)$ occurs at*

$$x_{\text{ext}} = \frac{\alpha}{\lambda(\alpha+1)} x^{(1)}_{1(n-1)} \approx \frac{\alpha}{\lambda(\alpha+1)} \cdot \frac{C}{n-1}. \tag{54}$$

Proof. Differentiating (32) with respect to x, we obtain

$$(T^n 1)'(x) \;=\; \Psi_n(x) \;=\; (-\alpha)^{n-1}(\alpha+1)\lambda\, e^{-\lambda x}\, L^{(1)}_{n-1}\left(\frac{\lambda(\alpha+1)x}{\alpha}\right). \tag{55}$$

The only factor in (55) that can change sign as x varies is $L^{(1)}_{n-1}$. Hence, the first sign change of $\Psi_n(x)$ occurs when the $L^{(1)}_{n-1}$ factor has its first root, which occurs at $x = x_{\text{ext}}$. This x value corresponds to the first local extremum of $(T^n 1)(x)$. ∎

Equation (35) gives the exact value of $|(T^n 1)(x)|$ at the $x = 0$ boundary, namely

$$|(T^n 1)(0)| = \alpha^n. \tag{56}$$

To estimate the value of $(T^n 1)(x)$ at its first local extremum, we refer to (51) and (54). By (54), we obtain the approximation

$$\frac{u(x_{\text{ext}})}{(n-1)^{\frac{1}{4}}} \to \left(\frac{1}{C\pi^2}\right)^{\frac{1}{4}} \;\approx\; 0.4026 \qquad (n \to \infty), \tag{57}$$

and (51) yields

$$|(T^n 1)(x_{\text{ext}})| \lesssim (0.0711 + 0.4026)\,\alpha^n \qquad (n \to \infty). \tag{58}$$

Note that the coefficient of α^n in (58) is less than 1. Therefore, if $L < L^*$, we conclude that the convergence to steady-state is slowest at the $x = 0$ boundary. That is,

$$\|T^n 1\|_\infty = |(T^n 1)(0)| = \alpha^n. \tag{59}$$

With η and y_0 as in (26), we can now estimate the number of beats required to reach approximate steady-state:

$$N_1 = N_1(\eta) = \frac{\ln(\eta) - \ln|y_0|}{\ln(\alpha)}. \tag{60}$$

The subscript on N emphasizes that we presently consider the first case, $L < L^*$. Note that N_1 does not depend upon L, and the convergence to steady-state is slowest at the pacing site, $x = 0$.

Case 2. $L > L^*$. In this case, the function $u(x)$ may achieve its maximum at the right boundary of the interval $[C/(n-1), L]$. By (51),

$$|(T^n 1)(L)| \lesssim \alpha^n \left(0.0711 + \frac{u(L)}{(n-1)^{\frac{1}{4}}}\right) \qquad (n \to \infty). \tag{61}$$

We wish to determine $N(\eta, L)$ such that

$$\alpha^n \left(0.0711 + \frac{u(L)}{(n-1)^{\frac{1}{4}}}\right) < \frac{\eta}{|y_0|} \qquad (\text{for all } n > N). \tag{62}$$

Taking logarithms and dividing through by $\ln\alpha$, we obtain

$$n + \frac{\ln\left(0.0711 + u(L)(n-1)^{-\frac{1}{4}}\right)}{\ln\alpha} > N_1, \tag{63}$$

where N_1 is given by (60) above. Motivated by inequality (63), we define the function

$$g(n) = n + \frac{\ln\left(0.0711 + u(L)(n-1)^{-\frac{1}{4}}\right)}{\ln\alpha} - N_1 \tag{64}$$

and estimate the value of n for which $g(n)$ has a root. Because N_1 is large if $\eta/|y_0| \ll 1$, we use it as an initial guess and calculate

$$g(N_1) = \frac{\ln\left(0.0711 + u(L)(N_1-1)^{-\frac{1}{4}}\right)}{\ln\alpha}. \tag{65}$$

An improved estimate for the root of $g(n)$ is then given by

$$N_1 - g(N_1) \;=\; N_1 - \frac{\ln\left(0.0711 + u(L)(N_1-1)^{-\frac{1}{4}}\right)}{\ln\alpha}. \tag{66}$$

Comparing (66) with the expression (60), we obtain the desired estimate for $N(\eta, L)$ by taking the maximum of these two expressions:

$$N(\eta, L) = N_1 - \frac{\left[\ln\left(0.0711 + u(L)(N_1-1)^{-\frac{1}{4}}\right)\right]_+}{\ln\alpha}, \tag{67}$$

where

$$[a]_+ = \begin{cases} a, & a > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{68}$$

In summary, the above approximations demonstrate that $|y_n(x)| = |y_0| \cdot |(T^n 1)(x)|$ is maximal near the boundaries of the interval $[0, L]$. The location of the maximum of $|(T^n 1)(x)|$ depends in part on whether the fiber length L exceeds a critical value L^*. The function $(T^n 1)(x)$ exhibits damped oscillatory behavior for $x < L^*$. Hence, if $L < L^*$, we conclude that $|(T^n 1)(x)|$ is maximal either at $x = 0$ or at the first local extremum of the function $(T^n 1)(x)$. Our computations rule out the latter case, and we find that $|(T^n 1)(x)|$ is always maximal at $x = 0$ if $L < L^*$. For $x > L^*$, the function $(T^n 1)(x)$ exhibits oscillations of growing amplitude and, in the worst case scenario, $|(T^n 1)(x)|$ is maximal at $x = L$. Our estimate of $N(\eta, L)$ is given by either (60) or (67) depending on whether L exceeds L^*.

A discussion of the physiological interpretation of the above results is provided in the next section.

4. Discussion and conclusions. We have described how a paced cardiac fiber responds when the pacing period is suddenly changed from B_{old} to B_{new}, providing the first analysis of the transient behavior resulting from such a pacing protocol. We estimated the number of beats $N(\eta, L)$ required for the spatial variation in DI to be small in the sense that

$$|D_n(x) - D^*_{\text{new}}| < \eta \quad \text{for all } x \in [0, L], \quad \text{for all } n \geq N.$$

The estimate is given by either (60) or (67) depending on whether the fiber length exceeds the critical value L^* defined by (53). According to our approximations, for $L < L^*$, the convergence to steady-state is slowest at the pacing site and the rate of convergence is determined by the slope α of the restitution curve evaluated at D^*_{new}.

To test these predictions, we performed numerical simulations of (11), (13), and (14) for a range of parameter values. In particular, we used the restitution and dispersion curves in the appendix (see (73), (74)), varying the parameters (69) within physiologically reasonable regimes for the mammalian ventricular action potential [1, 13] (e.g., peak conduction velocity of 60 ± 20 cm/sec). We also repeated the numerical simulations using simple exponential restitution and dispersion curves (see (3)), again for a range of parameter values. As expected, for short fiber lengths ($L < L^*$) the convergence is always slowest at the pacing site, and the transient lasts much longer as $\alpha \to 1^-$. Moreover, the estimate of $N(\eta, L)$ given by (60) typically provides a very accurate estimate (within several beats) of the actual number of beats required to achieve approximate steady-state. Not surprisingly, the estimate given by (60) breaks down if $|B_{\text{old}} - B_{\text{new}}|$ is large (on the order of hundreds of milliseconds) or if α is very close to 1.

For long fibers ($L > L^*$), the spatial DI profiles generated by numerical simulations are qualitatively similar to those shown in Figure 2(a)—in particular, the convergence $D_n(x) \to D^*_{\text{new}}$ is slowest at the far end of the fiber. In this case, (67) typically provides an accurate estimate of $N(\eta, L)$, with the same notable exceptions as in the case of short fibers (see preceding paragraph).

We remark that L^* is often so large that fibers of length $L > L^*$ are unrealistically long—on the order of tens of centimeters or even meters. For example, consider the restitution and dispersion curves in the appendix with parameter values given by (69).

Using $B_{\text{new}} = 266$ ms, one computes that $\alpha = 0.900$ and $\lambda = 0.127$ cm^{-1}, which by (53) yields a critical fiber length L^* in excess of 35 cm. Equation (53) suggests that the critical length blows up as we approach the bifurcation to alternans: $L^* \to \infty$ as the slope $\alpha = f'(D^*_{\text{new}}) \to 1^-$, a prediction consistent with all of our numerical simulations. Therefore, N is especially unlikely to exhibit any length dependence if we are pacing in a regime close to the onset of alternans.

In closing, we remark that extending our results to the case of alternans may be quite challenging. If $\alpha > 1$, the operator T no longer has the property that $\|T^n\varphi\|_\infty \to 0$ as $n \to \infty$ for all $\varphi \in C[0, L]$. Thus, a similar analysis of the alternans regime would require a substantially different approach, which we hope to provide in a future study.

Appendix. Sample restitution and dispersion curves. For the purpose of numerical simulation, we provide sample formulas for restitution and dispersion curves (73) and (74) for a particular choice of parameters. Using asymptotics, such formulas can be derived [4, 20] from the equations of an idealized ionic model [17, 20]; we omit the details here. Below, we measure DI values in ms and we use the following parameters:

$$\tau_{\text{in}} = 0.1\,\text{ms}, \quad \tau_{\text{out}} = 2.4\,\text{ms}, \quad \tau_{\text{open}} = 130\,\text{ms}, \quad \tau_{\text{close}} = 150\,\text{ms}, \quad \kappa = 10^{-3}\,\frac{\text{cm}^2}{\text{ms}}. \tag{69}$$

Let

$$h(DI) = 1 - (1 - h_{\min})\, e^{-DI/\tau_{\text{open}}} \tag{70}$$

and

$$V_\pm(DI) = \frac{1}{2}\left(1 \pm \sqrt{1 - \frac{h_{\min}}{h(DI)}}\right), \tag{71}$$

where

$$h_{\min} = \frac{4\tau_{\text{in}}}{\tau_{\text{out}}}. \tag{72}$$

Note that $h(DI)$ and $V_\pm(DI)$ are dimensionless. In all numerical simulations, we use the restitution function

$$f(DI) = \tau_{\text{close}} \ln\left[\frac{h(DI)}{h_{\min}}\right] \tag{73}$$

and the dispersion function

$$c(DI) = \max\left\{\left[\frac{1}{2}V_+(DI) - V_-(DI)\right]\sqrt{\frac{2\,\kappa\,h(DI)}{\tau_{\text{in}}}},\ 0\right\}. \tag{74}$$

Note that $f(DI)$ has units of ms and $c(DI)$ has units of cm/ms.

REFERENCES

[1] I. Banville and R. A. Gray, *Effect of action potential duration and conduction velocity restitution and their spatial dispersion on alternans and the stability of arrhythmias*, J. Cardiovasc. Electrophysiol., 13 (2002), pp. 1141–1149.

[2] M. R. Boyett and B. R. Jewell, *A study of the factors responsible for rate-dependent shortening of the action potential in mammalian ventricular muscle*, J. Physiol., 285 (1978), pp. 359–380.

[3] J. W. Cain, *Issues in the One-Dimensional Dynamics of a Paced Cardiac Fiber*, Ph.D. dissertation, Duke University, Durham, NC, 2005.

[4] J. W. Cain, E. G. Tolkacheva, D. G. Schaeffer, and D. J. Gauthier, *Rate-dependent propagation of cardiac action potentials in a one-dimensional fiber*, Phys. Rev. E (3), 70 (2004), pp. 061906.

[5] J. Cao, Z. Qu, Y. Kim, T. Wu, A. Garfinkel, J. N. Weiss, H. S. Karagueuzian, and P. Chen, *Spatiotemporal heterogeneity in the induction of ventricular fibrillation by rapid pacing: Importance of cardiac restitution properties*, Circ. Res., 84 (1999), pp. 1318–1331.

[6] E. M. Cherry and F. H. Fenton, *Suppression of alternans and conduction blocks despite steep APD restitution: Electrotonic, memory, and conduction velocity restitution effects*, Am. J. Physiol., 286 (2004), pp. H2332–H2341.

[7] M. Courtemanche, L. Glass, and J. P. Keener, *Instabilities of a propagating pulse in a ring of excitable media*, Phys. Rev. Lett., 70 (1993), pp. 2182–2184.

[8] M. Courtemanche, J. P. Keener, and L. Glass, *A delay equation representation of pulse circulation on a ring in excitable media*, SIAM J. Appl. Math., 56 (1996), pp. 119–142.

[9] E. Cytrynbaum and J. P. Keener, *Stability conditions for the traveling pulse: Modifying the restitution hypothesis*, Chaos, 12 (2002), pp. 788–799.

[10] N. Dunford and J. T. Schwartz, *Linear Operators. Part* I: *General Theory*, Interscience, New York, 1958.

[11] B. Echebarria and A. Karma, *Instability and spatiotemporal dynamics of alternans in paced cardiac tissue*, Phys. Rev. Lett., 88 (2002), pp. 208101.

[12] J. J. Fox, R. F. Gilmour, Jr., and E. Bodenschatz, *Conduction block in one dimensional heart fibers*, Phys. Rev. Lett., 89 (2002), pp. 198101–198104.

[13] L. H. Frame and M. B. Simson, *Oscillations of conduction, action potential duration, and refractoriness*, Circulation, 78 (1988), pp. 1277–1287.

[14] M. R. Guevara, G. Ward, A. Shrier, and L. Glass, *Electrical alternans and period doubling bifurcations*, in Proceedings of the 11th Computers in Cardiology Conference, IEEE Computer Society, Los Angeles, 1984, pp. 167–170.

[15] G. M. Hall, S. Bahar, and D. J. Gauthier, *Prevalence of rate-dependent behaviors in cardiac muscle*, Phys. Rev. Lett., 82 (1999), pp. 2995–2998.

[16] H. Ito and L. Glass, *Theory of reentrant excitation in a ring of cardiac tissue*, Phys. D, 56 (1992), pp. 84–106.

[17] A. Karma, *Spiral breakup in model equations of action potential propagation in cardiac tissue*, Phys. Rev. Lett., 71 (1993), pp. 1103–1107.

[18] J. P. Keener, *Waves in excitable media*, SIAM J. Appl. Math., 39 (1980), pp. 528–548.

[19] J. P. Keener and J. Sneyd, *Mathematical Physiology*, Springer-Verlag, New York, 1998.

[20] C. C. Mitchell and D. G. Schaeffer, *A two-current model for the dynamics of cardiac membrane*, Bull. Math. Bio., 65 (2003), pp. 767–793.

[21] G. R. Mines, *On dynamic equilibrium in the heart*, J. Physiol. (London), 46 (1913), pp. 349–383.

[22] J. C. Neu, R. S. Preissig, Jr., and W. Krassowska, *Initiation of propagation in a one-dimensional excitable medium*, Phys. D, 102 (1997), pp. 285–299.

[23] J. B. Nolasco and R. W. Dahlen, *A graphic method for the study of alternation in cardiac action potentials*, J. Appl. Physiol., 25 (1968), pp. 191–196.

[24] J. M. Pastore, S. D. Girouard, K. R. Laurita, F. G. Akar, and D. S. Rosenbaum, *Mechanism linking T-wave alternans to the genesis of cardiac fibrillation*, Circulation, 99 (1999), pp. 1499–1507.

[25] R. Plonsey and R. C. Barr, *Bioelectricity: A Quantitative Approach*, Plenum Press, New York, 1988.

[26] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* I: *Functional Analysis*, Academic Press, San Diego, 1980.

[27] J. Rinzel and K. Maginu, *Kinematic analysis of wave pattern formation in excitable media*, in Nonequilibrium Dynamics in Chemical Systems, A. Pacault and C. Vidal, eds., Springer-Verlag, Berlin, 1984, pp. 107–113.

[28] D. S. Rosenbaum, P. Albrecht, and R. J. Cohen, *Predicting sudden cardiac death from T wave alternans of the surface electrocardiogram: Promise and pitfalls*, J. Cardiovasc. Electrophysiol., 7 (1996), pp. 1095–1111.
[29] D. S. Rosenbaum, L. E. Jackson, J. M. Smith, H. Garan, J. N. Ruskin, and R. J. Cohen, *Electrical alternans and vulnerability to ventricular arrhythmias*, New Engl. J. Medicine, 330 (1994), pp. 235–241.
[30] H. Sedaghat, C. M. Kent, and M. A. Wood, *Criteria for the convergence, oscillation, and bistability of pulse circulation in a ring of excitable media*, SIAM J. Appl. Math., 66 (2005), pp. 573–590.
[31] S. Strogatz, *Nonlinear Dynamics and Chaos*, Perseus, Cambridge, MA, 1994.
[32] G. Szegő, *Orthogonal Polynomials*, 4th ed., AMS, Providence, RI, 1975.
[33] J. J. Tyson and J. P. Keener, *Singular perturbation theory of traveling waves in excitable media (a review)*, Phys. D, 32 (1988), pp. 327–361.
[34] M. A. Watanabe, F. H. Fenton, S. J. Evans, H. M. Hastings, and A. Karma, *Mechanisms for discordant alternans*, J. Cardiovasc. Electrophys., 12 (2001), pp. 196–206.
[35] A. R. Yehia, D. Jeandupeux, F. Alonso, and M. R. Guevara, *Hysteresis and bistability in the direct transition from* 1:1 *to* 2:1 *rhythm in periodically driven single ventricular cells*, Chaos, 9 (1999), pp. 916–931.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1793–1824

CYLINDER BUCKLING: THE MOUNTAIN PASS AS AN ORGANIZING CENTER*

JIŘÍ HORÁK†, GABRIEL J. LORD‡, AND MARK A. PELETIER§

Abstract. We revisit the classical problem of the buckling of a long thin axially compressed cylindrical shell. By examining the energy landscape of the perfect cylinder, we deduce an estimate of the sensitivity of the shell to imperfections. Key to obtaining this estimate is the existence of a mountain pass point for the system. We prove the existence on bounded domains of such solutions for almost all loads and then numerically compute example mountain pass solutions. Numerically the mountain pass solution with lowest energy has the form of a single dimple. We interpret these results and validate the lower bound against some experimental results available in the literature.

Key words. imperfection sensitivity, subcritical bifurcation, single dimple

AMS subject classifications. 35J50, 35J35, 35J60, 35G30, 35Q72

DOI. 10.1137/050635778

1. Introduction.

1.1. Buckling of cylinders under axial loading.

A classical problem in structural engineering is the prediction of the load-carrying capacity of an axially loaded cylinder. In addition to being a commonly used structural element, the axially loaded cylinder is the archetype of unstable, imperfection-sensitive buckling, and this has led to a large body of theoretical and experimental research.

In the decades before and after the Second World War, a central problem was understanding the large discrepancy between theoretical predictions and experimental observations, as shown in Figure 1.1. A variety of different explanations has been put forward, but with the experimental work of Tennyson [30] and the theoretical work of Almroth [1] it became clear that this discrepancy is mostly due to imperfections in loading conditions and in the shape of the specimens. Further experimental and theoretical work by many others has confirmed this conclusion [14, 33, 36].

For near-perfect cylinders, the linear and weakly nonlinear theories (see section 1.2) adequately describe the experimental buckling load[1] and the deformation just before failure (see, e.g., [4]). Cylinders used in practical applications, however, are far too imperfect, and thus the weakly nonlinear theory does not apply. From a practical point of view, the problem of predicting the failure load is still open.

There is good reason to believe that it will never be possible to accurately predict failure loads for cylinders that are used in practice. For simple materials, such as metals, it is believed that current numerical methods can describe the local material behavior with enough accuracy that correct prediction of the complete behavior of the cylinder—including its failure—is feasible. This could be achieved provided the geometrical and material imperfections, as well as the loading conditions, are determined

*Received by the editors July 12, 2005; accepted for publication (in revised form) April 6, 2006; published electronically August 22, 2006.
http://www.siam.org/journals/siap/66-5/63577.html
†Universität Köln, Köln, Germany (j.horak@math.uni-koeln.de).
‡Heriot-Watt University, Edinburgh, UK (g.j.lord@ma.hw.ac.uk).
§Technische Universiteit Eindhoven, Eindhoven, The Netherlands (m.a.peletier@tue.nl).
[1]In this paper the terms "experimental buckling load" and "failure load" are used interchangeably.

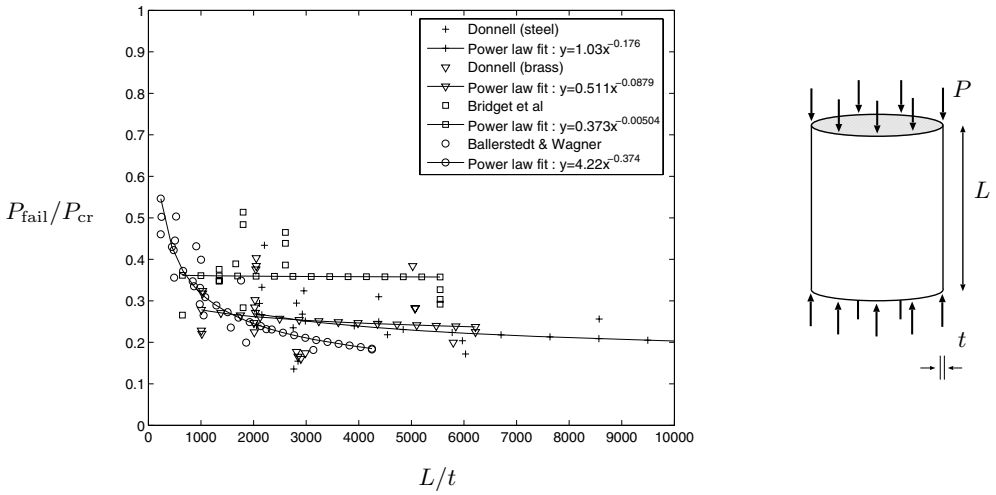


FIG. 1.1. *Experimental data from various research groups, all representing failure loads of axially loaded cylinders. The horizontal axis is the ratio of the cylinder length and the wall thickness; the vertical axis is the ratio of the failure load and the theoretical critical load as predicted for perfect cylinders. Note that all tested cylinders fail at loads significantly lower than that predicted by theory; the latter would correspond to failure load* $P_{\rm fail}/P_{\rm cr} = 1$*, and in some cases failure occurred at less than one-fifth of this value. Power-law fitting lines are added to emphasize the dependence of the failure load on the geometry. The data are from* [10, 7, 5].

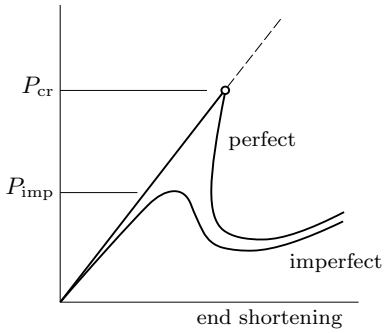


FIG. 1.2. *Illustration of perfect and imperfect bifurcation curves. For the unperturbed system, the undeformed state (the straight line) is always an equilibrium; this state loses stability at a bifurcation point at load* $P_{\rm cr}$*, and a curve of nonzero equilibria branches off. Perturbing the system generically converts the sharp bifurcation into a smooth transition. In the case of the cylinder, the postbuckling path is strongly unstable, and the perturbed path therefore has a lower limit load of* $P_{\rm imp}$.

in sufficient detail. The difficulty lies in the qualifier "in sufficient detail," since an extremely accurate measurement of the geometric imperfections would be necessary [4], and in the design phase both the loading conditions and the geometric and material imperfections in the finalized product are known only in vague terms. Therefore, in recent decades the attention of theoretical research has turned to characterizing the failure load in weaker ways, preferably in the form of a lower (safe) bound.

1.2. Characterizing sensitivity to imperfections. Viewed as a bifurcation problem, the buckling of the cylinder is a subcritical symmetry-breaking pitchfork bifurcation (Figure 1.2). Generically, imperfections in the structure eliminate the

bifurcation and round off the branch of solutions,[2] resulting in a turning point at a load P_{imp} strictly below the critical (bifurcation) load P_{cr} of the perfect structure. In an experiment in which the load is slowly increased, the system will fail (i.e., make a large jump in state space) at load P_{imp}.

Again, if the imperfections in geometry and loading are fully known, then calculation of P_{imp} is a practical rather than a theoretical problem, and we do not address this problem here. For the more difficult question of characterizing P_{imp} under incomplete information, various strategies have been proposed. A classical line of thought originates with Koiter [22], in which the imperfections are chosen a priori within certain finite-dimensional sets. Common choices are the sets spanned by the eigenvector at the bifurcation point of the perfect structure or by the eigenvectors associated with the first n bifurcations. This approach might be termed weakly nonlinear, as it is based on an expansion of the energy close to the bifurcation point, in the directions suggested by the bifurcation point itself. It gives predictions that are correct if the imperfections are very small—much smaller than those encountered in practice.

Since the a priori choice of imperfections is a weak point of this method, a natural step is to optimize over all possible perturbations. Deml and Wunderlich pioneered this approach, in which a numerical algorithm is used to find a "worst geometric imperfection" [9]. This "worst imperfection" is defined as that imperfection that produces a turning point of minimal load. Some constraint on the magnitude of allowable perturbations is necessary, of course, to prevent the running of a steam roller over the cylinder being interpreted as an admissible imperfection. The authors of [9] and [35] first suggest constraining the L^∞-norm of the perturbation displacement, but they immediately replace the L^∞-norm with an L^p-norm for computational convenience.

This method has an interesting aspect that is often glossed over in the engineering literature. By definition, the failure load obtained by this method is a lower bound for the failure load of all systems that have perturbations of lesser or equal magnitude. The measure of magnitude is defined by the choice of constraint. Therefore the choice of constraint on the imperfections is critical, since it implicitly defines a class of imperfections that produce either the same failure load or a higher one.

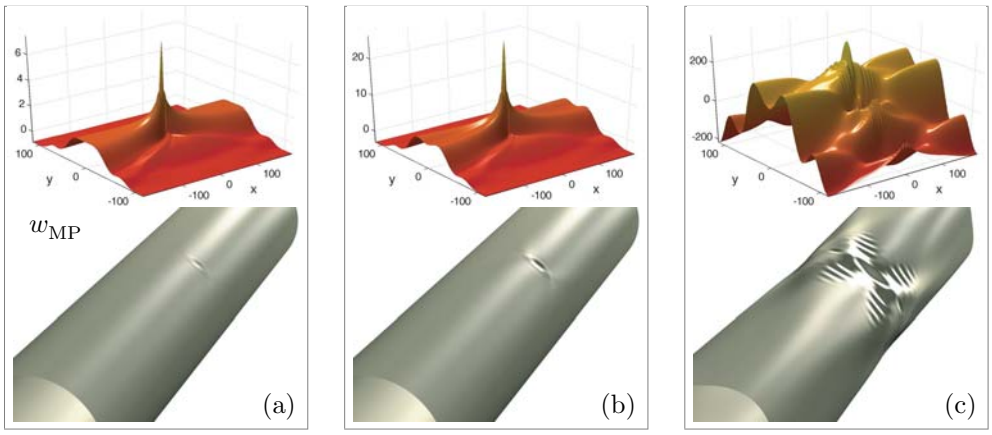
1.3. Main results. In this paper we follow a related, but distinct, line of reasoning. Instead of studying the actual behavior of imperfect cylinders, we deduce an estimate of the sensitivity to imperfections from the energy landscape of the perfect cylinder. The final result is a lower bound on the failure load similar to the above lower bound, and the approach gives additional insight into the problem.

The key result is the existence of a *mountain pass point*, an equilibrium state that is straddled between two valleys in the energy landscape; one valley surrounds the unbuckled state, and the other contains many buckled, large-deformation states.

This mountain pass point has a number of interesting properties, as follows:

1. It has the appearance of a *single-dimple solution*, a small buckle in the form of a single dent (see Figure 1.3(a)). Single-dimple deformations have appeared in engineering literature in a number of different ways (see section 6), but a theoretical understanding of this phenomenon is still lacking. Localization (concentration) of deformation is commonly known to appear in extended structures [21], and in the cylinder localization in the axial direction has been studied theoretically and numerically [20, 24, 25]. Whether localization

[2] Koiter actually used this elimination of a bifurcation point as a *definition* of "perfect system" and "perturbed system" [22].



(a) $\dots V(\lambda) = F_\lambda(w_{\mathrm{MP}}) \approx 4.84$
(b) $\dots F_\lambda(w) \approx -15.57$
(c) $\dots F_\lambda(w) \approx -5.5 \cdot 10^4$

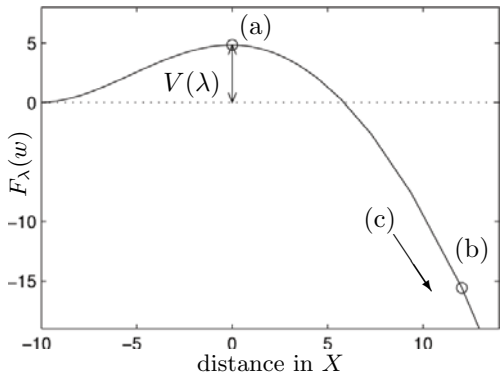


FIG. 1.3. *Part* (a) *shows a numerical computation of a solution* w_{MP} *that is a mountain pass point of the energy* F_λ *for a load* $\lambda = 1.5$. *We show the graph of the displacement* $w_{\mathrm{MP}}(x, y)$ *as a function of* (x, y) *as well as its rendering on a cylinder. At* w_{MP} *there exist two directions in the state space* X *in which the energy* F_λ *decreases. By perturbing* w_{MP} *in these directions and following a gradient flow of* F_λ, *we move away from* w_{MP}. *In one direction the dimple shrinks and disappears (not shown) and in the opposite its direction grows in amplitude and extent* ((b) *and* (c)).

is possible in the tangential direction has been an open problem for some time; it is interesting that our simulations for the perfect structure show solutions that are localized in both axial and tangential directions.

2. Like all mountain pass points, this single-dimple solution is unstable, in the sense that there are directions in state space in which the energy decreases. In one direction the dimple roughly shrinks and disappears, and in the opposite direction it grows and multiplies (Figures 1.3(b)–(c)). It is remarkable, however, that our numerical results indicate that the single-dimple solution has an alternative characterization as a *constrained global minimizer* (a global minimizer of the strain energy under prescribed end shortening).
3. The equations can be rescaled so that the only remaining parameters are the load level and the domain. The geometry of the mountain pass solution we calculate even appears to be independent of the domain size.

This mountain pass point is central in an estimate of the sensitivity to perturbations. For the system to escape from the neighborhood of the unbuckled state,

it must possess at least the energy associated with this mountain pass point. The mountain pass energy level is therefore an indication of the degree of stability of the unbuckled state. Implicitly it defines a class of perturbations for which the unbuckled state is stable. This approach is related to the "perturbation-energy" approach first suggested by Kröplin, Dinkler, and Hillmann [23], Duddeck et al. [11], and Wagenhuber and Duddeck [32], but differs in some essential points; see the discussion in section 5.

1.4. Methods. We use both analytical and numerical methods. In section 2 we introduce the von Kármán–Donnell equations, which form the basis of this paper, and rescale them in an appropriate manner. In section 3 we present the functional setting that we use, show that the energy functional has the geometry associated with a mountain pass, and prove the existence of mountain pass points (Lemma 3.6). There are certain interesting technical issues. By their localized nature, single-dimple solutions are most naturally defined on an unbounded domain; however, we are only able to prove existence of mountain pass points on bounded domains, and consequently we work on finite domains that become large in the limit of thin shells. Similarly, the noncoercive nature of the energy functional implies that we prove the existence of mountain pass points for almost all load levels (see Lemma 3.5).

In section 4 we turn to numerical investigation. We use a variety of different algorithms to find solutions of the discretized von Kármán–Donnell equations. With a discrete mountain pass algorithm we find solutions that are, by construction, mountain pass points. The solution of Figure 1.3(a) was found in this manner. With a constrained gradient flow we also find local minima of the strain energy under prescribed end shortening. Some of these solutions appear to coincide with those found by the discrete mountain pass algorithm, and the mountain pass solutions are stable under this gradient flow. These observations lead us to conjecture that the global mountain pass solution is also a global constrained minimizer of the strain energy. By a constrained version of the discrete mountain pass algorithm, we also find critical points of higher Morse index.

Section 5 is devoted to an interpretation of these results in the context of imperfection sensitivity, as mentioned above, and in section 6 we wrap up with the main conclusions.

2. The von Kármán–Donnell equations. We consider a cylindrical shell of radius R, thickness t, Young's modulus E, and Poisson's ratio ν that is subject to a compressive axial force P. In Appendix B we derive the dimensionless von Kármán–Donnell equations

$$\varepsilon^2\Delta^2\overline{w} + \bar{\lambda}\overline{w}_{xx} - \overline{\phi}_{xx} - 2[\overline{w},\overline{\phi}] = 0, \tag{2.1}$$

$$\Delta^2\overline{\phi} + [\overline{w},\overline{w}] + \overline{w}_{xx} = 0, \tag{2.2}$$

where subscripts x and y denote differentiation with respect to the spatial variables, the Laplacian Δ is given by $\Delta u = u_{xx} + u_{yy}$, and the bracket $[\cdot,\cdot]$ is defined as

$$[a,b] = \frac{1}{2}a_{xx}b_{yy} + \frac{1}{2}a_{yy}b_{xx} - a_{xy}b_{xy}.$$

The function $\overline{w}$ is the inward radial displacement measured from an unbuckled (fundamental) state, $\overline{\phi}$ is the Airy stress function, $\varepsilon^2 = t^2(192\pi^4R^2(1-\nu^2))^{-1}$, and the nondimensional load parameter is given by $\bar{\lambda} = P(8\pi^3ERt)^{-1}$. The unknowns $\overline{w}$ and $\overline{\phi}$ are defined on the two-dimensional spatial domain $(-\ell,\ell)\times(-1/2,1/2)$, where

$x \in (-\ell, \ell)$ is the axial and $y \in (-1/2, 1/2)$ is the tangential coordinate. Since the y-domain $(-1/2, 1/2)$ represents the circumference of the cylinder, the functions $\overline{w}$ and $\overline{\phi}$ are periodic in y; at the axial ends $x \in \{-\ell, \ell\}$ they satisfy the boundary conditions

$$\overline{w}_x = (\Delta \overline{w})_x = \overline{\phi}_x = (\Delta \overline{\phi})_x = 0.$$

For the experiments that we are interested in, the parameter ε is small, ranging from 10^{-2} to 10^{-4}. Here we rescale (2.1)–(2.2) such that the equations themselves are ε-independent, at the cost of a dependence on ε in the size of the spatial domain. Set

$$\overline{w} \mapsto \varepsilon w, \qquad \overline{\phi} \mapsto \varepsilon^2 \phi, \qquad x \mapsto \varepsilon^{1/2} x, \qquad y \mapsto \varepsilon^{1/2} y, \qquad \bar{\lambda} \mapsto \varepsilon \lambda, \tag{2.3}$$

so that the equations become

$$\Delta^2 w + \lambda w_{xx} - \phi_{xx} - 2\,[w, \phi] = 0, \tag{2.4}$$

$$\Delta^2 \phi + [w, w] + w_{xx} = 0. \tag{2.5}$$

The domain of definition of w and ϕ is now

$$\Omega := (-\ell \varepsilon^{-1/2}, \ell \varepsilon^{-1/2}) \times (-\tfrac{1}{2}\varepsilon^{-1/2}, \tfrac{1}{2}\varepsilon^{-1/2}),$$

which expands to $\mathbb{R}^2$ as $\varepsilon \to 0$. When not indicated otherwise, we choose the aspect ratio $2\ell = 1$; in section 4.3 we comment on the influence of domain size and aspect ratio.

The boundary conditions for w and ϕ now are

$$w \text{ is periodic in } y \quad \text{and} \quad w_x = (\Delta w)_x = 0 \text{ at } x = \pm \tfrac{1}{2}\varepsilon^{-1/2}, \tag{2.6a}$$

$$\phi \text{ is periodic in } y \quad \text{and} \quad \phi_x = (\Delta \phi)_x = 0 \text{ at } x = \pm \tfrac{1}{2}\varepsilon^{-1/2}. \tag{2.6b}$$

Equations (2.4)–(2.5) are related to the stored energy E and the average axial shortening S given by

$$E(w) := \frac{1}{2}\int_\Omega \left(\Delta w^2 + \Delta \phi^2\right) \qquad \text{and} \qquad S(w) := \frac{1}{2}\int_\Omega w_x^2. \tag{2.7}$$

Note that the function ϕ in (2.7) is determined from w by solving (2.5) with boundary conditions (2.6b); this uniquely defines ϕ up to an additive constant.

Solutions of (2.4)–(2.5) are stationary points of the total potential

$$F_\lambda(w) := E(w) - \lambda S(w). \tag{2.8}$$

This can be recognized as follows: If we substitute for w the perturbed function $w_\eta := w + \eta \tilde{w}$, then the perturbed Airy stress function ϕ_η solves

$$\Delta^2 \phi_\eta + [w, w] + 2\eta[w, \tilde{w}] + \eta^2 [\tilde{w}, \tilde{w}] + w_{xx} + \eta \tilde{w}_{xx} = 0.$$

Therefore $\phi_\eta = \phi + \eta \tilde{\phi} + O(\eta^2)$, where the perturbation $\tilde{\phi}$ solves

$$\Delta^2 \tilde{\phi} + 2[w, \tilde{w}] + \tilde{w}_{xx} = 0$$

with boundary condition (2.6b). Then

$$\begin{aligned} F'_\lambda(w)\cdot\tilde{w} &= \int_\Omega \Big[\Delta w\Delta\tilde{w} + \Delta\phi\Delta\tilde{\phi} - \lambda w_x\tilde{w}_x\Big] \\ &= \int_\Omega \Big[\Delta w\Delta\tilde{w} - \phi\big(2[w,\tilde{w}] + \tilde{w}_{xx}\big) - \lambda w_x\tilde{w}_x\Big] \\ &= \int_\Omega \Big[\Delta w\Delta\tilde{w} - \tilde{w}\big(2[w,\phi] + \phi_{xx}\big) - \lambda w_x\tilde{w}_x\Big], \end{aligned}$$

and this is a weak formulation of (2.4). Besides being stationary points of F_λ, solutions of (2.4)–(2.5) are also stationary points of E under the constraint of constant S; in this case λ is a Lagrange multiplier. We use both properties below.

3. The mountain pass: Overview. We briefly recall the general context of the Mountain Pass Theorem of Ambrosetti and Rabinowitz [2]. Let I be a functional defined on a Banach space X, and let w_1, w_2 be two distinct points in X. Consider the family Γ of all paths in X connecting w_1 and w_2 and define

$$c = \inf_{\gamma\in\Gamma}\max_{w\in\gamma} I(w), \tag{3.1}$$

that is, the infimum of the maxima of the functional I along paths in Γ. If $c > \max\{I(w_1), I(w_2)\}$, then the paths have to cross a "mountain range," and one may conjecture that there exists a critical point w_{MP} of I at the level c, called a mountain pass point.

We will apply this idea to the von Kármán–Donnell equations in the following way. We take for I the total potential F_λ (see (2.8)) at some fixed value of λ, and for the end point w_1 the origin. We will obtain a mountain pass solution by the following steps:

- MP1. We first show that $w_1 = 0$ is a strict local minimizer, or more precisely, that there exist $\varrho, \alpha > 0$ such that $F_\lambda(w) \geq \alpha$ for all w with $\|w\|_X = \varrho$ (Lemma 3.1).
- MP2. If ε is small enough, then there exists w_2 with $F_\lambda(w_2) \leq 0$ (Corollary 3.4).
- MP3. Given a sequence of paths γ_n that approximates the infimum in (3.1), we extract a (Palais–Smale) sequence of points $w_n \in \gamma_n$, each one close to the maximum along γ_n, and show that this sequence converges in an appropriate manner (Lemmas 3.5 and 3.6).

In this way it follows that there exists a mountain pass critical point w with $F_\lambda(w) = c$, provided that ε is sufficiently small (or that the domain is sufficiently large). For technical reasons (lack of coerciveness of the functional F_λ) this procedure can be performed only for almost all $0 < \lambda < 2$ (see Lemma 3.5).

In the rest of this section we detail the steps outlined above.

3.1. Choice of spatial domain. We are interested in mountain pass solutions of the system of equations (2.4)–(2.5) that are quasi independent of the domain size $\varepsilon^{-1/2}$, in the sense that they converge to a nontrivial solution on $\mathbb{R}^2$ as $\varepsilon \to 0$. This point of view suggests considering the problem on the whole of $\mathbb{R}^2$ rather than on a sequence of domains of increasing size; however, there are two reasons for not doing this. To start with, the numerical calculations described below are necessarily done on a bounded domain; more important, for the proof of existence of mountain pass points, boundedness of the domain is necessary. For these reasons we concentrate on bounded domains, while keeping the context of the unbounded domain in mind.

3.2. Functional setting and linearization. We introduce a functional setting for the functions w that is suggested by the linearization of the stored energy functional E. Writing $\phi = \phi_1 + \phi_2$, where

$$\Delta^2\phi_1 = -w_{xx} \qquad \text{and} \qquad \Delta^2\phi_2 = -[w,w], \tag{3.2}$$

we can expand the energy functional E as

$$E(w) = \frac{1}{2}\int_\Omega \Delta w^2 + \frac{1}{2}\int_\Omega \Delta\phi_1^2 + \int_\Omega \Delta\phi_1\Delta\phi_2 + \frac{1}{2}\int_\Omega \Delta\phi_2^2. \tag{3.3}$$

Since ϕ_2 is quadratic in w, the second derivative of E is given by

$$d^2E(0)\cdot u\cdot v = \int_\Omega \Delta u\Delta v + \int_\Omega \Delta\phi_1^u\Delta\phi_1^v,$$

where $\phi_1^{u,v}$ are obtained from u and v by replacing w with u or v in (3.2) and solving this equation for ϕ_1 with boundary conditions (2.6b). Inspired by this linearization of E, we define

$$X = \left\{\psi \in H^2(\Omega) : \psi_x\left(\pm\frac{1}{2}\varepsilon^{-1/2},\cdot\right) = 0,\ \psi \text{ is periodic in } y, \text{ and } \int_\Omega \psi = 0\right\}$$

with norm

$$\|w\|_X^2 = \int_\Omega \big(\,\Delta w^2 + \Delta\phi_1^2\,\big),$$

where $\phi_1 \in H^2(\Omega)$ is the unique solution of

$$\Delta^2\phi_1 = -w_{xx}, \qquad \phi_1 \text{ satisfies (2.6b)}, \qquad \text{and} \qquad \int_\Omega \phi_1 = 0.$$

This norm is equivalent to the H^2-norm on the set X, and with the appropriate inner product the space X is a Hilbert space.

We now address the requirements of the mountain pass theorem mentioned above in MP1–MP3.

3.3. The origin is a local minimizer. The norm in X is related in a natural manner to the shortening S, as demonstrated by the (sharp) estimate

$$\begin{aligned} 2S(w) &= \int_\Omega w_x^2 = -\int_\Omega w w_{xx} = \int_\Omega w\Delta^2\phi_1 \\ &= \int_\Omega \Delta w\Delta\phi_1 \le \frac{1}{2}\int_\Omega \Delta w^2 + \frac{1}{2}\int_\Omega \Delta\phi_1^2 = \frac{1}{2}\|w\|_X^2\,. \end{aligned} \tag{3.4}$$

This inequality strongly suggests that for $\lambda < 2$ the origin is a strict local minimum for the functional $F_\lambda(w) = E(w) - \lambda S(w)$.

LEMMA 3.1. *For any $\lambda < 2$, there exists $\varrho > 0$ such that*

$$\inf\big\{\ F_\lambda(w) : \|w\|_X = \varrho\ \big\} > 0.$$

Proof. Split $\phi = \phi_1 + \phi_2$ as in (3.2), and note that the function $\nabla\phi_1$ is bounded in L^∞ by the Sobolev imbedding $\{\psi \in H^3 : \int\psi = 0\} \hookrightarrow L^\infty$:

$$\|\nabla\phi_1\|_{L^\infty}^2 \le C\left\|\Delta^2\phi_1\right\|_{L^2}^2 = C\|w_{xx}\|_{L^2}^2\,.$$

The third term on the right-hand side of (3.3) can now be rewritten as

$$\int_\Omega \Delta\phi_1\Delta\phi_2 = \int_\Omega \phi_1[w,w] = \int_\Omega \phi_1(w_{xx}w_{yy} - w_{xy}^2) = \int_\Omega (\phi_{1y}w_x w_{xy} - \phi_{1x}w_x w_{yy}),$$

which we estimate by

$$\left|\int_\Omega \Delta\phi_1\Delta\phi_2\right| \le 2\,\|\nabla\phi_1\|_{L^\infty}\,\|w_x\|_{L^2}\,\|\Delta w\|_{L^2} \le C\,\|w_x\|_{L^2}\,\|\Delta w\|_{L^2}^2 \le C\sqrt{S(w)}\,\|w\|_X^2\,.$$

Since $\lambda < 2$, choose $0 < \varrho < (2-\lambda)/2C$ and define $\eta = (1/2)(1 - C\varrho - \lambda/2) > 0$. Then on the set $\mathcal{C} = \{w : \|w\|_X = \varrho\}$, using (3.4), we find that

$$\begin{aligned} F_\lambda(w) &= E(w) - \lambda S(w) \\ &\ge \frac{1}{2}\,\|w\|_X^2 - C\sqrt{S(w)}\,\|w\|_X^2 - \lambda S(w) \\ &\ge \frac{\varrho^2}{2} - C\varrho^3\frac{1}{2} - \lambda\frac{\varrho^2}{4} \\ &= \frac{\varrho^2}{2}\left(1 - C\varrho - \frac{\lambda}{2}\right) \\ &= \eta\varrho^2. \qquad \square \end{aligned}$$

This lemma implies that by choosing w_1 to be the origin we have shown condition MP1.

Remark 3.2. Although inequality (3.4) suggests that the origin should be a local minimizer for any domain Ω, bounded or not, the proof above applies only to bounded domains. F. Otto has constructed a proof of this result that is valid on any domain (private communication). Interestingly, this proof uses not only the cubic energy term $\int_\Omega \Delta\phi_1\Delta\phi_2$, but also the quartic term $\int_\Omega \Delta\phi_2^2$, and appears to break down without this latter term.

3.4. Periodic solutions exist with negative F_λ. To satisfy MP2 we show in this section that for any $\lambda > 0$, functions $w \in X$ exist, for which $F_\lambda(w) = E(w) - \lambda S(w) < 0$. To do this we construct a sequence of functions w_δ with specific scaling properties.

LEMMA 3.3. *There exists a sequence of functions w_δ, 1-periodic on $\mathbb{R}^2$, such that as $\delta \to 0$,*

1. $\displaystyle\int_{[-1/2,1/2]^2} w_{\delta x}^2 \longrightarrow c \quad$ *for some $c > 0$,*
2. $\displaystyle\int_{[-1/2,1/2]^2} \Delta w_\delta^2 = O(\delta^{-1})$

   and $\displaystyle\int_{[-1/2,1/2]^2} \Delta\phi_\delta^2 = O(\delta^{2-\alpha}) \quad$ *for any $\alpha > 0$.*

Here the functions w_δ and ϕ_δ solve (2.5) *with periodic boundary conditions. In addition, w_δ and ϕ_δ satisfy boundary conditions* (2.6) *on the boundary of $[-1/2,1/2]^2$.*

The proof, given in Appendix A, is inspired by the so-called *Yoshimura pattern* [37], a folding pattern by which a flat sheet of paper, or a cylindrical sheet of thin material, can be folded into a macroscopically cylindrical structure with zero Gaussian curvature but locally infinite total curvature (Figure 3.1). The functions w_δ are smoothed versions of the Yoshimura pattern, adapted to the geometrically

FIG. 3.1. *Yoshimura folding pattern.*

linear setting of the von Kármán–Donnell equations, and δ measures the width of the fold.

COROLLARY 3.4.

1. *Fix $\lambda > 0$. If ε is sufficiently small, then there exists $w \in X$ such that $F_\lambda(w) < 0$.*
2. *Fix ε sufficiently small. Then there exists $\lambda_0(\varepsilon) \in [0, 2)$ such that for all $\lambda > \lambda_0$, there exists $w \in X$ with $F_\lambda(w) < 0$.*

Proof. By scaling the functions w_δ of Lemma 3.3, the claims can be fulfilled as follows: Let $\delta = \varepsilon^{2/3}$, and set

$$\tilde{w}_\varepsilon(x, y) = \varepsilon^{-1} w_{\varepsilon^{2/3}}(x\varepsilon^{1/2}, y\varepsilon^{1/2}), \qquad \tilde{\phi}_\varepsilon(x, y) = \varepsilon^{-2}\phi_{\varepsilon^{2/3}}(x\varepsilon^{1/2}, y\varepsilon^{1/2});$$

then $\tilde{w}_\varepsilon \in X$, and (2.5) is invariant under this scaling; in addition, choosing $\alpha = 1/6$, we obtain

$$Q_\varepsilon := \frac{\displaystyle\int_{\Omega_\varepsilon} \left[\Delta\tilde{w}_\varepsilon^2 + \Delta\tilde{\phi}_\varepsilon^2\right]}{\displaystyle\int_{\Omega_\varepsilon} \tilde{w}_{\varepsilon x}^2} = O\big(\varepsilon^{1/6}\big) \qquad \text{as } \varepsilon \to 0.$$

Therefore $\lim_{\varepsilon\to 0} Q_\varepsilon = 0$, proving the first claim. For the second claim, we fix ε such that $Q_\varepsilon < 2$; then for all $\lambda > Q_\varepsilon$, $F_\lambda(\tilde{w}_\varepsilon) < 0$. ∎

3.5. Convergence of selected sequences. For given $\lambda \in (0, 2)$ and for sufficiently small $\varepsilon > 0$, the two previous sections provide two points: the origin $w_1 = 0$ that satisfies MP1, and a point w_2 with $F_\lambda(w_2) < 0$, such that

$$c(\lambda) := \inf_{\gamma\in\Gamma} \max_{w\in\gamma} F_\lambda(w) > 0, \tag{3.5}$$

where Γ is the set of curves connecting 0 and w_2,

$$\Gamma = \{\gamma \in C([0, 1]; X) : \gamma(0) = 0,\ \gamma(1) = w_2\}$$

(actually, Γ depends on λ through the dependence on w_2, but w_2 can be taken independently of λ in a neighborhood of a given $\lambda \in (0, 2)$).

We were unable to prove the classical Palais–Smale condition, which reads as follows:

> For any sequence $w_n \in X$ such that $F_\lambda(w_n) \to c$ and $F'_\lambda(w_n) \to 0$ in X', there exists a subsequence that converges in X.

The difficulty lies in the lack of coerciveness of the functional F_λ: the quotient $F_\lambda(w)/\|w\|_X^2$ is not bounded away from zero, implying that Palais–Smale sequences may be unbounded in X.

The "Struwe monotonicity trick" [29] provides a way of proving the boundedness of Palais–Smale sequences for at least *almost all* $\lambda \in (0,2)$. The pertinent observation is that for fixed w, $F_\lambda(w)$ is decreasing in λ; consequently, $c(\lambda)$ is a decreasing function of λ and therefore differentiable in almost all $\lambda \in (0,2)$. If $\gamma(t)$ is the highest point of a near-optimal curve γ at some λ_0, then $c'(\lambda_0)$ should be close to $-S(\gamma(t))$. Finiteness of c' at λ_0 thus implies that near–mountain pass points have bounded S, and this additional information suffices for the construction of bounded sequences.

LEMMA 3.5. *Let $\lambda \in (0,2)$ be such that $c'(\lambda)$ exists. Then there exists a bounded Palais–Smale sequence w_n, i.e., a sequence satisfying that*

1. *w_n is bounded in X;*
2. *$F'_\lambda(w_n) \to 0$ in X' and $F_\lambda(w_n) \to c(\lambda)$;*
3. *there exists a sequence of curves $(\gamma_n) \subset \Gamma$ such that $w_n \in \gamma_n([0,1])$ and $\max_{t\in[0,1]} F_\lambda(\gamma_n(t)) \to c(\lambda)$.*

In [28] this same argument was used to study mountain pass points for the related one-dimensional functional

$$J_\lambda(u) = \int_{\mathbb{R}} \Big\{ \frac{1}{2} u''^2 - \frac{\lambda}{2} u'^2 + F(u) \Big\},$$

where F is a nonnegative double- or single-well potential. The proof of Lemma 3.5 repeats verbatim the proof of [28, Prop. 5], and we omit it here.

LEMMA 3.6. *The sequence w_n given by Lemma 3.5 is compact in X, and a subsequence converges to a stationary point $w \in X$ of F_λ.*

Strictly speaking, the stationary point given by this lemma may not be a mountain pass point itself, in the sense that there may not be a curve $\gamma \in \Gamma$ of which w is the highest point. Property 3 of Lemma 3.5, however, states that w has an approximate mountain pass character.

Proof. We extract a subsequence that converges weakly in X and strongly in H^1 and L^∞ to a limit w. Defining ϕ_{1n} and ϕ_{2n} by (3.2), we find that the right-hand sides in (3.2) are bounded in L^2 and L^1, and therefore that ϕ_{1n} and ϕ_{2n} converge strongly (up to extracting a subsequence, which we do without changing notation) in H^2 to functions $\phi_{1,2}$. Both functions $\phi_{1,2}$ are again related to w by (3.2); for ϕ_2 this follows from remarking that for given $\zeta \in C_c^\infty(\Omega)$,

$$\lim_{n\to\infty} \int_\Omega \zeta[w_n, w_n] = \lim_{n\to\infty} \int_\Omega w_n[\zeta, w_n] = \int_\Omega w[\zeta, w] = \int_\Omega \zeta[w, w],$$

so that the right-hand side converges in the sense of distributions. Similarly, it follows from the strong H^2-convergence of $\phi_n = \phi_{1n} + \phi_{2n}$ that

$$\lim_{n\to\infty} \int_\Omega \phi_n[w_n, w - w_n] = \lim_{n\to\infty} \int_\Omega w_n[\phi_n, w - w_n] = 0.$$

To show that w_n converges strongly in X, note that the derivative $F'_\lambda(w_n)$ can be characterized as

$$F'_\lambda(w_n) \cdot v = \int_\Omega \Delta w_n \Delta v - \int_\Omega \phi_n\big(v_{xx} + 2[w_n, v]\big) - \lambda \int_\Omega w_{nx} v_x.$$

We now calculate

$$\lim_{n\to\infty}\left\{\int_\Omega \Delta w^2 - \int_\Omega \Delta w_n^2\right\} = \lim_{n\to\infty}\int_\Omega \Delta w_n \Delta(w-w_n)$$
$$= \lim_{n\to\infty}\left\{F_\lambda'(w_n)\cdot(w-w_n) + \int_\Omega \phi_n\big((w-w_n)_{xx} + 2[w_n, w-w_n]\big) + \lambda\int_\Omega w_{nx}(w-w_n)_x\right\} = 0.$$

The strong convergence of w_n in X now follows from the uniform convexity of X. ∎

4. Numerical results.

4.1. Description of the algorithm. Our goal in this section is to find, numerically, critical points of F_λ. Although we will focus on mountain pass points described above and sketch the method used to find them, numerical approximations of other critical points of F_λ will be shown as well. More details on all the numerical methods used are given in a companion paper [17].

In order to employ the mountain pass algorithm, we discretize (2.4)–(2.8) using finite differences. The algorithm was first proposed in [8] for a second-order elliptic problem in one dimension. It was later used in [18] for a fourth-order problem in two dimensions.

The main idea of the algorithm is illustrated in Figure 4.1. We take a discretized path connecting $w_1 = 0$ with a point w_2 such that $F_\lambda(w_2) < 0$. After finding the point z_m at which F_λ is maximal along the path, this point is moved a small distance in the direction of the steepest descent of F_λ. Thus the path has been deformed and the maximum of F_λ lowered. This deforming of the path is repeated until the maximum along the path cannot be lowered any more: a critical point w_{MP} has been reached.

Figure 4.2(a) shows a numerical solution of (2.4)–(2.5) obtained by this algorithm with $\lambda = 1.1$. The graph in each panel shows the radial displacement w as a function

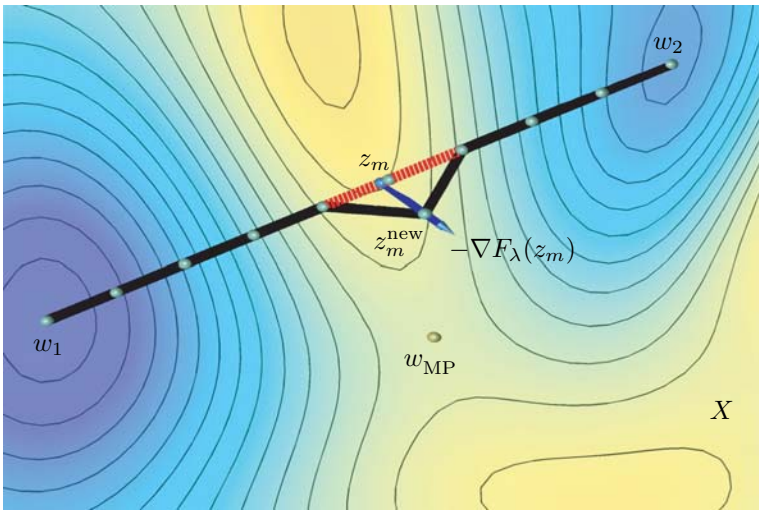


FIG. 4.1. *Deforming the path in the main loop of the mountain pass algorithm: point z_m is moved a small distance in the direction $-\nabla F_\lambda(z_m)$ and becomes z_m^{new}. This step is repeated until the mountain pass point w_{MP} is reached.*

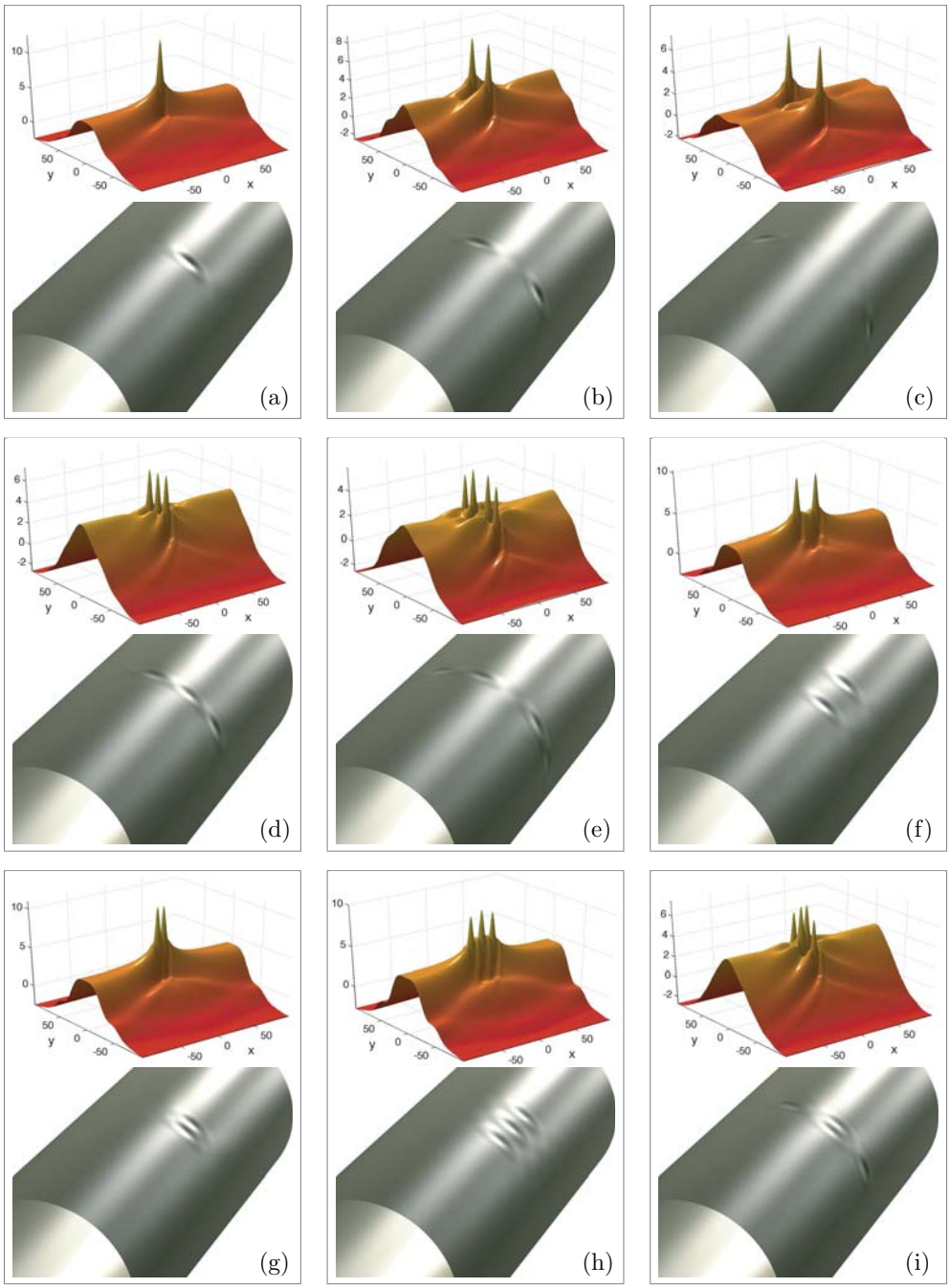


FIG. 4.2. *Numerical solutions found using the (constrained) mountain pass algorithm, constrained steepest descent method, and the Newton algorithm. The figures show both the graph of* $w(x, y)$ *and its rendering on a cylinder.*

of x and y. Rendered on a cylinder, this solution represents a single dimple, as can be seen below the graphs.

We restrict our computations to functions that are even about the x- and y-axes, i.e., to the subspace

$$\mathscr{S} = \{\psi \in X : \psi(x,y) = \psi(-x,y),\ \psi(x,y) = \psi(x,-y)\},$$

thus reducing the computational domain Ω to one quarter, e.g., $(0, \frac{1}{2}\varepsilon^{-1/2})\times(0, \frac{1}{2}\varepsilon^{-1/2})$. The boundary conditions (2.6a) then become

$$w_x = (\Delta w)_x = 0 \text{ for } x \in \{0, \tfrac{1}{2}\varepsilon^{-1/2}\} \quad \text{and} \quad w_y = (\Delta w)_y = 0 \text{ for } y \in \{0, \tfrac{1}{2}\varepsilon^{-1/2}\}.$$

This symmetry assumption has many numerical advantages, but on the other hand it a priori excludes solutions that do not belong to $\mathscr{S}$.

For the mountain pass algorithm, we always use the unbuckled state $w_1 = 0$ as the first end point of the paths. The choice of the second end point w_2 has a nontrivial influence on the solution to which the mountain pass algorithm converges. Corollary 3.4 guarantees the existence of $w_2 \in \mathscr{S}$ with $F_\lambda(w_2) < 0$; in the numerical implementation, however, we found such a w_2 by a steepest descent method (rather than by taking the function constructed in the proof of Lemma 3.3): starting from a function w_0 that has one peak located in the center of the domain Ω, we solved the initial value problem

$$\frac{d}{dt}w(t) = -\nabla F_\lambda(w(t)), \quad w(0) = w_0, \tag{4.1}$$

on an interval $(0,T)$ until $F_\lambda(w(T)) < 0$. We then defined $w_2 = w(T)$.

A different choice of w_2 (or, more precisely, of the starting point w_0 of (4.1)) can lead to a different solution of the problem, as Figures 4.2(b) and (f) show. Here w_0 was chosen to have two peaks with centers on the axes $x = 0$ and $y = 0$, respectively. The algorithm then converged to a numerical solution with two dimples in the circumferential and axial directions, respectively.

Note that the numerical solution w_{MP} selected by the mountain pass algorithm has the mountain pass property in a certain neighborhood only: there exists a ball $B_\rho(w_{\mathrm{MP}})$ and two points $\tilde{w}_1, \tilde{w}_2 \in B_\rho(w_{\mathrm{MP}})$ such that

$$F_\lambda(w_{\mathrm{MP}}) = \inf_{\gamma\in\tilde{\Gamma}} \max_{w\in\gamma} F_\lambda(w) > \max\{F_\lambda(\tilde{w}_1), F_\lambda(\tilde{w}_2)\},$$

where $\tilde{\Gamma}$ is the set of curves in $B_\rho(w_{\mathrm{MP}})$ connecting $\tilde{w}_1$ and $\tilde{w}_2$. The reason for this is that the algorithm deforms a certain initial path connecting w_1 and w_2 which is fixed. In order to recover the global character, one would need to run the algorithm for all possible initial paths.

The rest of the numerical solutions shown in Figure 4.2 were obtained under a prescribed value of shortening S by the constrained steepest descent method and the constrained mountain pass algorithm [16, 17].

4.2. Calculation of $F_\lambda(w_{\mathrm{MP}})$. In the preceding sections we showed that

1. for a sufficiently large domain Ω, a function w_2 on Ω exists with $F_\lambda(w_2) < 0$;
2. for each such function w_2 and for almost all $0 < \lambda < 2$, a mountain pass solution $w_{\mathrm{MP}} = w_{\mathrm{MP}}(\lambda, \Omega, w_2)$ exists.

Different end points w_2 may give rise to different mountain pass points, as we have observed in the numerical experiments described above. We therefore define the mountain pass energy function V on $(0,2)$ by

$$V(\lambda,\Omega) := \inf_{w_2}\{\, F_\lambda\big(w_{\mathrm{MP}}(\lambda,\Omega,w_2)\big) : F_\lambda(w_2) < 0\,\}. \tag{4.2}$$

For a given λ, the value of $V(\lambda,\Omega)$ is the lowest height (or energy level) at which one may pass from the origin to a point with negative total potential F_λ. We now derive some of its properties and calculate it numerically.

LEMMA 4.1 (properties of $V(\lambda,\Omega)$).

1. *For sufficiently large Ω there exists $\lambda_0(\Omega) \geq 0$ such that $V(\lambda,\Omega) < \infty$ for almost all $\lambda \in (\lambda_0, 2)$.*
2. *V is a decreasing function of λ.*
3. *For sufficiently large Ω, there exists $c(\Omega) > 0$ such that*

$$V(\lambda,\Omega) \leq c(2-\lambda)^3$$

for sufficiently small $2-\lambda > 0$.

Proof. Part 1 is a reformulation of the main result of section 3, making use of Corollary 3.4. For part 2 we remark that for each fixed w, $F_\lambda(w)$ is a decreasing function of λ; the infimum of a set of decreasing functions is again decreasing.

For part 3, let us set

$$E(w) = E_2(w) + E_3(w) + E_4(w),$$

where

$$E_2(w) := \frac{1}{2}\,\|w\|_X^2 = \frac{1}{2}\int_\Omega \big(\Delta w^2 + \Delta\phi_1^2\big), \quad E_3(w) := \int_\Omega \Delta\phi_1\Delta\phi_2,$$
$$\text{and} \quad E_4(w) := \frac{1}{2}\int_\Omega \Delta\phi_2^2,$$

where ϕ_1 and ϕ_2 are determined from w by (3.2) (see also (3.3)). Note that E_n has homogeneity n, i.e., $E_n(\mu w) = \mu^n E_n(w)$.

A classical result in the engineering literature of cylinder buckling (see, e.g., [20]) states that there exists a periodic function w on $\mathbb{R}^2$ such that

$$E_2(w) = 2S(w) \qquad \text{and} \qquad E_3(w) + E_4(w) < 0.$$

Here and below we consider the integrals that define $E_n(w)$ and $S(w)$ as taken over a single period cell. For sufficiently small $2-\lambda > 0$ the inequality above gives that $F_\lambda(w) = (2-\lambda)S(w) + E_3(w) + E_4(w)$ is negative, implying that w is an admissible end point w_2 for definition (4.2) of $V(\lambda,\Omega)$, and the connecting line segment $\{\mu w : 0 \leq \mu \leq 1\}$ is therefore an admissible curve in Γ. Consequently,

$$V(\lambda) \leq \sup_{0\leq\mu\leq 1} F_\lambda(\mu w) = \sup_{0\leq\mu\leq 1} \mu^2(2-\lambda)S(w) + \mu^3 E_3(w) + \mu^4 E_4(w).$$

The supremum on the right-hand side is obtained at

$$\mu = \frac{3\,|E_3(w)|}{8E_4(w)}\left\{1 - \sqrt{1 - \frac{32(2-\lambda)S(w)E_4(w)}{9E_3(w)^2}}\right\}$$
$$= \frac{2S(w)}{3\,|E_3(w)|}\,(2-\lambda) + o(1) \quad \text{as } \lambda \to 2,$$

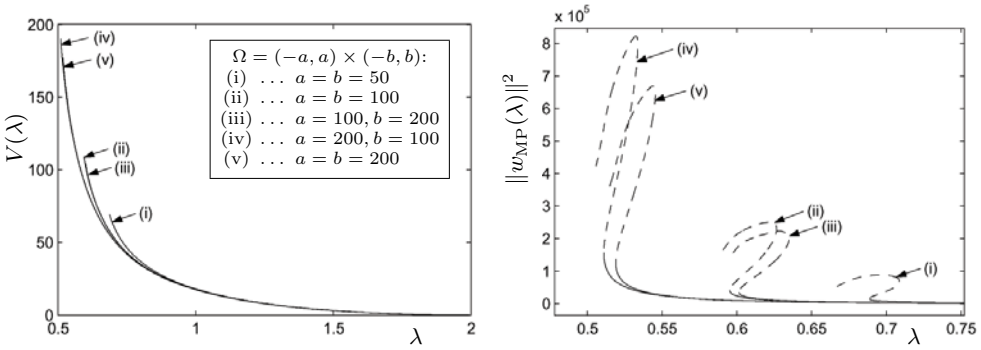


FIG. 4.3. *Left: The mountain pass energy $V(\lambda, \Omega)$ found numerically for various sizes of domain Ω. Right: The solid line shows the same computation as on the left, but plotted for the norm of $w_{\mathrm{MP}}(\lambda)$ squared. The dashed curve was obtained by continuation of the solid curve; the solutions on the dashed curve do not represent mountain pass points.*

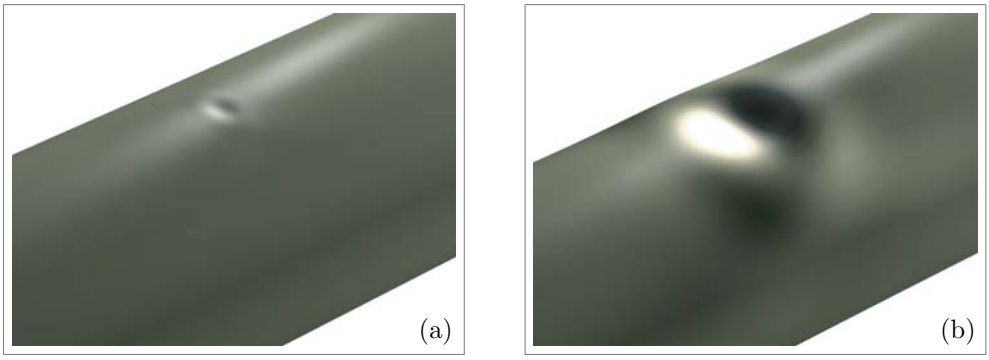


FIG. 4.4. *The numerical mountain pass solution w_{MP} of the scaled equations* (2.4)–(2.5) *for a given value of λ rendered on two cylinders of the same radius R but a different thickness t:* (a) $t/R = 0.003$, (b) $t/R = 0.04$.

implying that the claim holds for periodic functions. The generalization to nonperiodic functions on large domains Ω (i.e., for small ε) is made by filling the domain with a large number of periodic cells of the function w and connecting the function smoothly to the boundary of Ω. □

Figure 4.3 shows graphs of the mountain pass energy $V(\lambda, \Omega)$ computed for various sizes of domain Ω. For each domain, the mountain pass algorithm was employed to compute w_{MP} for several values of λ. These mountain pass solutions were then continued in λ using numerical path following.

4.3. Influence of the domain. The localized nature of the solutions calculated above suggests that they should be independent of domain size, in the sense that for a sequence of domains of increasing size the solutions converge (for instance, pointwise on compact subsets). Such a convergence would also imply convergence of the associated energy levels. Similarly, we would expect that the aspect ratio of the domain is of little importance in the limit of large domains.

We have tested these hypotheses by computing mountain pass solutions on domains of different sizes and aspect ratios. Generally solutions on different domains compare well; the maximal difference between the second derivatives of w is two or

three orders of magnitude smaller than the supremum norm of the same derivative (the details of this comparison are given in [17]).

Here we include only a calculation of the mountain pass energy level $V(\lambda, \Omega)$ for different aspect ratios and sizes of domain Ω (see Figure 4.3).

The comparison of solutions computed on different domains and their respective energies suggests that for each λ we are indeed dealing with a single, localized function defined on $\mathbb{R}^2$, of which our computed solutions are finite-domain adaptations. In the rest of this paper we adopt this point of view, and consequently we will write $V(\lambda)$ instead of $V(\lambda, \Omega)$.

A consequence of this point of view is that dimples in cylinders with different geometric parameters are mapped to the same rescaled solution, or equivalently, that the same single-dimple solution of (2.4)–(2.5) corresponds to differently sized dimples on an actual cylinder, as a function of the parameters (see Figure 4.4).

5. Interpretation: Imperfection sensitivity. We now turn to the relevance of the mountain pass in the context of a loading problem. This relevance can be best understood in the context of imperfections in the loading conditions (rather than geometric imperfections) such as in the case of a (small) lateral loading.

Under a small lateral load, an equilibrium w_0, which is a local minimum of the functional F_λ, may be perturbed into an equilibrium $\tilde{w}_0$ of a perturbed functional $\tilde{F}_\lambda$. Since w_0 is a local minimum, $F_\lambda(\tilde{w}_0) > F_\lambda(w_0)$; i.e., with respect to the unperturbed system, $\tilde{w}_0$ has a higher total potential than w_0. The level of F_λ that is reached is a measure of the magnitude of the imperfection—a different measure than is commonly used, but one that has distinct advantages.

By definition, the number $V(\lambda)$ is the lowest energy level at which it is possible to move between the basins of attraction of w_1 and w_2 (Figure 5.1). If the loading imperfection is interpreted, as above, as a mechanism capable of maintaining the system at a higher energy level than that of the neighboring fundamental minimizer, then the number $V(\lambda)$ is critical: as long as the imperfection is so small that the energy is never raised by more than $V(\lambda)$, the new stationary point will be part of the same basin of attraction as w_1. For larger imperfections, however, it becomes possible to leave the fundamental basin of attraction, resulting in a large jump in state space.

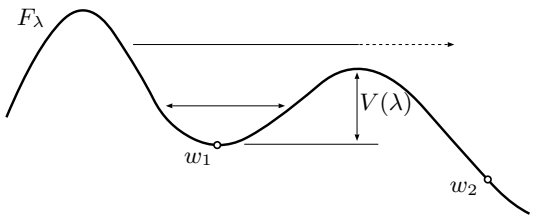


FIG. 5.1. *In order to leave the basin of attraction of w_1, the surplus energy should exceed $V(\lambda)$.*

This line of reasoning provides a natural measure of the size of imperfections, namely, the maximal increase in energy (in the perfect structure) that an imperfection can achieve. It also provides a natural measure of the stability of the unbuckled state, since a higher mountain pass energy level implies a larger class of loading imperfections under which the state remains in the fundamental basin of attraction. This observation allows us to connect systems with different geometrical characteristics and compare their relative sensitivity to imperfections.

5.1. Calibrating the mountain pass energy. Comparing cylinders of varying geometries requires a common measure of imperfection sensitivity. It is not a priori clear which measure to take; e.g., one might consider either the mountain pass energy itself or the average spatial density of this energy, which will result in different comparisons for cylinders of different wall volumes. Here we choose to rescale the mountain pass energy level by the other energy level present in the loaded cylinder, i.e., the energy that is stored in the homogeneous compression of the unbuckled shell.

This calculation can be done in two slightly different ways. The first and most straightforward is to rescale the dimensional mountain pass energy (see (B.11) and (2.3)),

$$64\pi^6 EtR^2\varepsilon^3 V(\lambda) = \frac{Et^4}{8(3(1-\nu^2))^{3/2}R}\, V(\lambda),$$

by the elastic strain energy stored in the full length of the compressed cylinder of length L,

$$\frac{L}{4\pi ERt}\, P^2 = \frac{\pi t^3 EL}{12(1-\nu^2)R}\, \lambda^2,$$

to give an energy ratio, or a rescaled mountain pass energy level,

$$\alpha = \frac{1}{2\pi\sqrt{3(1-\nu^2)}}\, \frac{t}{L}\, \frac{V(\lambda)}{\lambda^2}. \tag{5.1}$$

From this expression and the calculation shown in Figure 4.3, curves may be drawn in a plot of load versus the ratio L/t (see Figure 5.2). Note that to obtain this figure from Figure 4.3 the curve $V(\lambda)$ was fitted to extend the range of λ. Figure 5.2 shows the following two remarkable features:

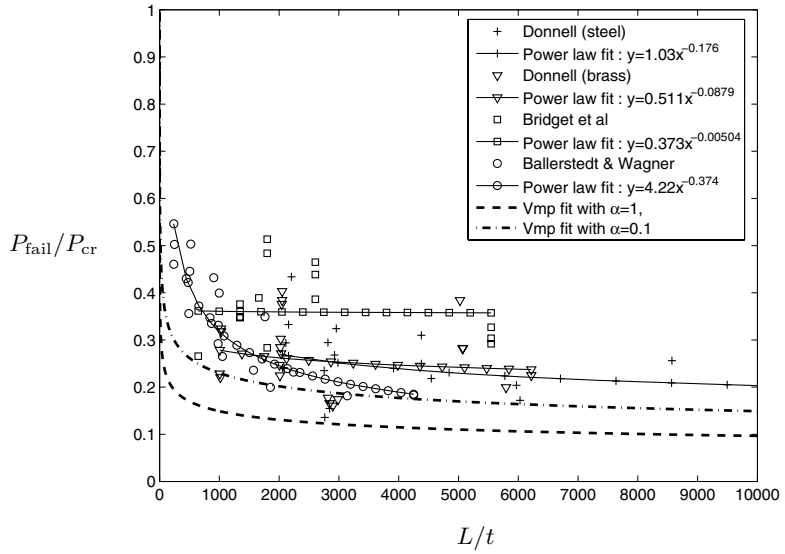


FIG. 5.2. *Shown are the same data as in Figure* 1.1*, with the addition of two curves of constant* $\alpha = 1$*,* $\alpha = 0.1$*, where* α *is given in* (5.1)*. Note that the load at which the mountain pass energy equals the stored energy in the prebuckled cylinder* ($\alpha = 1$) *appears to be a lower bound on the data.*

1. The general trend of the constant-α curves is very similar to the trend of the experimental data.
2. The $\alpha = 1$ curve, which indicates the load at which the mountain pass energy equals the stored energy in the prebuckled cylinder, appears to be a lower bound on the data.

One may also consider an alternative way of rescaling energy. The cylinder is a long structure, and it is not clear to what extent the length of the structure is relevant for the imperfection sensitivity. It may be reasonable to compare the energy of the mountain pass with the stored energy contained in a representative section of the cylinder; the radius R provides a natural length scale for such a representative section.

Similar to Figure 5.2, Figure 5.3 presents curves of constant β, where β is the ratio of mountain pass energy to stored energy in a section of length $2\pi R$:

$$\beta = \frac{1}{4\pi^2\sqrt{3(1-\nu^2)}}\,\frac{t}{R}\,\frac{V(\lambda)}{\lambda^2}. \tag{5.2}$$

Once again, to obtain Figure 5.3 we fitted $V(\lambda)$ from Figure 4.3 to extend the range of λ.

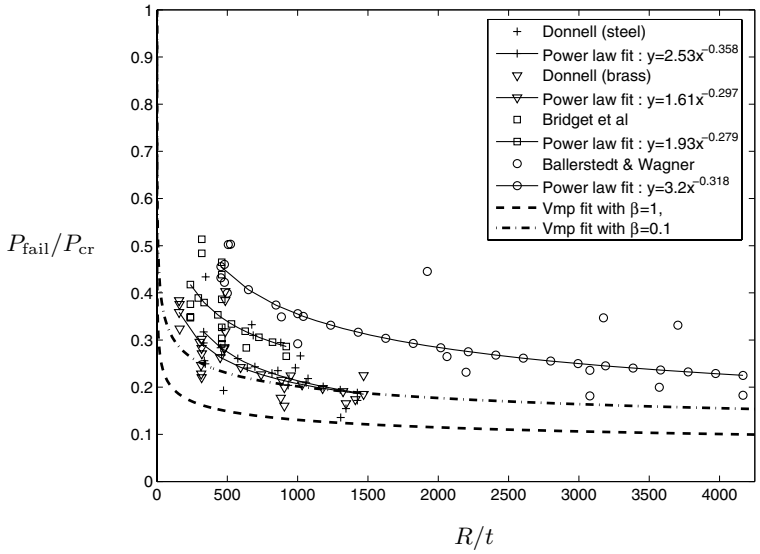


FIG. 5.3. *Experimental data and fit to V_{mp} with $\beta = 1$ and $\beta = 0.1$ in (5.2). Again, the load at which the mountain pass energy equals the stored energy in a representative portion of the prebuckled cylinder ($\beta = 1$) appears to be a lower bound on the data.*

6. Discussion and conclusions. The mathematical results and their interpretation in the context of a loading problem have brought about a number of new and improved insights.

6.1. The cylinder has doubly localized solutions. The subcritical nature of the bifurcation in Figure 1.2 strongly suggests that equilibria exist with deformation localized to a small portion of the cylinder length. In [20, 24, 25] such solutions are indeed calculated numerically and investigated analytically; these solutions are periodic around the cylinder and have exponential decay in the axial direction.

The localization in the axial direction demonstrated by these solutions is consistent with results on simpler systems, such as the laterally supported strut [21, 26]. The behavior of the cylinder in the tangential direction is not as well understood. The lack of localization in the simply supported flat plate [13] suggests that the cylinder should also prefer tangentially delocalized solutions, as do most of the experiments. The single- and multiple-dimple solutions of this paper, however, clearly demonstrate that doubly localized solutions do exist, and that some of these can be stable under constrained shortening.

6.2. The mountain pass is a single-dimple solution. The fact that the mountain pass solution exists follows essentially from two features, the local minimality of the unbuckled state and the existence of a large-deflection state of lower energy. The former is a simple consequence[3] of the subcritical load level, but the latter is based on an essential property of the cylinder: for a sequence of cylinders for which $R/t \to \infty$, the nondimensionalized load-carrying capacity (the highest load at which the unbuckled state is not only a local but also a global energy minimizer) decreases to zero. This property was demonstrated implicitly by Hoff, Madsen, and Mayers [15], and Lemma 3.3 provides a simplified proof of this result and a simple sequence of functions that illustrates the property.

However, the fact that the mountain pass solution is localized, and even is the most localized solution that is possible—a single dimple—is interesting in its own right and provides a complementary view of the discussion of localization above. A different way of formulating this result is that "creating the first dimple is the major obstacle"; afterwards one may increase the size of the dimple and add further dimples without ever returning to the same high energy level. In itself this interpretation points to a relationship between single dimples and imperfection sensitivity.

6.3. Single dimples in other contexts. Interestingly, single dimples have appeared in the literature in a number of seemingly unrelated ways:

- In the celebrated high-speed camera images of Eßlinger [12], the first visible deformation is a single, well-developed dimple halfway between the ends of the cylinder. New dimples quickly appear next to this first dimple, and the deformation then spreads around the cylinder in an axial direction. It is remarkable, though, that the first visible deformation is a single dimple.
- Some of the "worst" imperfections calculated by Deml and Wunderlich [9] and Wunderlich and Albertin [35] are in the form of a single dimple; as the load decreases, the dimple contracts and becomes even more concentrated.
- Hühne et al. [19] assert that single dimples are also *realistic* and *stimulating* imperfections in the sense of [34].
- Zhu, Mandal, and Calladine [38] base their analysis of the scaling behavior of the experimental buckling load on the behavior of a single dimple in other structural situations (such as the point-loaded cylinder and the sphere under uniform external pressure).

Note that the single-dimple appearances above are of three different types. Eßlinger's dimple is an experimental observation; the dimples of Wunderlich and coworkers and of Hühne et al. are geometric imperfection profiles; and the dimples studied by Zhu, Mandal, and Calladine are only analogies, since they are solutions of different problems.

[3] On a finite domain this consequence is indeed simple; on an infinite domain it appears that not only the third-order term but also the fourth-order term in the energy has to be taken into account, as remarked in section 3.

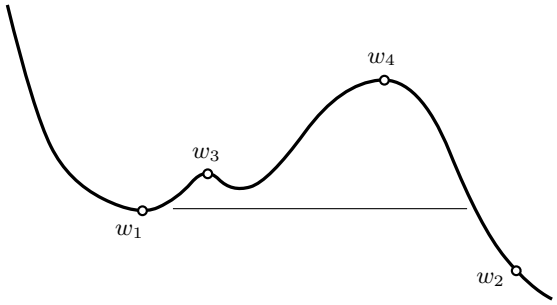


FIG. 6.1. *While a local algorithm for finding a critical point may settle on a minor critical point such as w_3, the mountain pass algorithm, by its global setup, will converge to the essential obstacle w_4.*

6.4. Scale-invariance of the localized solutions. It is an interesting observation that the von Kármán–Donnell equations can be rescaled to depend only on the (rescaled) load level. For localized solutions, for which the boundary plays no role of importance, this implies that the set of solutions reduces to a one-parameter family. This not only allows for efficient computation of the behavior of such solutions but also gives interesting insight into the relationship between dimples in cylinders of varying geometry (see, e.g., Figure 4.4).

Naturally the scale-invariance is expected to break upon replacing the von Kármán–Donnell equations with a different (probably more detailed) shell model. Nonetheless, it may be reasonably expected that much of the understanding of the relationship between cylinders of different geometries remains roughly correct.

We certainly also expect that the large-scale geometry of the energy landscape does not depend on the specific model of the cylinder. Using a discrete mountain pass algorithm to find mountain pass points therefore does not depend on the von Kármán–Donnell equations and should give similar results regardless of which shell model is used.

6.5. Connection with sensitivity to imperfections and "perturbation energy." Kröplin, Dinkler, and Hillmann [23]; Duddeck et al. [11]; and Wagenhuber and Duddeck [32] were the first to suggest an estimate of the stability of the unbuckled state in terms of the ratio of a "perturbation energy" (*Störenergie*) to the prebuckling energy. In early papers [23, 11] the perturbations are still fixed rather than determined, but from both the introduction and the final results in [32] it may be deduced that an optimization is done over all perturbations (although this is simultaneously contradicted on page 333 of [32]). Unfortunately, these papers do not provide enough details for determining exactly what the authors calculate.

There is one aspect in which our method can clearly be seen to differ from these earlier approaches. The discrete mountain pass algorithm takes into account *global* features of the energy landscape and provides a global measure of the separation barrier between two states that lie far apart. This is different from the papers mentioned above, in which the method uses only local information (reflected, for instance, in the assumption that the equilibria in question lie on the same bifurcation branch). This difference is illustrated in Figure 6.1, where a local analysis might find stationary point w_2, but the mountain pass algorithm will find the more important obstacle w_4.

Appendix A. Proof of Lemma 3.3. Lemma 3.3 states that *there exists a*

sequence of functions w_δ, 1-periodic on $\mathbb{R}^2$, such that

$$\int_{[-1/2,1/2]^2} w_{\delta x}^2 \sim 1, \qquad \int_{[-1/2,1/2]^2} \Delta w_\delta^2 = O(\delta^{-1}), \tag{A.1}$$

$$\text{and} \qquad \int_{[-1/2,1/2]^2} \Delta \phi_\delta^2 = O(\delta^{2-\alpha}) \quad \text{as} \quad \delta \to 0$$

for any $\alpha > 0$. Here the function ϕ_δ solves (2.5) *with periodic boundary conditions. In addition, w_δ and ϕ_δ satisfy* (2.6) *on the boundary of $[-1/2, 1/2]^2$.*

The proof consists of three parts. In the first part we construct the functions w_δ; in the second part we study the symmetry properties and the support of the right-hand side of (2.5); and in the third part we show that this sequence has the asserted scaling.

A.1. Construction of w_δ. Let f_δ be given by

$$f_\varepsilon''(s) = \begin{cases} \dfrac{1}{4\delta}, & \operatorname{dist}(s,\mathbb{Z}) < \delta, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{with} \qquad f_\delta(0) = f_\delta'(0) = 0.$$

Note that f is even and that $f(1) = 1/4$. Define

$$w_\delta(x,y) = f_\delta(y+x) + f_\delta(y-x) - \frac{1}{2} f_\delta(2x) - \frac{1}{2} y^2.$$

We shall drop the subscript δ and simply write w and f.

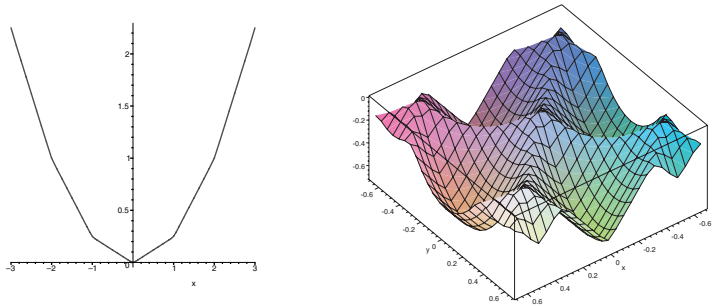


FIG. A.1. *The functions f and $-w$; on the right the plotting area is slightly larger than one period.*

The function w is periodic on $\mathbb{R}^2$ with period 1 in each direction. To show this, we prove that the first two derivatives match up on opposite sides of $[-1/2, 1/2] \times [-1/2, 1/2]$ as follows:

- By the symmetry of f, the function w is even in both x and y. Consequently w takes the same values on $(1/2, y)$ and $(-1/2, y)$; the same holds for $(x, \pm 1/2)$.
- For the comparison of the first derivatives, we calculate

$$\int_{-1/2}^{1/2} w_{xx}(x,y)\,dx = \int_{-1/2}^{1/2} f''(y+x)\,dx + \int_{-1/2}^{1/2} f''(y-x)\,dx - 2\int_{-1/2}^{1/2} f''(2x)\,dx = 0,$$

implying that $w_x(-1/2,y) = w_x(1/2,y)$; by the symmetry of w it follows that $w_x(-1/2,y) = w_x(1/2,y) = 0$. Similarly, using the definition of f_δ we find that

$$\int_{-1/2}^{1/2} w_{yy}(x,y)\,dy = \int_{-1/2}^{1/2} f''(y+x)\,dy + \int_{-1/2}^{1/2} f''(y-x)\,dy - 1 = 0,$$

implying that $w_y(x,-1/2) = w_y(x,1/2) = 0$.

Periodicity on $\mathbb{R}^2$ then follows from the remark that all second derivatives of w are periodic with period 1 in x and y.

A.2. Support, symmetry, and boundary conditions. Next we investigate the right-hand side of (2.5). We find

$$\begin{aligned}[w,w] + w_{xx} &= \big\{(f''(y+x) + f''(y-x) - 2f''(2x))(f''(y+x) + f''(y-x) - 1)\\ &\qquad - (f''(y+x) - f''(y-x))^2\big\}\\ &\qquad + (f''(y+x) + f''(y-x) - 2f''(2x))\\ &= 4f''(y+x)f''(y-x) - 2f''(2x)(f''(y+x) + f''(y-x)).\end{aligned}$$

This expression has a zero integral over $[-1/2,1/2]^2$. This follows from the periodicity of w,

$$\int_{-1/2}^{1/2}\int_{-1/2}^{1/2} w_{xx}w_{yy}\,dxdy = \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} w_{xy}^2\,dxdy, \tag{A.2}$$

by partial integration. More is true, however; we analyze the support of $[w,w] + w_{xx}$ in $[-1/2,1/2]^2$ in more detail.

The value of f'' is either $(4\delta)^{-1}$ or zero; in order to determine $[w,w] + w_{xx}$ it is therefore sufficient to calculate the measures of the pairwise intersections of the supports of $f''(y+x)$, $f''(y-x)$, and $f''(2x)$ as follows:

- The intersection of the supports of $f''(y+x)$ and $f''(y-x)$ has total area $4\delta^2$ (see Figure A.2).
- The intersection of the supports of $f''(y+x)$ and $f''(2x)$ also has total area $4\delta^2$ (see Figure A.3).

Since the support of $[w,w] + w_{xx}$ is concentrated on a discrete set of points, let us examine the behavior at one of these points. For small δ the support forms disjoint sets in $[-1/2,1/2]^2$, and we can restrict our attention to the origin alone.

If $|s| < 1/2$, then $f_\delta''(s)$ can be written as

$$f_\delta''(s) = \frac{1}{\delta}\,g\left(\frac{s}{\delta}\right),$$

where

$$g(\sigma) = \begin{cases} \dfrac{1}{4}, & |\sigma| < 1,\\ 0 & \text{otherwise.}\end{cases}$$

Therefore, as long as $|(x,y)| < 1/4$, then

$$\begin{aligned}&4f_\delta''(y+x)f_\delta''(y-x) - 2f_\delta''(2x)(f_\delta''(y+x) + f_\delta''(y-x))\\ &\qquad = \frac{4}{\delta^2}\,g\left(\frac{y+x}{\delta}\right)g\left(\frac{y-x}{\delta}\right) - \frac{2}{\delta^2}g\left(\frac{2x}{\delta}\right)\left[g\left(\frac{y+x}{\delta}\right) + g\left(\frac{y-x}{\delta}\right)\right]\\ &\qquad = \frac{1}{\delta^2}\,F\big(\delta^{-1}(x,y)\big),\end{aligned} \tag{A.3}$$

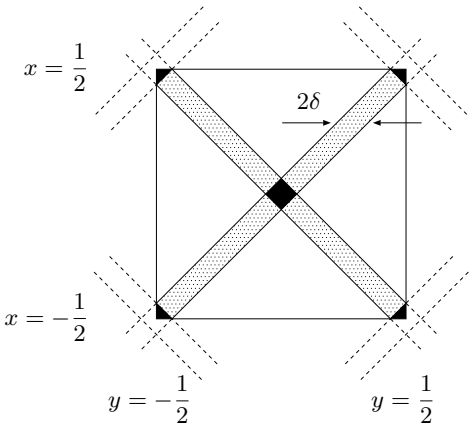


FIG. A.2. *The areas of the black regions add up to* $4\delta^2$.

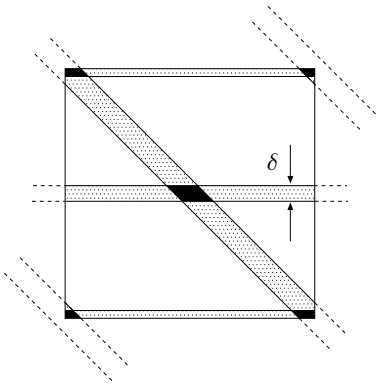


FIG. A.3. *The areas of the black regions add up to* $4\delta^2$.

where we introduce a new function F, which does not depend on δ, to summarize the line above. Note that $\operatorname{supp} F \subset [-2,2]^2$. Note also that by (A.2) the function F has zero integral; in addition, since f_δ'' is even, the function $4f_\delta''(y+x)f_\delta''(y-x) - 2f_\delta''(2x)(f_\delta''(y+x) + f_\delta''(y-x))$ is also even in x and in y. Therefore

$$\int_{\mathbb{R}^2} xF\big((x,y)\big)\,dxdy = \int_{\mathbb{R}^2} yF\big((x,y)\big)\,dxdy = 0.$$

This property will be used below.

The assertion also states that the functions w and ϕ satisfy (2.6) on the boundary of $[-1/2,1/2]^2$. We first note that w and ϕ are periodic in the following sense:

$$w(x \pm 1/2, y \pm 1/2) = w(x,y) \qquad \text{and} \qquad \phi(x \pm 1/2, y \pm 1/2) = \phi(x,y). \tag{A.4}$$

For w this is a simple consequence of the functional form of w; for ϕ it is a consequence of the uniqueness of solutions of (2.5) under periodic boundary conditions. The periodicity of w and ϕ in the y-direction in (2.6) then follows from a repeated application of (A.4). Similarly, the symmetry conditions in x in (2.6) follow from a combination of the symmetry of w and ϕ around $\{y = 0\}$ in combination with (A.4).

A.3. Scaling properties. We now use the information gathered above to show that the sequence w_δ has the scaling properties of (A.1). All function spaces are on $[-1/2,1/2]^2$.

First, f'_δ remains bounded on bounded sets as $\delta \to 0$; therefore $\int_\Omega w_{\delta x}^2$ converges to a finite, positive value. In addition, all second derivatives of w_δ remain bounded in L^1, so that

$$\|\Delta w_\delta\|_{L^1} \le C.$$

The second derivative f''_δ is bounded by $1/4\delta$, so that we can estimate

$$\|\Delta w_\delta\|_{L^2}^2 \le \|\Delta w_\delta\|_{L^1} \|\Delta w_\delta\|_{L^\infty} \le \frac{C}{\delta}.$$

Turning to ϕ_δ, we start by remarking that $[w_\delta, w_\delta] + w_{\delta xx}$ is bounded in L^1, since

$$\int_{|(x,y)|<1/4} \big|[w_\delta, w_\delta] + w_{\delta xx}\big| = \frac{1}{\delta^2} \int_{|(x,y)|<2\delta} \big|F\big(\delta^{-1}(x,y)\big)\big| = O(1).$$

Since $W^{2,p} \hookrightarrow L^\infty$ for all $p > 1$, the solution of

$$\Delta^2 \psi = h$$

satisfies

$$\|\Delta\psi\|_{L^{p'}} = \sup_\zeta \frac{\int_\Omega \Delta\psi\Delta\zeta}{\|\Delta\zeta\|_{L^p}} = \sup_\zeta \frac{\int_\Omega h\zeta}{\|\Delta\zeta\|_{L^p}} \le C\,\|h\|_{L^1} \frac{\|\zeta\|_{L^\infty}}{\|\zeta\|_{W^{2,p}}} \le C\,\|h\|_{L^1},$$

so that

$$\|\phi_\delta\|_{W^{2,p'}} \le C\,\|[w_\delta, w_\delta] + w_{\delta xx}\|_{L^1} \le C,$$

where $1/p + 1/p' = 1$. Using $W^{2,p'} \hookrightarrow C^{1,1-2/p'}$ we find

$$\|\phi_\delta\|_{C^{1,1-2/p'}} \le C\,\|\phi_\delta\|_{W^{2,p'}} \le C.$$

Writing, locally at the origin,

$$\phi_\delta(x,y) = \phi_\delta(0,0) + \nabla\phi_\delta(0,0)\cdot(x,y) + O\big(|(x,y)|^{2(1-1/p')}\big),$$

we find, by multiplying (2.5) by ϕ_δ and integrating,

$$\begin{aligned}
\|\Delta\phi_\delta\|^2_{L^2([-1/2,1/2]^2)} &= 2\int_{[-1/4,1/4]^2} \phi_\delta\big\{[w_\delta, w_\delta] + w_{\delta xx}\big\} \\
&= 2\frac{\phi_\delta(0,0)}{\delta^2} \int_{[-1/4,1/4]^2} F\big(\delta^{-1}(x,y)\big) \\
&\qquad + 2\frac{\nabla\phi_\delta(0,0)}{\delta^2} \cdot \int_{[-1/4,1/4]^2} (x,y) F\big(\delta^{-1}(x,y)\big) + O\big(\delta^{2(1-1/p')}\big) \\
&= O\big(\delta^{2(1-1/p')}\big),
\end{aligned}$$

since the zeroth and first moments of F are zero. Since p' may be chosen arbitrarily large, this estimate concludes the proof.

Appendix B. Derivation of the von Kármán–Donnell equations. The common aim of the many elastic shell theories is to approximate three-dimensional elasticity by a reduced description in which the unknowns are functions of not three but two spatial variables; see, for example, [6]. For the von Kármán–Donnell cylinder the central approximation is the *director Ansatz*, which states that a normal to the center surface remains normal through deformation. By this Ansatz the displacement is fully characterized by the displacement vector (u, v, w), a function of the two in-plane spatial variables x and y, where u, v, and w are the displacements in the axial (x-), tangential (y-), and radial directions, respectively. Apart from some rescaling, the function w is the same as the unknown w in the rest of this paper.

In the formulation of section 2 the unknowns u and v are replaced with the Airy stress function ϕ, which is derived by minimization with respect to the displacements u and v for fixed w. This minimization argument is well known in the context of the von Kármán plate theory and can be found in many textbooks. Determining the boundary conditions that the function ϕ satisfies, however, is not straightforward (see also the discussion in [27]), and it is for this reason that we now describe the argument in detail. The main goal is to show that the function ϕ is periodic in the tangential direction.

B.1. Energy and shortening. All quantities in this appendix are dimensional. We assume a cylinder of thickness t, length L, and radius R, and we set $\Omega = [0, L] \times [0, 2\pi R]$. The stored energy given by [31] is

$$E_1 = \frac{t^3 E}{24(1-\nu^2)} \int_\Omega \Delta w^2 + \frac{t}{2E} \int_\Omega \big[(\sigma_{11}+\sigma_{22})^2 - 2(1+\nu)(\sigma_{11}\sigma_{22} - \sigma_{12}^2)\big]$$

for a linear material of Young's modulus E and Poisson's ratio ν. Under the assumption of plane stress, the stress and strain tensors are related by

$$\sigma = \frac{E}{1-\nu^2}\big[(1-\nu)\varepsilon + \nu \operatorname{tr}\varepsilon\, I\big] = \frac{E}{1-\nu^2}\begin{pmatrix} \varepsilon_{11}+\nu\varepsilon_{22} & (1-\nu)\varepsilon_{12} \\ (1-\nu)\varepsilon_{12} & \varepsilon_{22}+\nu\varepsilon_{11} \end{pmatrix}, \tag{B.1}$$

and under the small-angle approximation the strain tensor can be expressed in the displacements as

$$\varepsilon = \begin{pmatrix} u_x + \frac{1}{2}w_x^2 & \frac{1}{2}u_y + \frac{1}{2}v_x + \frac{1}{2}w_x w_y \\ \frac{1}{2}u_y + \frac{1}{2}v_x + \frac{1}{2}w_x w_y & v_y + \frac{1}{2}w_y^2 - \rho w \end{pmatrix}. \tag{B.2}$$

These choices for the energy and for the stress and strain tensors are very similar to those for a flat plate. The intrinsic curvature of the cylinder, of magnitude $\rho = 1/R$, appears only in the last term of ε_{22}, $-\rho w$, which expresses the fact that radial displacement creates extensional strain in the y-direction.

The average axial shortening is given by

$$S_1 = -\frac{1}{2\pi R}\int_0^{2\pi R} \big[u(L,y) - u(0,y)\big]\, dy = -\frac{1}{2\pi R}\int_\Omega u_x \, dxdy,$$

and an equilibrium (u, v, w) at load level P is a stationary point of the total potential $V_1 = E_1 - PS_1$.

B.2. Boundary conditions. At the boundaries $y = 0, 2\pi R$ it is natural to assume that u, v, and w are periodic, but at $x = 0, L$ there is a certain amount of choice.

The boundary conditions on w (see (2.6a)) are

$$w_x = (\Delta w)_x = 0 \qquad \text{at } x = 0, L, \tag{B.3}$$

and these conditions signify a fixed angle ($w_x = 0$) and zero radial force ($(\Delta w)_x = 0$). They may also be understood as symmetry boundary conditions, as in the case of a sequence of cylinders stacked on top of each other. For u and v we assume boundary conditions

$$u_y = 0 \qquad \text{and} \qquad \sigma_{12} = 0 \qquad \text{at } x = 0, L, \tag{B.4}$$

which signify that the ends of the cylinder are rigid in the x-direction and that there is no friction between the cylinder and the apparatus holding it. Note that the pair of boundary conditions $\sigma_{12} = 0$ and $(\Delta w)_x = 0$ together states that the loading apparatus exerts only axial forces on the cylinder.

The boundary conditions on w are invariant under the addition of a constant to w, i.e., under the replacement of w with $w + c$; for stationary points we may exploit this fact.

LEMMA B.1. *If (u, v, w) is a stationary point of $E_1 - PS_1$ under boundary conditions* (B.3)–(B.4)*, then*

$$\int_\Omega \sigma_{22} = 0. \tag{B.5}$$

Proof. Under the replacement $w \mapsto w + c$, we have

$$\frac{d\sigma}{dc} = -\frac{E\rho}{1-\nu^2}\begin{pmatrix} \nu & 0 \\ 0 & 1 \end{pmatrix},$$

and therefore

$$\begin{aligned} 0 &= \frac{d}{dc}(E_1 - PS_1) \\ &= -\frac{t\rho}{1-\nu^2}\int_\Omega \big[(\sigma_{11} + \sigma_{22})(\nu + 1) - (1 + \nu)(\sigma_{11} + \nu\sigma_{22})\big] = -t\rho \int_\Omega \sigma_{22}. \end{aligned}$$

∎

B.3. Derivation of the Airy stress function ϕ. The energy (2.7) and the Airy stress function ϕ are derived from the total potential $E_1 - PS_1$ by minimization with respect to the displacements u and v for fixed w. Performing this minimization on the second term in E_1 yields the classical plate equilibrium equations

$$\sigma_{11x} + \sigma_{12y} = 0 \qquad \text{and} \qquad \sigma_{12x} + \sigma_{22y} = 0.$$

Note that the derivative of S_1 with respect to u only creates boundary terms. By applying three times the well-known characterization of divergence-free vector fields as rotations of scalar fields (see, e.g., [3, Thm. XII.3.5]) we obtain the *local* existence of a function ϕ satisfying

$$\sigma_{11} = E\phi_{yy}, \qquad \sigma_{12} = -E\phi_{xy}, \qquad \text{and} \qquad \sigma_{22} = E\phi_{xx}, \tag{B.6}$$

where we use the traditional scaling of ϕ by Young's modulus.

B.4. Boundary conditions on ϕ. The existence of the function ϕ is the result of a local differential-geometric argument, and as such gives no reason for ϕ to be periodic in y. The following theorem shows that after a normalization transformation, the function ϕ can indeed be assumed to be periodic in y, and may be taken to satisfy the same boundary conditions as the function w.

THEOREM B.2. *If u, v, and w are periodic in y and satisfy boundary conditions* (B.3)–(B.4)*, then there exists a function ϕ that satisfies*

$$\sigma_{11} - \frac{1}{|\Omega|}\int_\Omega \sigma_{11} = E\phi_{yy}, \qquad \sigma_{12} = -E\phi_{xy}, \qquad \text{and} \qquad \sigma_{22} = E\phi_{xx}; \tag{B.7}$$

is periodic in y; and satisfies boundary conditions

$$\phi_x = (\Delta\phi)_x = 0 \qquad \text{at } x = 0, L. \tag{B.8}$$

Remark B.3. Mechanically the normalization of ϕ with $\int_\Omega \sigma_{11}$ means that ϕ represents the *deviation* from the unbuckled in-plane stress state.

Proof. As discussed above, there exists a function ϕ satisfying (B.6); we will construct in stages a new function $\hat{\phi}$ which satisfies (B.7) and the boundary conditions.

We first convert condition (B.6) into (B.7). Set

$$p(x) := \frac{1}{2\pi R}\int_0^{2\pi R} \phi_{yy}(x,y)\,dy = \frac{1}{2\pi R}\big[\phi_y(x,2\pi R) - \phi_y(x,0)\big].$$

Since the second derivatives of ϕ can be expressed in terms of derivatives of u, v, and w, the second and higher derivatives of ϕ are automatically periodic in y. Therefore

$$\frac{d}{dx}p(x) = \frac{1}{2\pi R}\big[\phi_{xy}(x,2\pi R) - \phi_{xy}(x,0)\big] = 0,$$

implying that p is actually independent of x. (A mechanical argument provides the same result: $Etp(x) = t\int_0^{2\pi R}\sigma_{11}(x,y)dy$ is the total force applied at a virtual cut at level x, and mechanical equilibrium implies that this force is independent of x.) Therefore

$$|\Omega|\,p = 2\pi R\int_0^L p\,dx = \int_\Omega \phi_{yy} = \frac{1}{E}\int_\Omega \sigma_{11},$$

so that the new function

$$\tilde{\phi}(x,y) := \phi(x,y) - \frac{p}{2}y^2$$

satisfies (B.7). Note that this implies

$$\int_\Omega \tilde{\phi}_{yy} = 0. \tag{B.9}$$

We now turn to the periodicity in the y-direction. It remains to show that $\tilde{\phi}$, $\tilde{\phi}_x$, and $\tilde{\phi}_y$ are the same at $y = 0$ and $y = 2\pi R$. Again the periodicity of the second derivatives implies that

$$\frac{d^2}{dx^2}\big[\tilde{\phi}(x,2\pi R) - \tilde{\phi}(x,0)\big] = \tilde{\phi}_{xx}(x,2\pi R) - \tilde{\phi}_{xx}(x,0) = 0,$$

so that $\tilde{\phi}(x, 2\pi R) - \tilde{\phi}(x, 0) = ax + b$ for some $a, b \in \mathbb{R}$. Defining

$$\hat{\phi}(x, y) := \tilde{\phi}(x, y) - \frac{by}{2\pi R} = \phi(x, y) - \frac{p}{2}y^2 - \frac{by}{2\pi R},$$

the function $\hat{\phi}$ still satisfies (B.7), and

$$\hat{\phi}(x, 2\pi R) - \hat{\phi}(x, 0) = ax.$$

Finally, we find that

$$a = \hat{\phi}_x(0, 2\pi R) - \hat{\phi}_x(0, 0) = \int_0^{2\pi R} \hat{\phi}_{xy}(0, y)\, dy = \frac{1}{E}\int_0^{2\pi R} \sigma_{12}(0, y)\, dy \overset{(B.4)}{=} 0,$$

and therefore that $\hat{\phi}(x, 2\pi R) - \hat{\phi}(x, 0) = 0$ for all x. The same follows for $\hat{\phi}_x(x, 2\pi R) - \hat{\phi}_x(x, 0)$ by differentiation.

To show that ϕ_y also matches,

$$\frac{d}{dx}\big[\hat{\phi}_y(x, 2\pi R) - \hat{\phi}_y(x, 0)\big] = \big[\phi_{xy}(x, 2\pi R) - \phi_{xy}(x, 0)\big] = 0,$$

and therefore $\phi_y(x, 2\pi R) - \phi_y(x, 0)$ is constant in x; by (B.9) this constant is zero. This proves that $\hat{\phi}$ satisfies (B.7) and is periodic in y.

We finally discuss the boundary conditions at $x = 0, L$, and we follow the line of reasoning of [27]. By (B.4) and (B.1), $\varepsilon_{12} = 0$ at $x = 0, L$, so that by (B.3) and (B.4),

$$v_{xy} = \frac{\partial}{\partial y}(2\varepsilon_{12} - u_y - w_x w_y) = 0 \qquad \text{at } x = 0, L.$$

Therefore

$$\varepsilon_{22x} = v_{xy} + w_y w_{xy} - \rho w_x \overset{(B.4)}{=} 0 \qquad \text{at } x = 0, L.$$

Using $E\varepsilon_{22} = \sigma_{22} - \nu\sigma_{11}$, we then find

$$\hat{\phi}_{xxx} - \nu\hat{\phi}_{xyy} = \phi_{xxx} - \nu\phi_{xyy} = \frac{1}{E}\frac{d}{dx}(\sigma_{22} - \nu\sigma_{11}) = \varepsilon_{22x} = 0,$$

and by adding $(1 + \nu)\hat{\phi}_{xyy} = -(1/E)(1 + \nu)\sigma_{12y} = 0$ it follows that

$$(\Delta\hat{\phi})_x = \hat{\phi}_{xxx} + \hat{\phi}_{xyy} = 0 \qquad \text{at } x = 0, L,$$

which proves one part of (B.8).

From $\hat{\phi}_{xy} = -\sigma_{12}/E = 0$ we find that

$$\hat{\phi}_x(0, y) = c_0 \qquad \text{and} \qquad \hat{\phi}_x(L, y) = c_L \qquad \text{for all } y \in [0, 2\pi R].$$

Writing

$$2\pi R(c_L - c_0) = \int_\Omega \hat{\phi}_{xx} = \frac{1}{E}\int_\Omega \sigma_{22} \overset{(B.5)}{=} 0$$

we find that $c_L = c_0$. Now the function

$$\overline{\phi}(x, y) := \hat{\phi}(x, y) - c_0 x = \phi(x, y) - \frac{p}{2}y^2 - \frac{by}{2\pi R} - c_0 x$$

satisfies (B.7) and (B.8) and is periodic in y. This concludes the proof. ∎

Remark B.4. It is instructive to note that the periodicity of ϕ is a result of the specific choice of boundary conditions, and will in fact not hold if different boundary conditions are taken. For instance, if a tangential shear stress τ is applied at the cylinder ends (i.e., the cylinder is loaded under torsion), then the coefficient a in the derivation above will not vanish, and ϕ_x will not be periodic in y.

B.5. Putting it all together. By an elementary but lengthy calculation we find that ϕ, as provided by Theorem B.2, satisfies the equation

$$\Delta^2\phi + \rho w_{xx} + [w,w] = 0 \qquad \text{in } \Omega, \tag{B.10}$$

and that the second term in E_1 can be written as

$$\frac{tE}{2}\int_\Omega \Big[\Delta\phi^2 - 2(1+\nu)[\phi,\phi]\Big].$$

By the boundary conditions given by Theorem B.2 the second term vanishes, and the total stored energy functional can therefore be written as

$$E_2(w) := \frac{t^3E}{24(1-\nu^2)}\int_\Omega \Delta w^2 + \frac{tE}{2}\int_\Omega \Delta\phi^2.$$

Note that this energy is a function of w alone; the function ϕ in this definition is assumed to be given by (B.10), with the boundary conditions of Theorem B.2. Similarly, we rewrite the average shortening as

$$\begin{aligned} S_2(w) := S_1(u) &= -\frac{1}{2\pi R}\int_\Omega u_x \\ &\overset{(B.2)}{=} -\frac{1}{2\pi R}\int_\Omega \left[\varepsilon_{11} - \frac{1}{2}w_x^2\right] \\ &\overset{(B.1)}{=} -\frac{1}{2\pi RE}\int_\Omega \big[\sigma_{11} - \nu\sigma_{22}\big] + \frac{1}{4\pi R}\int_\Omega w_x^2 \\ &\overset{(B.7)}{=} -\frac{1}{2\pi R}\int_\Omega \big[\phi_{yy} - \nu\phi_{xx}\big] + \frac{1}{4\pi R}\int_\Omega w_x^2 \\ &= \frac{1}{4\pi R}\int_\Omega w_x^2. \end{aligned}$$

A stationary point w of $E_2 - PS_2$ satisfies the Euler equation

$$\frac{t^2}{12(1-\nu^2)}\Delta^2 w + \frac{P}{2\pi REt}w_{xx} - \rho\phi_{xx} - 2[w,\phi] = 0,$$

where again ϕ is related to w by (B.10). With the nondimensionalization

$$w = 4\pi^2 R\,\overline{w}, \qquad \phi = 16\pi^4R^2\,\overline{\phi}, \qquad x \mapsto 2\pi R\,x, \qquad y \mapsto 2\pi R\,y,$$

we then obtain (2.4) and (2.5), and the dimensional energy E_2 and average shortening S_2 above can be expressed in these variables as

$$E_2 = \frac{\pi^2t^3E}{6(1-\nu^2)}\int \Delta\overline{w}^2 + 32\pi^6tER^2\int \Delta\overline{\phi}^2, \qquad S_2 = 4\pi^3R\int \overline{w}_x^2. \tag{B.11}$$

REFERENCES

[1] B. O. Almroth, *Postbuckling behaviour of axially compressed circular cylinders*, AIAA J., 1 (1963), pp. 627–633.

[2] A. Ambrosetti and P. H. Rabinowitz, *Dual variational methods in critical point theory and applications*, J. Funct. Anal., 14 (1973), pp. 349–381.

[3] S. S. Antman, *Nonlinear Problems of Elasticity*, 2nd ed., Appl. Math. Sci. 107, Springer, New York, 2005.

[4] J. Arbocz and C. D. Babcock, *The effect of general imperfections on the buckling of cylindrical shells*, ASME J. Appl. Mech., 36 (1969), pp. 28–38.

[5] W. Ballerstedt and H. Wagner, *Versuche über die Festigkeit dünner unversteifter Zylinderunter Schub- und Längskräften*, Luftfahrtforschung, 13 (1936), pp. 309–312.

[6] Z. P. Bažant and L. Cedolin, *Stability of Structures: Elastic, Inelastic, Fracture and Damage Theories*, Oxford University Press, New York, 1991.

[7] F. J. Bridget, C. C. Jerome, and A. B. Vosseller, *Some new experiments on buckling of thin-wall construction*, Trans. ASME Aero. Eng., 56 (1934), pp. 569–578.

[8] Y. S. Choi and P. J. McKenna, *A mountain pass method for the numerical solution of semilinear elliptic problems*, Nonlinear Anal., 20 (1993), pp. 417–437.

[9] M. Deml and W. Wunderlich, *Direct evaluation of the worst imperfection shape in shell buckling*, Comput. Methods Appl. Mech. Engrg., 149 (1997), pp. 201–222.

[10] L. H. Donnell, *A new theory for buckling of thin cylinders under axial compression and bending*, Trans. ASME Aero. Eng., 56 (1934), pp. 795–806.

[11] H. Duddeck, B. Kröplin, D. Dinkler, J. Hillmann, and W. Wagenhuber, *Nonlinear computations in civil engineering structures*, in DFG Colloquium, Springer-Verlag, Berlin, 1989 (in German).

[12] M. Eßlinger, *Hochgeschwindigkeitsaufnahmen vom Beulvorgang dünnwandiger, axialbelasteter Zylinder*, Der Stahlbau, 39 (1970), pp. 73–76.

[13] P. R. Everall and G. W. Hunt, *Mode jumping in the buckling of struts and plates: A comparative study*, Int. J. Non-Linear Mech., 35 (2000), pp. 1067–1079.

[14] D. J. Gorman and R. M. Evan-Iwanowski, *An analytical and experimental investigation of the effects of large prebuckling deformations on the buckling of clamped thin-walled circular cylindrical shells subjected to axial loading and internal pressure*, Develop. in Theor. and Appl. Mech., 4 (1970), pp. 415–426.

[15] N. Hoff, W. A. Madsen, and J. Mayers, *Post-buckling equilibrium of axially compressed circular cylindrical shells*, AIAA J., 4 (1966), pp. 126–133.

[16] J. Horák, *Constrained mountain pass algorithm for the numerical solution of semilinear elliptic problems*, Numer. Math., 98 (2004), pp. 251–276.

[17] J. Horák, G. J. Lord, and M. A. Peletier, *Numerical Variational Methods Applied to Cylinder Buckling*, in preparation.

[18] J. Horák and P. J. McKenna, *Traveling waves in nonlinearly supported beams and plates*, in Nonlinear Equations: Methods, Models and Applications (Bergamo, 2001), Progr. Nonlinear Differential Equations Appl. 54, Birkhäuser, Basel, 2003, pp. 197–215.

[19] C. Hühne, R. Zimmerman, R. Rolfes, and B. Geier, *Sensitivities to geometrical and loading imperfections on buckling of composite cylindrical shells*, in Proceedings of the European Conference on Spacecraft Structures, Materials and Mechanical Testing, Toulouse, 2002.

[20] G. W. Hunt and E. Lucena Neto, *Localized buckling in long axially-loaded cylindrical shells*, J. Mech. Phys. Solids, 39 (1991), pp. 881–894.

[21] G. W. Hunt, M. A. Peletier, A. R. Champneys, P. D. Woods, M. A. Wadee, C. J. Budd, and G. J. Lord, *Cellular buckling in long structures*, Nonlinear Dynam., 21 (2000), pp. 3–29.

[22] W. T. Koiter, *On the Stability of Elastic Equilibrium*, Ph.D. thesis, Technische Hogeschool, Delft (Technological University of Delft), Holland, 1945; English translation issued as Tech. report NASA-TT-F-10833, NASA Center for Aerospace Information, Hanover, MD, 1967.

[23] B. Kröplin, D. Dinkler, and J. Hillmann, *An energy perturbation method applied to nonlinear structural analysis*, Comput. Methods Appl. Engrg., 52 (1985), pp. 885–897.

[24] G. J. Lord, A. R. Champneys, and G. W. Hunt, *Computation of localized post buckling in long axially-compressed cylindrical shells*, Philos. Trans. Roy. Soc. Lond. Ser. A, 355 (1997), pp. 2137–2150.

[25] G. J. Lord, A. R. Champneys, and G. W. Hunt, *Computation of homoclinic orbits in partial differential equations: An application to cylindrical shell buckling*, SIAM J. Sci. Comput., 21 (1999), pp. 591–619.

[26] M. A. Peletier, *Sequential buckling: A variational analysis*, SIAM J. Math. Anal., 32 (2001), pp. 1142–1168.

[27] D. Schaeffer and M. Golubitsky, *Boundary conditions and mode jumping in the buckling of a rectangular plate*, Commun. Math. Phys., 69 (1979), pp. 209–236.
[28] D. Smets and G. J. B. van den Berg, *Homoclinic solutions for Swift-Hohenberg and suspension bridge type equations*, J. Differential Equations, 184 (2002), pp. 78–96.
[29] M. Struwe, *Variational Methods*, Springer-Verlag, Berlin, 1990.
[30] R. C. Tennyson, *An Experimental Investigation of the Buckling of Circular Cylindrical Shells in Axial Compression Using the Photoelastic Technique*, Tech. Report 102, University of Toronto, Toronto, ON, Canada, 1964.
[31] T. von Kármán and H. S. Tsien, *The buckling of thin cylindrical shells under axial compression*, J. Aeronautical Sci., 8 (1941), pp. 303–312.
[32] W. Wagenhuber and H. Duddeck, *Numerischer Stabilitätsnachweis dünner Schalen mit dem Konzept der Störenergie*, Arch. Appl. Mech., 61 (1991), pp. 327–343.
[33] V. I. Weingarten, E. J. Morgan, and P. Seide, *Elastic stability of thin-walled cylindrical and conical shells under axial compression*, AIAA J., 3 (1965), pp. 500–505.
[34] T. A. Winterstetter and H. Schmidt, *Stability of circular cylindrical steel shells under combined loading*, Thin-Walled Structures, 40 (2002), pp. 893–909.
[35] W. Wunderlich and U. Albertin, *Analysis and load carrying behaviour of imperfection sensitive shells*, Int. J. Numer. Methods Engrg., 47 (2000), pp. 255–273.
[36] N. Yamaki, *Elastic Stability of Circular Cylindrical Shells*, Appl. Math. Mech. 27, North–Holland, Amsterdam, 1984.
[37] Y. Yoshimura, *On the Mechanism of Buckling of a Circular Shell under Axial Compression*, Tech. Report 1390, NACA, Washington, DC, 1955.
[38] E. Zhu, P. Mandal, and C. R. Calladine, *Buckling of thin cylindrical shells: An attempt to resolve a paradox*, Int. J. Mech. Sci., 44 (2002), pp. 1583–1601.

SIAM J. APPL. MATH.
Vol. 66, No. 5, pp. 1825–1837

CONVERGENCE OF STRONG SHOCK IN A VAN DER WAALS GAS*

RAJAN ARORA†‡ AND V. D. SHARMA†

Abstract. Strong cylindrical and spherical shock waves, collapsing at the center (or axis) of symmetry, are studied for a Van der Waals gas. The perturbation technique applied in this paper provides a global solution to the implosion problem, yielding the results for Guderley's local self-similar solution, which is valid only in the vicinity of the center/axis of implosion. The similarity exponents are found along with the corresponding amplitudes in the vicinity of the shock-collapse. The flow parameters and the shock trajectory have been computed for different values of the adiabatic coefficient and the Van der Waals excluded volume.

Key words. shock waves, convergence, Van der Waals gas

AMS subject classifications. 35L60, 35L67, 74J30, 76L05, 76N15

DOI. 10.1137/050634402

1. Introduction.

The study of shock waves is motivated by its application in a variety of fields such as aerodynamics, astrophysics, nuclear science, and plasma physics, and shock waves generated by spherical and cylindrical pistons in a gas have received much attention in the past four decades. Converging shock waves have been a field of growing interest since the early 1940s from both mathematical and physical points of view. They are one of the means of generating high pressure and high temperatures at the center/axis of convergence. Applications range from nuclear weapons to plasmas; in medical science, shock wave lithotripsy is used to treat kidney stone disease. This work has been motivated by the fact that such shocks are an essential part of the mechanism responsible for sonoluminescence, that is, the light which under certain conditions is emitted from a bubble of gas trapped in a liquid and compressed by incident sound waves.

A theoretical investigation of shock wave behavior near the center of convergence was first done by Guderley in 1942 [1]. We also note the work of Lazarus and Richtmyer [2], Van Dyke and Guttmann [3], Hafner [4], Wu and Roberts [5], and Madhumita and Sharma [6] as major contributions towards the investigation of the implosion problem.

Wu and Roberts have found that for small values of the Van der Waals excluded volume (b), there is only one branch of similarity solutions that resembles the Guderley solutions and is well described by the CCW (Chester–Chisnell–Whitham)approximation [7]; however, for larger values of b, they predict another branch of the solution, which is distinct from the Guderley-type branch and which cannot be obtained by the CCW approximation.

In the present paper, we successfully apply the technique of Van Dyke and Guttmann [3] to the shock implosion problem in a Van der Waals gas, which provides a global solution to the imploding shock problem. This global solution is valid

*Received by the editors June 24, 2005; accepted for publication (in revised form) April 28, 2006; published electronically August 22, 2006. This work was supported by CSIR, India vide reference number 9/87(235)/99-EMR-I and project grant 25(0122)/02/EMR-II, and ISRO-IITB STC (via project 05-IS001).

http://www.siam.org/journals/siap/66-5/63440.html

†Department of Mathematics, Indian Institute of Technology, Bombay, Powai, Mumbai-400076, India (rajan@math.iitb.ac.in, vsharma@math.iitb.ac.in).

‡Current address: BITS PILANI-GOA CAMPUS, Zuarinagar-403726, Goa, India.

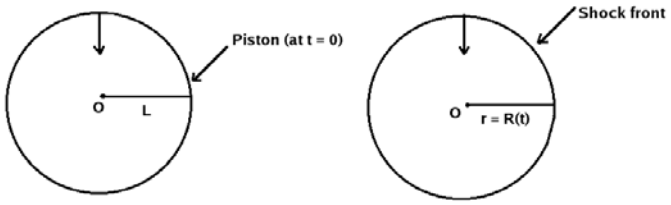


FIG. 1. *The spherical piston is initially of radius L. At time $t = 0$ the piston begins to contract with a very large constant velocity V, driving ahead of it a spherical shock of radius $R(t)$ collapsing at the point of implosion O, where $R(t)$ is to be determined.*

throughout the flow field almost up to the instant of collapse. From this global solution we are able to extract with good accuracy the Guderley local solution at the instant of shock collapse; this is in excellent agreement with a solution branch obtained by Wu and Roberts. Wu and Roberts reported that a second solution branch may exist within the range of the first solution branch, and for certain values of γ and b the similarity solution may cease to exist; for instance, it is reported that for $\gamma = 5/3$, one solution arises in the range $0 \leq b \leq 0.05$, while the other solution exists in the range $0.005 \leq b \leq 0.035$, and no similarity solution could be obtained in the range $0.0001 \leq b \leq 0.05$ for $\gamma = 6/5$. The present method does not exhibit any computational difficulty, enables us to compute successfully the similarity exponent for all such values of b and γ, and yields only one solution that matches well with the Guderley solution.

Guderley's local solution provides us only the first dominant similarity exponent, but we have found, through this technique, the other similarity exponents and the corresponding amplitudes.

We consider a spherical or cylindrical piston that is filled with a Van der Waals gas of constant density. The piston collapses with constant inward speed, greater than the speed of sound, generating a strong shock wave collapsing at the center/axis of symmetry; see Figure 1.

Computations are performed to obtain the similarity exponents and the corresponding amplitudes in the vicinity of the shock-collapse. Computations are carried out for different values of adiabatic coefficient γ and Van der Waals excluded volume b. Distributions of the flow variables at the rear of the shock wave are presented. The results are given for cylindrical and spherical fronts propagating into the medium. Our results match well with those obtained by Wu and Roberts [5] and Guderley's method [1].

2. Basic equations with solutions analytic in time. We consider a cylindrical or spherical piston of initial radius L and filled with Van der Waals gas. Let the initial conditions be given by

$$v = 0, \quad \rho = \rho_0, \quad p = p_0, \tag{1}$$

where v, ρ, and p are, respectively, the outward radial velocity, density, and pressure; ρ_0 and p_0 are appropriate positive constants. At time $t = 0$, the container starts to contract with very large constant velocity V; this produces a cylindrical or spherical shock wave whose radius $R(t)$ is to be determined. The equations of one-dimensional

adiabatic motion of a Van der Waals gas are (Wu and Roberts [5])

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho v)}{\partial r} + \frac{m\rho v}{r} = 0, \tag{2}$$

$$\frac{\partial v}{\partial t} + v\,\frac{\partial v}{\partial r} + \frac{1}{\rho}\,\frac{\partial p}{\partial r} = 0, \tag{3}$$

$$\frac{\partial p}{\partial t} + v\,\frac{\partial p}{\partial r} - \frac{\gamma p}{\rho(1-b\rho)}\left(\frac{\partial \rho}{\partial t} + v\,\frac{\partial \rho}{\partial r}\right) = 0, \tag{4}$$

where t is the time, r is the distance of the particle from the center of symmetry, b is the Van der Waals excluded volume, γ is the adiabatic exponent, and m takes values 1 and 2 depending on whether the piston is cylindrical or spherical. The Rankine–Hugoniot (R-H) conditions just behind the shock wave are given by (Wu and Roberts [5])

$$v = \frac{2(1-b)}{(\gamma+1)}\dot{R}, \quad \rho = \frac{(\gamma+1)}{(\gamma-1+2b)}\,\rho_0, \quad p = \frac{2(1-b)\rho_0}{(\gamma+1)}\,(\dot{R})^2 \quad \text{at} \quad r = R(t). \tag{5}$$

The condition of no flow through the piston yields

$$v = -V \qquad \text{at} \quad r = L - Vt. \tag{6}$$

For our convenience we measure the distance $x = L - r$ inward, and let $u = -v$ be the corresponding inward velocity. Making use of the conical nature of the basic flow, we introduce a new variable

$$z = \frac{2}{(\gamma-1)}\left(\frac{x}{Vt} - k\right) \tag{7}$$

that varies from $-2(k-1)/(\gamma-1)$ at the piston to unity at the basic position of the shock wave, where the constant k is to be determined.

Now we nondimensionalize the variables ρ, u, p, b, x, and t by referring lengths to L, speed to V, density to ρ_0, pressure to $\rho_0 V^2$, Van der Waals excluded volume b to $1/\rho_0$, and time to L/V. Then the differential equations (2), (3), and (4) become

$$\left(1 - \left(k + \frac{1}{2}(\gamma-1)z\right)t\right)\left(\rho\frac{\partial u}{\partial z} + \left(u - k - \frac{1}{2}(\gamma-1)z\right)\frac{\partial \rho}{\partial z} + \frac{1}{2}(\gamma-1)t\frac{\partial \rho}{\partial t}\right) = \frac{1}{2}(\gamma-1)mt\rho u, \tag{8}$$

$$\rho\left(u - k - \frac{1}{2}(\gamma-1)z\right)\frac{\partial u}{\partial z} + \frac{1}{2}(\gamma-1)t\rho\frac{\partial u}{\partial t} + \frac{\partial p}{\partial z} = 0, \tag{9}$$

$$\left(u - k - \frac{1}{2}(\gamma-1)z\right)\left(\rho(1-b\rho)\frac{\partial p}{\partial z} - \gamma p\frac{\partial \rho}{\partial z}\right) + \frac{1}{2}(\gamma-1)t\left(\rho(1-b\rho)\frac{\partial p}{\partial t} - \gamma p\frac{\partial \rho}{\partial t}\right) = 0; \tag{10}$$

the boundary conditions (5) and (6) become

$$u = \frac{2(1-b)}{\gamma+1}\dot{X}, \qquad \rho = \frac{(\gamma+1)}{(\gamma-1+2b)}, \qquad p = \frac{2(1-b)}{(\gamma+1)}(\dot{X})^2 \tag{11}$$

at $z = (2/(\gamma-1))\,(X/t-k)$; and

$$u = 1 \qquad \text{at} \quad z = \frac{-2(k-1)}{(\gamma-1)}. \tag{12}$$

Assuming that the solution is analytic in time, we expand the unknown position of the shock wave in a Taylor series as

$$X(t) = \sum_{n=1}^{\infty} X_n t^n, \tag{13}$$

and similarly we expand the flow variables as

$$u = \sum_{n=1}^{\infty} U_n(z) t^{n-1}, \qquad \rho = \sum_{n=1}^{\infty} R_n(z) t^{n-1}, \qquad p = \sum_{n=1}^{\infty} P_n(z) t^{n-1}. \tag{14}$$

Substituting (14) into (8), (9), (10) and using (11), (12), (13), we find, on equating the terms on both the sides which are independent of t,

$$\begin{gathered} U_1 = 1, \qquad R_1 = \frac{(\gamma+1)}{(\gamma-1+2b)}, \qquad P_1 = \frac{(\gamma+1)}{2(1-b)}, \\ X_1 = \frac{(\gamma+1)}{2(1-b)}, \qquad k = \frac{2+b(\gamma-1)}{2(1-b)}. \end{gathered} \tag{15}$$

The coefficients U_2, R_2, and P_2 for the second approximation satisfy the following first-order linear ordinary differential equations:

$$R_1 U_2' - \left(k - 1 + \frac{1}{2}(\gamma-1)z\right) R_2' + \frac{1}{2}(\gamma-1) R_2 = \frac{1}{2}(\gamma-1) m R_1, \tag{16}$$

$$\frac{1}{2}(\gamma-1) R_1 U_2 - \left(k - 1 + \frac{1}{2}(\gamma-1)z\right) R_1 U_2' + P_2' = 0, \tag{17}$$

$$\begin{aligned} &\frac{(\gamma-1)}{2}\Big((1-bR_1)R_1 P_2 - \gamma P_1 R_2\Big) \\ &- \left(k - 1 + \frac{1}{2}(\gamma-1)z\right)\Big((1-bR_1)R_1 P_2' - \gamma P_1 R_2'\Big) = 0, \end{aligned} \tag{18}$$

where the prime denotes the derivative with respect to the variable z and where U_1, R_1, and P_1 are given in (15).

The boundary condition (12) on the piston yields

$$U_n(z) = 0 \qquad \text{at} \quad z = \frac{-2(k-1)}{\gamma-1} \qquad \text{for} \quad n = 2, 3, 4, \ldots . \tag{19}$$

Also the boundary conditions (11) give

$$U_2(1) = \frac{4(1-b)}{(\gamma+1)} X_2, \qquad R_2(1) = 0, \qquad P_2(1) = 4X_2. \tag{20}$$

From (16)–(19) we find that

$$U_2''(z) = R_2''(z) = P_2''(z) = 0, \tag{21}$$

which together with the boundary conditions (19) and (20) yields

$$
\begin{aligned}
U_2 &= \frac{\gamma m}{2(2\gamma-1)}\left(\frac{b(\gamma+1)}{(1-b)}+(\gamma-1)z\right),\\
R_2 &= \frac{(1-b)(\gamma-1)^2\ (\gamma+1)m}{(\gamma-1+2b)^2\ (2\gamma-1)}(1-z),\\
P_2 &= \frac{\gamma(\gamma+1)(\gamma-1+2b)m}{2(1-b)^2\ (2\gamma-1)},\\
X_2 &= \frac{\gamma(\gamma+1)(\gamma-1+2b)m}{8(1-b)^2\ (2\gamma-1)}.
\end{aligned}
\tag{22}
$$

The forms of U_2, R_2, and P_2 suggest that in higher approximations the coefficients U_n, R_n, and P_n are polynomials in z of degree $n-1$, of the following form:

$$
U_n(z)=\sum_{j=1}^{n} U_{nj}z^{j-1},\quad R_n(z)=\sum_{j=1}^{n} R_{nj}z^{j-1},\quad P_n(z)=\sum_{j=1}^{n} P_{nj}z^{j-1}. \tag{23}
$$

Substituting these values into the differential equations (8)–(10) and the shock conditions (11), and equating the like powers of z as well as t, we have for the nth approximation a system of $3n+1$ linear algebraic equations in $3n+1$ coefficients U_{nj}, R_{nj}, P_{nj} $(j=1,\ 2,\ldots,n)$ and X_n; solving the system for the third approximation we obtain the position of the shock wave

$$
\begin{aligned}
X(t) = t\ &(\gamma+1)/[2(1-b)] + t^2 m\gamma(\gamma+1)(\gamma-1+2b)/[8(1-b)^2(2\gamma-1)]\\
&+ t^3 m(\gamma+1)(\gamma-1+2b)[1+\gamma^3(4-21m)-\gamma m+\gamma^2(13m-9)\\
&\qquad +\gamma^4(12+13m)+b(6\gamma-20\gamma^2 m-2(m+1)+\gamma^3(26m-8))]\\
&/[48(1-b)^3(2\gamma-1)^2(7\gamma-5)]+\cdots\ .
\end{aligned}
\tag{24}
$$

The above result fully recovers the ideal gas ($b=0$) case discussed by Van Dyke and Guttmann [3].

3. Computation of coefficients X_n, U_n, R_n, and P_n. We have written a program using MATHEMATICA to find $X_n,\ U_n,\ R_n$, and P_n for $n \geq 3$. Table 1 consists of coefficients X_n, where n is between 1 and 44. We have performed the computations for cylindrical ($m=1$) and spherical ($m=2$) pistons with adiabatic coefficient $\gamma = 6/5,\ 7/5$, and $5/3$ and with Van der Waals excluded volume $b =$ 0.00006, 0.0004, 0.003, 0.01, 0.05, 0.1, 0.2, and 0.25. Although we have shown the coefficients rounded up to 15 significant digits, we have carried out all computations using 32 significant digits. We provide the graphs for the computed values of the flow variables $\rho,\ u$, and p, starting from the piston and extending to the shock before its collapse and almost up to the instant of collapse; see the figures in section 6.

4. Determination of the radius of convergence. We observe that all the coefficients found in Table 1 are positive, indicating that the nearest singularity of the shock position $X(t)$ is positive. We shall investigate whether this is the Guderley singularity, corresponding to collapse of the shock onto the axis/center. There is a steady increase in the coefficients, which indicates that the radius of convergence is less than unity. We obtain this radius of convergence assuming that the nearest singularity has the form

$$
R(t)=1-X(t)=1-\sum X_n t^n \sim A_1\left(1-\frac{t}{t_c}\right)^{\alpha_1} \quad \text{as} \quad t\to t_c, \tag{25}
$$

TABLE 1
Coefficients X_n in series (13) *for the position of shock wave.*

n	$\gamma = 7/5, m = 2, b = 0.05$	$\gamma = 7/5, m = 2, b = 0.25$	$\gamma = 5/3, m = 1, b = 0.1$
1	1.263157894736842	1.600000000000000	1.481481481481481
2	0.258541089566020	0.746666666666667	0.254752106603959
3	0.282984058072749	1.157925925925926	0.232952155612029
4	0.297254514031253	1.973345808966862	0.241628413821622
5	0.346287950377869	3.793975156923648	0.287682236139467
6	0.450095621010967	7.858055489704435	0.376557168684981
7	0.619867833831531	17.04730147416937	0.523971811959874
8	0.883302887805608	38.25248460415753	0.760931520063466
9	1.294810363112074	88.06648708646051	1.140863116560825
10	1.942296378922512	206.8465365162004	1.753305093348872
11	2.966335408193292	493.6947783445748	2.748153507103557
12	4.596608715830178	1193.969317804952	4.377341945653990
13	7.210481012551941	2919.509564663995	7.066309120004236
14	11.42886787507827	7205.784068093584	11.53688671222373
15	18.27711349478003	17928.08815485652	19.01928838320196
16	29.45533764428738	44916.74824487822	31.61894150826695
17	47.79235369600088	113221.7267883896	52.95313351205065
18	78.00968744834458	286938.8725873484	89.25898997965009
19	128.0112062907959	730683.3403143386	151.3267352842264
20	211.0646531133836	$1.868665974467891 \times 10^6$	257.8811596017432
21	349.4989923596837	$4.797463012661039 \times 10^6$	441.5096443797921
22	580.9826547894644	$1.235976372907586 \times 10^7$	759.0768191918469
23	969.2034915083856	$3.194404376452741 \times 10^7$	1310.058447571227
24	1622.067138806888	$8.280016914921135 \times 10^7$	2268.877537744009
25	2722.755803054768	$2.151936445394558 \times 10^8$	3942.038759268135
26	4582.815424832222	$5.606506312830268 \times 10^8$	6869.249418002710
27	7733.024164841269	$7.319951612955246 \times 10^9$	12002.67727977264
28	13079.14069840282	$3.830844261311359 \times 10^9$	21025.20275517070
29	22169.19843835613	$1.004375975806659 \times 10^{10}$	36916.24995769399
30	37652.79278127936	$2.638064893846673 \times 10^{10}$	64959.16474126197
31	64071.18083168314	$6.940800757561001 \times 10^{10}$	114537.0187267555
32	109217.9678176402	$1.829031909760342 \times 10^{11}$	202337.8355764119
33	186484.9035575728	$4.827004102876312 \times 10^{11}$	358082.1048635953
34	318910.0433006593	$1.275673997131769 \times 10^{12}$	634768.8550780901
35	546170.3804084596	$3.375761764298510 \times 10^{12}$	$1.127024565053689 \times 10^6$
36	936671.6937018691	$8.944197339840730 \times 10^{12}$	$2.003996195125300 \times 10^6$
37	$1.608469470233142 \times 10^6$	$2.372567637366618 \times 10^{13}$	$3.568373840198949 \times 10^6$
38	$2.765501931287699 \times 10^6$	$6.300514903398902 \times 10^{13}$	$6.362398213749396 \times 10^6$
39	$4.760392533025991 \times 10^6$	$1.674898222785370 \times 10^{14}$	$1.135842433844710 \times 10^7$
40	$8.203389006623327 \times 10^6$	$4.456895014606242 \times 10^{14}$	$2.030177893358934 \times 10^7$
41	$1.415147182792444 \times 10^7$	$1.187097535288542 \times 10^{15}$	$3.632812463622128 \times 10^7$
42	$2.443686268564865 \times 10^7$	$3.164684422415196 \times 10^{15}$	$6.507598489969976 \times 10^7$
43	$4.223808002817224 \times 10^7$	$8.443953622567328 \times 10^{15}$	$1.166930858348849 \times 10^8$
44	$7.307327247673245 \times 10^7$	$2.254838311490953 \times 10^{16}$	$2.094572381224061 \times 10^8$

where α_1 and A_1 are the leading exponent and amplitude, respectively.

Hence,

$$\frac{X_n}{X_{n-1}} \sim \frac{1}{t_c}\left(1 - \frac{1+\alpha_1}{n}\right) \quad \text{as} \quad n \to \infty. \tag{26}$$

We assume that, when n is large, the ratio X_n/X_{n-1} approximates the value of $1/t_c$. We construct a sequence X_n/X_{n-1} $(n = 35, 36, \ldots, 44)$ and thereafter refine this estimate of $1/t_c$ by forming a Neville table (Gaunt and Guttmann [8]). Neville's algorithm is a recursive procedure in which the given set of data points are interpolated by a Lagrange polynomial. In our case we fit an rth degree polynomial in $1/n$ to $r+1$

TABLE 2
The Neville table for estimating $1/t_c$ for $m = 2$, $\gamma = 7/5$, and $b = 0.05$.

n	e_n^0	Linear	Quadratic	Cubic	Quartic
40	1.72325894	1.79776720	1.79782272	1.79782923	1.79783062
41	1.72507628	1.79776993	1.79782320	1.79782937	1.79783071
42	1.72680715	1.79777249	1.79782365	1.79782951	1.79783083
43	1.72845756	1.79777489	1.79782407	1.79782965	1.79783097
44	1.73003300	1.79777714	1.79782446	1.79782978	1.79783108

TABLE 3
The Neville table for estimating α_1 for $m = 2$, $\gamma = 7/5$, and $b = 0.05$.

n	e_n^0	Linear	Quadratic	Cubic	Quartic
40	0.65914284	0.66051342	0.66067537	0.66069908	0.66069431
41	0.65917646	0.66052140	0.66067705	0.66069835	0.66069158
42	0.65920866	0.66052888	0.66067850	0.66069734	0.66068783
43	0.65923953	0.66053590	0.66067972	0.66069603	0.66068327
44	0.65926914	0.66054248	0.66068073	0.66069447	0.66067880

TABLE 4
The Neville table for estimating t_c for $m = 2$, $\gamma = 7/5$, and $b = 0.05$.

n	Linear	Quadratic	Cubic	Quartic	Quintic
40	0.55623236	0.55622829	0.55622655	0.55622593	0.55622563
41	0.55623215	0.55622815	0.55622649	0.55622589	0.55622560
42	0.55623195	0.55622803	0.55622643	0.55622585	0.55622557
43	0.55623177	0.55622791	0.55622637	0.55622582	0.55622555
44	0.55623159	0.55622781	0.55622632	0.55622578	0.55622553

points $e_n^0, e_{n-1}^0, \ldots, e_{n-r}^0$. Given a sequence e_n^0, we can construct a triangular array of elements e_n^0, where n labels the rows and $r = 0, 1, 2, 3, \ldots, n$ the columns. The elements of the rth column are generated from the $(r-1)$th column by using the formula [8]

$$e_n^r = \frac{n e_n^{r-1} - (n-r) e_{n-1}^{r-1}}{r}, \tag{27}$$

where e_n^r is the intercept on the $1/n = 0$ axis of the rth degree curve. Hence, $r = 1$ corresponds to the linear intercepts, $r = 2$ to the quadratic intercepts, $r = 3$ to the cubic intercepts, and so on. The Neville table so formed provides a refined estimate for e_n^0. For constructing a Neville table for $1/t_c$, we compute the sequence X_n/X_{n-1} ($n = 35, 36, \ldots, 44$) and take this as the initial sequence e_n^0; this gives the first column of Table 2. Then we use (27) to compute the sequence e_n^1 ($n = 35, 36, \ldots, 44$), and this forms the second column of Table 2. Again using the sequence e_n^1, we construct the sequence e_n^2 ($n = 35, 36, \ldots, 44$), and this forms the third column of Table 2. All other columns of the Neville table are formed similarly, which show that the sequences $e_n^0, e_n^1, e_n^2, \ldots$ approach a limiting value of $1/t_c$. From Tables 2, 3, and 4, the values of $1/tc$, α_1, and t_c for a spherical piston with $\gamma = 7/5$ and $b = 0.05$ are 1.79783, 0.66069, and 0.556226, respectively, and the corresponding values for an ideal gas ($b = 0$) are 1.609021, 0.7171, and 0.621496, respectively.

To verify that the nearest singularity of the shock position $X(t)$ corresponds to collapse of the shock wave onto the axis, we calculate the time t_0 for $X(t)$ to reach unity. We construct a Neville table, Table 4, using these values of t_0 as the initial sequence e_n^0. We find from Table 4 that $t_c = 0.556226$. Therefore the values of t_c obtained from Tables 2 and 4 are almost the same. The time of collapse of the shock

TABLE 5
Values of t_c, the time taken by the shock to collapse.

γ	b	t_c(cylindrical)	t_c(spherical)
7/5	0.05	0.656381	0.556226
7/5	0.1	0.603085	0.498700
7/5	0.2	0.507272	0.402430
7/5	0.25	0.463878	0.361770
5/3	0.1	0.534592	0.440227
5/3	0.2	0.456106	0.365164

wave for different values of γ, b, and m are listed in Table 5, showing thereby that for given values of γ and m, increase in b causes t_c to decrease.

5. Guderley's local solution. Ours is a global solution that is valid throughout the flow field almost up to the instant of collapse. We now extract Guderley's local singular behavior, which is valid in the neighborhood of collapse. According to Guderley's conjecture, the radius of the shock in the vicinity of collapse is prescribed by an expansion:

$$R(t) = 1 - X(t) \sim \sum_{j=1} A_j \left(1 - \frac{t}{t_c}\right)^{\alpha_j}. \tag{28}$$

Guderley [1] computed only the first exponent α_1; the other exponents α_j and amplitudes A_j were unknown. Van Dyke and Guttmann [3] calculated all the real exponents and the corresponding amplitudes for an ideal gas using the method given by Baker and Hunter [9].

Using our estimates for t_c, given in Table 5, we rewrite the series (13) for $X(t)$ in powers of the new variable τ defined by $t = t_c[1 - \exp(-\tau)]$, then multiply the nth term in (28) by $n!$, and then finally sum over n to obtain the series for the auxiliary function

$$\mathcal{R}(\tau) = \sum_{j=1} \frac{A_j}{(1 + \alpha_j \tau)}, \tag{29}$$

which is meromorphic and has simple poles at $\tau = -1/\alpha_j$ with corresponding residues A_j/α_j. Quantities α_j and A_j are evaluated by forming Pade approximants (Baker [10]) to $\mathcal{R}(\tau)$. The $[(N-1)/N]$ Pade approximant is a rational function approximation to a Taylor's series expansion. The idea of Pade approximants is to replace a power series $S(w)$ by a rational function $P_{N-1}(w)/Q_N(w)$, where $P_{N-1}(w)$ and $Q_N(w)$ are polynomials of degree $N-1$ and N, respectively. The coefficients of $P_{N-1}(w)$ and $Q_N(w)$ are determined by equating the like powers of w in the equation

$$S(w) - \frac{P_{N-1}(w)}{Q_N(w)} = O(w^{2N}),$$

where N can be varied up to 22 at the most, since we have a series (13) with 44 terms; the entire computational work has been carried out using MATHEMATICA. The exponents (α_j) and the amplitudes (A_j) are listed in the Table 6.

6. Results and conclusion. We have found a global solution to the imploding shock problem in a Van der Waals gas using a technique proposed by Van Dyke and Guttmann [3]. From this global solution the Guderley's local solution is extracted at

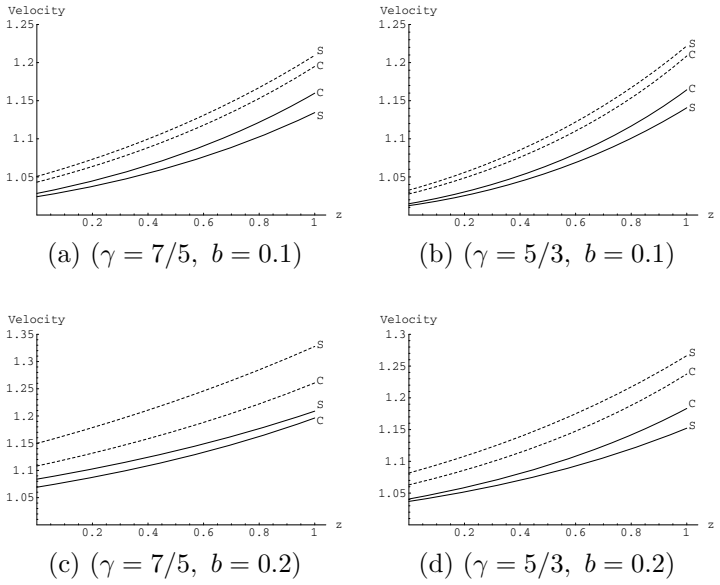
TABLE 6
Similarity exponents and corresponding amplitudes for different m, γ, and b.

γ	b	m	Exponents	Amplitudes
7/5	0.05	1	$\alpha_1 = 0.802926498593542$	$A_1 = 0.976706645353648$
			$\alpha_2 = 1.708109679894251$	$A_2 = 0.017363259018156$
			$\alpha_3 = 2.905971391360548$	$A_3 = 0.007469693787882$
7/5	0.05	2	$\alpha_1 = 0.661544877048253$	$A_1 = 0.957960578620838$
			$\alpha_2 = 1.357267903704529$	$A_2 = 0.029717449271207$
			$\alpha_3 = 2.528943256821996$	$A_3 = 0.010409103488760$
7/5	0.1	1	$\alpha_1 = 0.774110465554056$	$A_1 = 0.969168370834679$
			$\alpha_2 = 1.484914398015026$	$A_2 = 0.022014927795792$
7/5	0.1	2	$\alpha_1 = 0.596033290273788$	$A_1 = 0.832554002704591$
			$\alpha_2 = 0.844650847225558$	$A_2 = 0.141755825214615$
			$\alpha_3 = 1.729973633423482$	$A_3 = 0.020534644575970$
			$\alpha_4 = 2.636485340516882$	$A_4 = 0.005214793099547$
7/5	0.2	1	$\alpha_1 = 0.728344030222535$	$A_1 = 0.944360829848564$
			$\alpha_2 = 0.854013320875237$	$A_2 = 0.030299813597965$
			$\alpha_3 = 1.865228830929922$	$A_3 = 0.025341011312357$
7/5	0.2	2	$\alpha_1 = 0.537973817651407$	$A_1 = 0.919138233082343$
			$\alpha_2 = 1.104103329053359$	$A_2 = 0.081120376413760$
			$\alpha_3 = 2.090398604049229$	$A_3 = 0.009175944106706$
7/5	0.25	1	$\alpha_1 = 0.715770379510055$	$A_1 = 0.981989769804047$
			$\alpha_2 = 1.502387377484738$	$A_2 = -0.133407209059492$
			$\alpha_3 = 1.584581969860033$	$A_3 = 0.151461717963143$
7/5	0.25	2	$\alpha_1 = 0.507834057982443$	$A_1 = 0.858272687100006$
			$\alpha_2 = 0.931959563156305$	$A_2 = 0.126778038058150$
			$\alpha_3 = 1.847831095895297$	$A_3 = 0.012312326098207$
			$\alpha_4 = 2.845711931806961$	$A_4 = 0.001309688628189$
5/3	0.1	1	$\alpha_1 = 0.779399116466193$	$A_1 = 0.992086393567176$
			$\alpha_2 = 2.283144023168935$	$A_2 = 0.006703005432861$
5/3	0.1	2	$\alpha_1 = 0.632769630031156$	$A_1 = 0.985811255634619$
			$\alpha_2 = 1.039338542001711$	$A_2 = 0.002857291885896$
			$\alpha_3 = 2.251608063705212$	$A_3 = 0.011338631914952$
5/3	0.2	1	$\alpha_1 = 0.751708101570398$	$A_1 = 1.000011462992213$
			$\alpha_2 = 1.534154947593532$	$A_2 = -0.012389118761152$
5/3	0.2	2	$\alpha_1 = 0.594130566317584$	$A_1 = 0.998215606732302$
			$\alpha_2 = 1.559319200690666$	$A_2 = 0.022661556238679$
			$\alpha_3 = 2.804029955386938$	$A_3 = 0.001506304684029$

the instant of collapse. Guderley's local solution provides us only the first dominant similarity exponent, but we have found, through this technique, the other three less dominant similarity exponents and the corresponding amplitudes. These similarity exponents and the corresponding amplitudes are listed in Table 6. The values of the leading exponents α_1 for cylindrical and spherical symmetry for $\gamma = 6/5, 7/5$, and $5/3$, and for different values of b (0.00006, 0.0004, 0.003, 0.01, 0.05, 0.1, 0.2, and 0.25) are shown in Table 7; these values compare well with the numerical results obtained by Wu and Roberts [5] and Guderley [1]. We observe that the present method shows no evidence of another branch of solutions that arises in the work of Wu and Roberts. It may be recalled that Wu and Roberts [5] have reported that they were unable to determine the similarity exponents for $\gamma = 6/5$ and the values of b lying in the range $[0.0001, 0.05)$; indeed, we have successfully obtained the exponents in this range and compared them with the corresponding exponents found by Guderley's method in Table 7. Wu and Roberts also reported that for $\gamma = 5/3$ there exists one branch of solution for b lying in the range $[0, 0.05]$ and another branch of solution for b lying in the range $[0.005, 0.035]$; however, the present approach reveals only one solution for

TABLE 7
The computed values of the leading similarity exponents and the results obtained by other authors.

γ	b	m	Computed α_1	Guderley [1]	Wu and Roberts [5]
6/5	0.00006	2	0.757447160385346	0.75700	
6/5	0.0004	2	0.7572638179401117	0.756182	
6/5	0.01	2	0.7344920172176496	0.733259	
7/5	0.05	1	0.8029264985935416	0.802916	
7/5	0.05	2	0.661544877048253	0.661411	0.661
7/5	0.1	1	0.774110465554056	0.774049	
7/5	0.1	2	0.596033290273788	0.590510	
7/5	0.2	1	0.7283440302225347	0.730108	
7/5	0.2	2	0.5379738176514068	0.531159	
7/5	0.25	1	0.71577037951005527	0.715962	
7/5	0.25	2	0.5078340579824431	0.53448	0.534
5/3	0.003	2	0.6864958952645261	0.686495	
5/3	0.025	2	0.6697932669755673	0.659800	
5/3	0.1	1	0.7793991164661934	0.779357	
5/3	0.1	2	0.6327696300311559	0.632929	
5/3	0.2	1	0.7517081015703977	0.751790	
5/3	0.2	2	0.594130566317583	0.594017	



(a) ($\gamma = 7/5,\ b = 0.1$) (b) ($\gamma = 5/3,\ b = 0.1$)

(c) ($\gamma = 7/5,\ b = 0.2$) (d) ($\gamma = 5/3,\ b = 0.2$)

FIG. 2. *Velocity profiles for cylindrical* (C) *and spherical* (S) *symmetries. The solid and dashed curves correspond to the time of collapse $t = t_c$ and the time just before collapse $t = t_{bc}$, respectively.* (a) *represents the flow patterns for* $\gamma = 7/5,\ b = 0.1$*; for the cylindrical* (C) *symmetry:* $t_c = 0.6031, t_{bc} = 0.58$*; for the spherical* (S) *symmetry:* $t_c = 0.4987,\ t_{bc} = 0.48$. (b) *represents the flow patterns for* $\gamma = 5/3,\ b = 0.1$*; for* (C)*:* $t_c = 0.5346,\ t_{bc} = 0.52$*; for* (S)*:* $t_c = 0.4402,\ t_{bc} = 0.42$. (c) *represents the flow patterns for* $\gamma = 7/5,\ b = 0.2$*; for* (C)*:* $t_c = 0.5073,\ t_{bc} = 0.48$*; for* (S)*:* $t_c = 0.4024,\ t_{bc} = 0.38$. (d) *represents the flow patterns for* $\gamma = 5/3,\ b = 0.2$*; for* (C)*:* $t_c = 0.4561,\ t_{bc} = 0.44$*; for* (S)*:* $t_c = 0.3652,\ t_{bc} = 0.35$.

the above ranges of b and γ.

The flow variables, velocity, density, and pressure are also computed for different values of γ and b. They are shown in the Figures 2, 3, and 4. We observe from Table

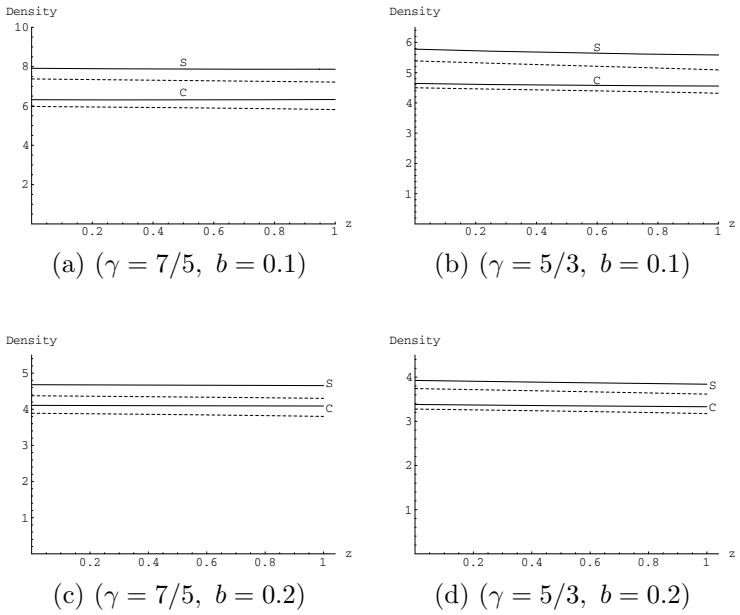


FIG. 3. *Density profiles for cylindrical* (C) *and spherical* (S) *symmetries. The solid and dashed curves correspond to the time of collapse* $t = t_c$ *and the time just before collapse* $t = t_{bc}$*, respectively.* (a) *represents the flow patterns for* $\gamma = 7/5,\ b = 0.1$*; for the cylindrical* (C) *symmetry:* $t_c = 0.6031,\ t_{bc} = 0.58$*; for the spherical* (S) *symmetry:* $t_c = 0.4987,\ t_{bc} = 0.48$*;* (b) *represents the flow patterns for* $\gamma = 5/3,\ b = 0.1$*; for C:* $t_c = 0.5346,\ t_{bc} = 0.52$*; for S:* $t_c = 0.4402,\ t_{bc} = 0.42$*;* (c) *represents the flow patterns for* $\gamma = 7/5,\ b = 0.2$*; for C:* $t_c = 0.5073,\ t_{bc} = 0.48$*; for S:* $t_c = 0.4024,\ t_{bc} = 0.38$*;* (d) *represents the flow patterns for* $\gamma = 5/3,\ b = 0.2$*; for C:* $t_c = 0.4561,\ t_{bc} = 0.44$*; for S:* $t_c = 0.3652,\ t_{bc} = 0.35$.

6 that increase in either of the parameters b or m, or decrease in γ, causes the leading similarity exponent α_1 to decrease and consequently the shock acceleration to increase as it approaches the center/axis. We also notice that the shock is continuously accelerated, since α_1 is always less than one; in fact, the shock speed becomes unbounded as $t \to t_c$, but less rapidly than $(t - t_c)^{-1}$. Figures 2(a)–2(d) and 3(a)–3(d) show that the velocity decreases monotonically behind the shock as we move towards the piston, whereas the density increases in the region behind the shock; this is because of geometrical convergence or area contraction of the shock, which causes velocity to decrease and density to increase. We also observe from these figures that an increase in b causes the particle velocity to increase and density to decrease. Figures 4(a)–4(d) indicate that the gas pressure remains constant in most of the region except in the vicinity of the front, where it exhibits a maximum; this is due to the fact that the gas, which is highly compressed by the shock, gets cooled down in the region behind the shock. It is also observed that an increase in b causes the gas pressure to increase in the region behind the shock. Table 5 shows that, for given values of γ and m, an increase in b causes t_c to decrease; the results are depicted in Figures 5(a)–5(d), which show that an increase in b causes the time of shock collapse to decrease, i.e., the shock reaches the center/axis much faster with the increase of Van der Waals excluded volume b.

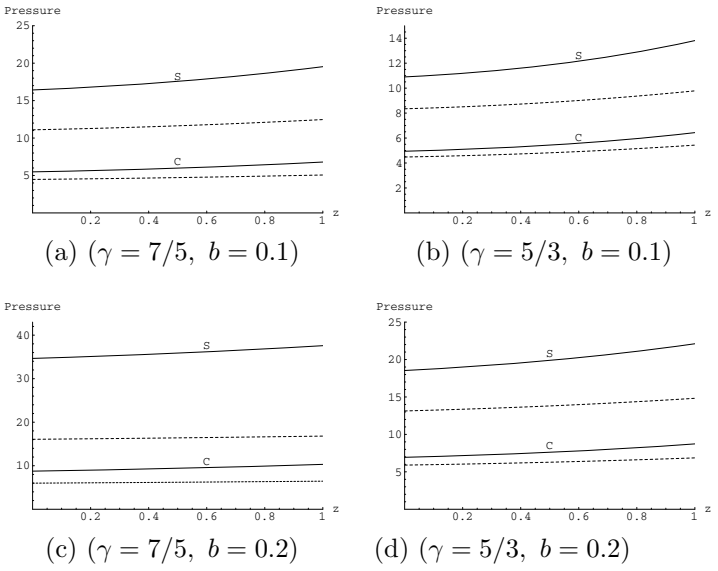


FIG. 4. *Pressure profiles for cylindrical* (C) *and spherical* (S) *symmetries. The solid and dashed curves correspond to the time of collapse* $t = t_c$ *and the time just before collapse* $t = t_{bc}$*, respectively.* (a) *represents the flow patterns for* $\gamma = 7/5,\ b = 0.1$*; for the cylindrical* (C) *symmetry:* $t_c = 0.6031,\ t_{bc} = 0.58$*; for the spherical* (S) *symmetry:* $t_c = 0.4987,\ t_{bc} = 0.48$*;* (b) *represents the flow patterns for* $\gamma = 5/3,\ b = 0.1$*; for C:* $t_c = 0.5346,\ t_{bc} = 0.52$*; for S:* $t_c = 0.4402,\ t_{bc} = 0.42$*;* (c) *represents the flow patterns for* $\gamma = 7/5,\ b = 0.2$*; for C:* $t_c = 0.5073,\ t_{bc} = 0.48$*; for S:* $t_c = 0.4024,\ t_{bc} = 0.38$*;* (d) *represents the flow patterns for* $\gamma = 5/3,\ b = 0.2$*; for C:* $t_c = 0.4561,\ t_{bc} = 0.44$*; for S:* $t_c = 0.3652,\ t_{bc} = 0.35$.

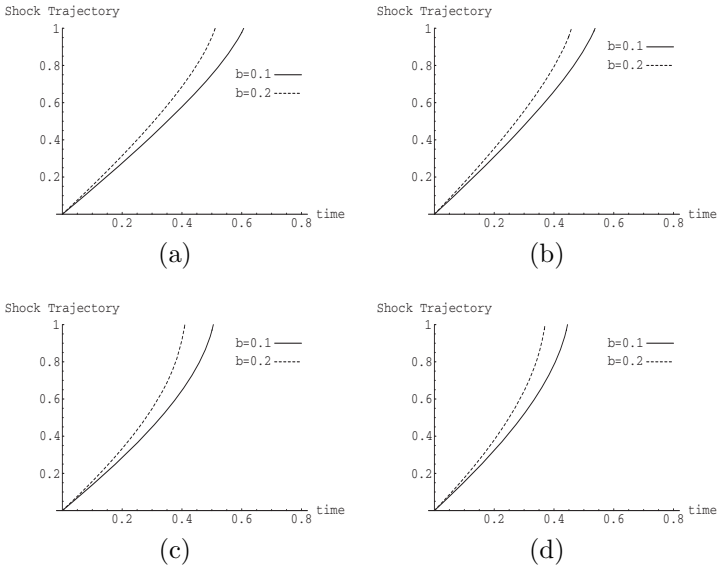


FIG. 5. *Shock Trajectory: The solid and dashed curves correspond to* $b = 0.1$ *and* $b = 0.2$*, respectively.* (a) $\gamma = 7/5,\ m = 1$. (b) $\gamma = 5/3,\ m = 1$. (c) $\gamma = 7/5,\ m = 2$. (d) $\gamma = 5/3,\ m = 2$.

Acknowledgments. We thank the reviewers of this paper for their useful remarks.

REFERENCES

[1] G. Guderley, *Starke kugelige und zylindrische Verdichtungsstosse in der Nahe des Kugelmittelpunktes bzw der Zylinderachse*, Luftfahrtforschung, 19 (1942), pp. 302–312.

[2] R. B. Lazarus and R. D. Richtmyer, *Similarity Solutions for Converging Shocks*, Los Alamos Scientific Laboratory Report, LA-6823-MS, Los Alamos, NM, 1977.

[3] M. Van Dyke and A. J. Guttmann, *The converging shock wave from a spherical or cylindrical piston*, J. Fluid Mech., 120 (1982), pp. 451–462.

[4] P. Hafner, *Strong convergent shock waves near the center of convergence: A power series solution*, SIAM J. Appl. Math., 48 (1988), pp. 1244–1261.

[5] C. C. Wu and P. H. Roberts, *Structure and stability of a spherical shock wave in a Van der Waals gas*, Quart. J. Mech. Appl. Math., 49 (1996), pp. 501–543.

[6] G. Madhumita and V. D. Sharma, *Propagation of strong converging shock waves in a gas of variable density*, J. Engrg. Math., 46 (2003), pp. 55–68.

[7] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley-Interscience, New York, 1974.

[8] D. S. Gaunt and A. J. Guttmann, *Asymptotic analysis of coefficients*, in Phase Transitions and Critical Phenomena, C. Domb and M. S. Green, eds., Academic, New York, 1974, vol. 3, pp. 181–243.

[9] G. A. Baker and D. L. Hunter, *Methods of series analysis* II. *Generalized and extended methods with applications to the Ising model*, Phys. Rev., B7 (1973), pp. 3377–3392.

[10] G. A. Baker, *The theory and application of the Pade approximant method*, in Advances in Theoretical Physics, K. A. Brueckner, ed., Academic, New York, 1965, vol. 1, pp. 1–58.

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 1839–1852

ON THE SOLUTION OF LONG'S EQUATION WITH SHEAR*

MAYER HUMI†

Abstract. Long's equation describes two dimensional stratified flow over terrain. Its numerical solutions under various approximations were investigated by many authors under the assumption that the base flow field is without shear. Special attention was paid to the properties of the gravity waves that are predicted to be generated as a result. In this paper we address, analytically, the nature and properties of these solutions when shear is present and derive some constraints on the possible generation of gravity waves under these circumstances.

Key words. gravity waves, Long's equation, shear

AMS subject classifications. 76B60, 76E05, 76E30, 86A10

DOI. 10.1137/050627794

1. Introduction. Long's equation [1, 2, 3, 4] models the flow of stratified incompressible fluid (in the Boussinesq approximation) in two dimensions over terrain. When the base state of the flow (that is, the unperturbed flow field far upstream) is without shear, the numerical solutions (in the form of steady lee waves) of this equation in various settings and approximations were studied by many authors [5, 6, 7, 8, 9, 10, 11, 12, 13]. The most common approximation in these studies was to set Brunt–Väisälä frequency to a constant or a step function over the computational domain. Moreover, the values of the parameters β and μ which appear in this equation were set to zero. In this (singular) limit the nonlinear terms and one of the leading second order derivatives in the equation drop out and the equation reduces to that of a linear harmonic oscillator over a two dimensional domain. Careful studies [8] showed that these approximations are justified unless wave breaking is present in the solution [9].

Long's equation also provides the theoretical framework for the analysis of experimental data [15, 16, 17] under the assumption of shearless base flow. (An assumption which, in general, is not supported by the data.) An extensive list of references appears in [18, 19, 20].

An analytic approach to the study of this equation and its solutions was initiated recently by the author [14]. We showed that for a base flow without shear and under rather mild restrictions the nonlinear terms in the equation can be simplified. We also identified the "slow variable" that controls the nonlinear oscillations in this equation and, using phase averaging approximation, derived a formula for the attenuation of the stream function perturbation with height. This result is generically related to the presence of the nonlinear terms in Long's equation.

The objective of this paper is to study the nature of the solutions to Long's equation when shear is present in the base flow and Brunt–Väisälä frequency is a continuous function of height. Using conditions which depend solely on the base flow and Brunt–Väisälä frequency we characterize the qualitative nature of the perturbations from the base flow and how their amplitude varies with height. These results

*Received by the editors March 27, 2005; accepted for publication (in revised form) May 22, 2006; published electronically August 22, 2006.
http://www.siam.org/journals/siap/66-6/62779.html
†Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609 (mhumi@wpi.edu).

are independent of the actual detailed description of the terrain that caused these perturbations. Furthermore we derive conditions under which these perturbations are not oscillatory; i.e., no gravity waves are generated by the flow. To the best of our knowledge this issue was never considered in the literature before (in the context of Long's equation).

The plan of the paper is as follows: In section 2 we present a short review of the derivation of Long's equation and the solution of its linearized version. In section 3 we derive constraints on the solutions of this equation in a general setting and in particular in the presence of shear. In section 4 solutions to this equation with different shear profiles are studied explicitly. In section 5 we carry out simulations of Long's equation for shearless and shear base flows. We end in section 6 with summary and conclusions.

2. Long's equation: A short review. In two dimensions (x, z) the flow of a steady inviscid and incompressible stratified fluid (in the Boussinesq approximation) is modeled by the following equations:

$$u_x + w_z = 0, \tag{2.1}$$

$$u\rho_x + w\rho_z = 0, \tag{2.2}$$

$$\rho(uu_x + wu_z) = -p_x, \tag{2.3}$$

$$\rho(uw_x + ww_z) = -p_z - \rho g, \tag{2.4}$$

where subscripts indicate differentiation with respect to the indicated variable, $\mathbf{u} = (u, w)$ is the fluid velocity, ρ is its density, p is the pressure, and g is the acceleration of gravity.

We can nondimensionalize these equations by introducing

$$\bar{x} = \frac{x}{L}, \quad \bar{z} = \frac{N_0}{U_0} z, \quad \bar{u} = \frac{u}{U_0}, \quad \bar{w} = \frac{LN_0}{U_0^2} w,$$
$$\bar{\rho} = \frac{\rho}{\rho_0}, \quad \bar{p} = \frac{N_0}{gU_0\rho_0} p, \tag{2.5}$$

where L represents a characteristic length and U_0, ρ_0 represent, respectively, the free stream velocity and density. N_0 is the characteristic Brunt–Väisälä frequency

$$N_0^2 = -\frac{g}{\rho_0} \frac{d\rho_0}{dz}. \tag{2.6}$$

In these new variables, (2.1)–(2.4) take the following form (for brevity we drop the bars):

$$u_x + w_z = 0, \tag{2.7}$$

$$u\rho_x + w\rho_z = 0, \tag{2.8}$$

$$\beta\rho(uu_x + wu_z) = -p_z, \tag{2.9}$$

$$\beta\rho(uw_x + ww_z) = -\mu^{-2}(p_z + \rho), \tag{2.10}$$

where

$$\beta = \frac{N_0 U_0}{g}, \tag{2.11}$$

$$\mu = \frac{U_0}{N_0 L}. \tag{2.12}$$

β is the Boussinesq parameter [13] which controls stratification effects (assuming $U_0 \neq 0$), and μ is the long wave parameter which controls dispersive effects (or the deviation from the hydrostatic approximation). In the limit $\mu = 0$ the hydrostatic approximation is fully satisfied [20].

In view of (2.7) we can introduce a stream function ψ so that

$$u = \psi_z, \quad w = -\psi_x. \tag{2.13}$$

From (2.8) and (2.13) we infer that $\rho = \rho(\psi)$ and (after some algebra) derive the following equation for ψ [13]:

$$\psi_{zz} + \mu^2 \psi_{xx} - N^2(\psi)\left[z + \frac{\beta}{2}(\psi_z^2 + \mu^2 \psi_x^2)\right] = G(\psi), \tag{2.14}$$

where

$$N^2(\psi) = -\frac{\rho_\psi}{\beta \rho} \tag{2.15}$$

is the nondimensional Brunt–Väisälä frequency. $G(\psi)$ is some unknown function which is determined from the base flow, which henceforth we assume to be a function of z only. To carry out this determination we consider (2.14) at $x = -\infty$ and express the left-hand side of this equation in terms of ψ only (assuming that disturbances do not propagate far upstream [19]). Equation (2.14) is referred to as Long's equation.

For example, if we let

$$\psi(-\infty, z) = z, \tag{2.16}$$

i.e., consider a shearless base flow with $u(-\infty, z) = 1$, then

$$G(\psi) = -N^2(\psi)\left(\psi + \frac{\beta}{2}\right) \tag{2.17}$$

and (2.14) becomes

$$\psi_{zz} + \mu^2 \psi_{xx} - N^2(\psi)\left[z - \psi + \frac{\beta}{2}(\psi_z^2 + \mu^2 \psi_x^2 - 1)\right] = 0. \tag{2.18}$$

However, it is evident that different profiles for the base flow at $x = -\infty$ will lead to different forms of $G(\psi)$ (for examples, see section 4).

For a general base flow in an unbounded domain over topography with shape $f(x)$ and maximum height H the following boundary conditions are imposed on ψ:

$$\psi(-\infty, z) = \psi^0(z), \tag{2.19}$$

$$\psi(x, \epsilon f(x)) = \text{constant}, \quad \epsilon = \frac{H N_0}{U_0}, \tag{2.20}$$

where the constant in (2.20) is (usually) set to zero. As to the boundary condition on $\psi(\infty, z)$ we observe that Long's equation contains no dissipation terms and therefore only radiation boundary conditions can be imposed in this limit. Similarly at $z = \infty$ it is customary to impose (following [7]) radiation boundary conditions. For the perturbation from the shearless base flow

$$\phi = \psi - z, \tag{2.21}$$

(2.18) becomes

$$\phi_{zz} - \alpha^2\phi_z^2 + \mu^2(\phi_{xx} - \alpha^2\phi_x^2) - N^2(\phi)(\beta\phi_z - \phi) = 0, \tag{2.22}$$

where

$$\alpha^2 = \frac{N^2(\psi)\beta}{2}. \tag{2.23}$$

Since ψ is set to zero at the bottom topography, the corresponding (approximate) boundary condition on ϕ for small ϵ will be

$$\phi(x, 0) = -\epsilon f(x). \tag{2.24}$$

Thus for small amplitude topography the boundary condition can be applied at $z = 0$.

In the limits $\beta = 0$, $\mu = 0$ and when $N(\psi)$ is a constant whose value over the domain is N, (2.22) reduces to a linear equation

$$\phi_{zz} + N^2\phi = 0. \tag{2.25}$$

We observe that the limit $\beta = 0$ can be obtained by letting either $U_0 \to 0$ or $N_0 \to 0$. In the following we assume that this limit is obtained as $U_0 \to 0$ (so that stratification persists in this limit). The general solution of (2.25) is

$$\phi(x, z) = p(x)\cos(Nz) + q(x)\sin(Nz), \tag{2.26}$$

where the functions $p(x), q(x)$ have to be chosen so that the boundary conditions derived from (2.19), (2.20) and the radiation boundary conditions are satisfied. These boundary conditions lead in general to an integral equation for $p(x)$ and $q(x)$:

$$q(x)\cos(\epsilon N f(x)) + H[q(x)]\sin(\epsilon N f(x)) = -\epsilon f(x), \tag{2.27}$$

where $H[q(x)]$ is the Hilbert transform of q(x). This equation has to be solved numerically [6, 7].

It is clear from the form of the general solution given by (2.26) that it represents a wave in the z-direction, and the properties of this wave (under varied physical conditions) were investigated by the authors mentioned in section 1. It should be observed, however, that (2.25) is a "singular limit" of Long's equation, as one of the leading second order derivatives drops when $\mu = 0$ and the nonlinear terms drop when $\beta = 0$. This approximation and its limitations were considered numerically and analytically [6, 7, 14] and were found to be justified under the assumption that the base flow is shearless. It is used in the actual analysis of atmospheric data [16, 17, 18].

3. Properties of solutions to Long's equation with shear. In this section we address the nature and properties of the perturbation from the base flow when shear is present. To simplify our notation and treatment we set $\mu = 1$ since we can always scale x as $\bar{x} = x/\mu$ (and drop the bars on $\bar{x}$). We also assume that

$$\lim_{x\to-\infty} \psi(x,z) = \psi^0(z). \tag{3.1}$$

Long's equation with shear is then

$$(\psi_{zz} - \alpha^2\psi_z^2) + (\psi_{xx} - \alpha^2\psi_x^2) - N^2(\psi)z = G(\psi), \tag{3.2}$$

and in the limit $x \to -\infty$, $G(\psi)$ must satisfy

$$G(\psi^0) = \psi^0_{zz} - N^2(\psi^0)\left[z + \frac{\beta}{2}(\psi^0_z)^2\right]. \tag{3.3}$$

To treat the perturbation from the base state we write

$$\psi(x,z) = \psi^0(z) + \phi(x,z). \tag{3.4}$$

Substituting this in (3.2) and linearizing using (3.3) we obtain

$$\nabla^2\phi - 2\alpha^2\psi^0_z\phi_z - (N^2)'(\psi^0)z\phi - G'(\psi^0)\phi = 0, \tag{3.5}$$

where primes denote differentiation with respect to ψ.

Since ψ^0 is a function of z only, this equation is separable and we can deduce the properties of its solution by applying separation of variables. Introducing

$$\phi(x,z) = \chi(x)\eta(z), \tag{3.6}$$

we obtain

$$\chi(x)_{xx} + \lambda\chi(x) = 0 \tag{3.7}$$

and

$$\eta(z)_{zz} - 2\alpha^2\psi^0_z\eta_z - [\lambda + (N^2)'(\psi^0)z + G'(\psi^0)]\eta = 0, \tag{3.8}$$

where λ is the separation of variables constant. Equation (3.8) can be rewritten as

$$\frac{d}{dz}\left(e^{-2\alpha^2\psi^0}\frac{d}{dz}\eta(z)\right) - H(z)\eta = 0, \tag{3.9}$$

where

$$H(z) = e^{-2\alpha^2\psi^0}[\lambda + (N^2)'(\psi^0)z + G'(\psi^0)]. \tag{3.10}$$

Equations (3.9)–(3.10) demonstrate that the properties of the perturbation as a function z depend only on λ (i.e., the wave number in the x direction) and the initial state of the flow.

To obtain further information about the properties of η we observe that $e^{-2\alpha^2\psi^0} \geq 0$ and it is possible to apply to (3.9) the comparison theorem of Strum and Picone [21]. A direct application of these theorems leads to the following result.

Assume that on the interval $[a, z]$

$$0 < m \leq e^{-2\alpha^2\psi^0} \leq M, \tag{3.11}$$

$$k \leq H(z) \leq K. \tag{3.12}$$

Then the following hold:

1. If $0 < k$, the solution of (3.9) is not oscillatory (no waves on the interval [a,z]).
2. If $k < 0$ and

$$\frac{-\pi^2}{(z-a)^2} < \frac{k}{m}, \tag{3.13}$$

then the solution of (3.9) is not oscillatory (no waves).

3. A sufficient condition for (3.9) to have an oscillatory solution with n zeros is that

$$\frac{K}{M} \leq \frac{-n^2\pi^2}{(z-a)^2}. \tag{3.14}$$

That is, the wavenumber of the wave will increase as K becomes more negative. We observe also that the estimate for k depends on the value of λ. As $0 < \lambda$ increases, this estimate for the lower bound of $H(z)$ will increase and when k satisfies the inequality (3.13), the solution for η will become nonoscillatory (that is, the wave is trapped). This demonstrates the "coupling" between the horizontal wavenumber of oscillations and the nature of the solution in the vertical direction.

To obtain further insight into the nature of the solution for $\eta(z)$ we multiply (3.8) by η and integrate over [0,z]. Using integration by parts we obtain

$$[\eta\eta' - \alpha^2\psi_z^0\eta^2]\Big|_0^z = \int_0^z \{(\eta')^2 + [\alpha^2\psi_{zz}^0 + F(z)]\eta^2\}dz, \tag{3.15}$$

where

$$F(z) = \frac{1}{2}[\lambda + (N^2)'(\psi^0)z + G'(\psi^0)]. \tag{3.16}$$

Assuming that $\eta(0) = 0$ (i.e., the amplitude of the perturbation at ground level is 0), (3.15) can be written as

$$\frac{1}{2}\frac{d\eta^2}{dz}(z) = \alpha^2\psi_z^0(z)\eta^2(z) + \int_0^z (\eta')^2 dz + \int_0^z [\alpha^2\psi_{zz}^0 + F(z)]\eta^2 dz. \tag{3.17}$$

Hence we conclude that the amplitude of the perturbation will increase with height if

$$\psi_z^0(z) > 0, \;\; \alpha^2\psi_{zz}^0 + F(z) > 0 \tag{3.18}$$

on the interval $[0, z]$.

To proceed we now invoke the (modified) Poincaré inequality on the interval $[0, z]$. This inequality states that if η is smooth enough and $\eta(0) = 0$, then

$$\int_0^z (\eta')^2(s)ds \geq \frac{\pi^2}{4z^2}\int_0^z \eta^2(s)ds. \tag{3.19}$$

(For the proof of this inequality see the appendix.)

To apply this inequality we rewrite (3.15) (assuming $\eta(0) = 0$) as

$$\frac{1}{2}\frac{d\eta^2}{dz}(z) \le \alpha^2\psi_z^0(z)\eta^2(z) + \int_0^z (\eta')^2 dz + max[\alpha^2\psi_{zz}^0 + F(z)]\int_0^z \eta^2 dz. \tag{3.20}$$

Using the Poincaré inequality to estimate the integral of η^2 yields

$$\frac{1}{2}\frac{d\eta^2}{dz}(z) \le \alpha^2\psi_z^0(z)\eta^2(z) + \left\{1 + \frac{4z^2}{\pi^2}max[\alpha^2\psi_{zz}^0 + F(z)]\right\}\int_0^{z_0} (\eta')^2 dz. \tag{3.21}$$

Hence if

$$\psi_z^0 < 0, \ \left\{1 + \frac{4z^2}{\pi^2}max[\alpha^2\psi_{zz}^0 + F(z)]\right\} < 0 \tag{3.22}$$

on the interval $[0, z]$, then the perturbation η will decay with height.

We observe that the conditions (3.18) and (3.22) depend only on the properties of the base flow and the variation of N^2 with height.

3.1. Some special cases.

1. $\psi^0(z) = z$ and $N =$ constant.

This is essentially the only case that has been treated in the literature on Long's equation. It represents a shearless base state with constant Brunt–Väisälä frequency.

In this case $G(\psi)$ is given by (2.17), and (3.8) reduces to

$$\eta_{zz} - 2\alpha^2\eta_z + (N^2 - \lambda)\eta = 0, \tag{3.23}$$

whose solution is [14]

$$\eta = Ae^{\alpha^2 z}\cos(mz + \gamma), \tag{3.24}$$

where $m = \sqrt{(N^2 - \lambda)}$ and γ is a constant.

2. $\alpha = 0$.

When α is very small we can neglect the second term in (3.8) which reduces then to

$$\eta_{zz} = 2F(z)\eta. \tag{3.25}$$

Introducing

$$v = \frac{\eta_z}{\eta} = \frac{d\ln|\eta|}{dz}, \tag{3.26}$$

this equation becomes

$$v_z + v^2 = 2F(z). \tag{3.27}$$

Hence $v_z = 2F(z) - v^2 \le 2F(z)$. We can conclude therefore that if $F(z) < 0$, then v and hence $\frac{d\ln|\eta|}{dz}$ are decreasing with height; i.e., the perturbation is being dissipated. On the other hand, if $F(z) > 0$, then v is increasing when $2F(z) - v^2$ is positive and decreasing when this quantity is negative, and therefore there will be oscillations in the amplitude of η.

4. Some examples with shear. In this section we consider some examples whose base flow is not shearless and derive explicitly the corresponding equations for the perturbations. We use then analytic methods to explore the properties of the solutions to these equations.

4.1. $\psi^0 = z^2$, N = constant. For this base flow $u = z$; that is, u increases linearly with height. Using (3.3) we find that

$$G(\psi) = 2 - N^2(\psi)[\psi^{1/2} + 2\beta\psi] \tag{4.1}$$

and Long's equation (3.2) for ψ (with $\mu = 1$) becomes

$$(\psi_{zz} - \alpha^2\psi_z^2) + (\psi_{xx} - \alpha^2\psi_x^2) - N^2(\psi)z = 2 - N^2(\psi)[\psi^{1/2} + 2\beta\psi]. \tag{4.2}$$

To derive an equation for a perturbation from the base flow, we set $\psi = z^2 + \phi(x, z)$. Substituting this in (4.2) and linearizing, we obtain

$$\nabla^2\phi - 4\alpha^2 z\frac{\partial\phi}{\partial z} + \left(\frac{N^2}{2z} + 4\alpha^2\right)\phi = 0. \tag{4.3}$$

This equation is separable, and we can consider three types of solutions:
1. $\phi(x, z) = e^{-kx}\eta(z)$ with $k > 0$.
2. $\phi(x, z) = \sin(kx)\eta(z)$ with $k > 0$.
3. $\phi(x, z) = \eta(z)$ (that is, a vertical perturbation, $k = 0$).

In all three cases we obtain for $\eta(z)$ the following equation:

$$\eta'' - 4\alpha^2 z\eta' + \left(4\alpha^2 \pm k^2 + \frac{N^2}{2z}\right)\eta = 0. \tag{4.4}$$

The solution of this equation is given by Heun biconfluent functions [22]. To explore analytically when the solution $\eta(z)$ is oscillatory, we rewrite (4.4) in the form

$$[e^{-2\alpha^2 z^2}\eta']' + \left(4\alpha^2 \pm k^2 + \frac{N^2}{2z}\right)e^{-2\alpha^2 z^2}\eta = 0. \tag{4.5}$$

This equation has the same form as (3.9), and therefore we can apply the oscillation theorems of Strum and Picone. In fact, since $0 < e^{-2\alpha^2 z^2} \leq 1$, oscillations can occur only if $(4\alpha^2 \pm k^2 + \frac{N^2}{2z})$ is positive enough. (See (3.11), (3.12), and (3.14) and note that there is a minus sign in front of $H(z)$ in (3.9).) This can happen for proper values of N^2 and α^2 in cases 1 and 3 above. It can also happen when k is small in the second case, but $4\alpha^2 + \frac{N^2}{2z}$ is large enough. Furthermore, one can show numerically that in all cases the amplitude of the perturbation grows with height. Thus the perturbation will always feed on the energy of the base flow. (However, it should be kept in mind that the amplitude of the perturbation cannot grow indefinitely. Once it violates the assumptions made to derive Long's equation and the approximations that led to (4.4), this solution becomes invalid.)

4.2. $G(\psi) = 0$, N = constant. In this case, instead of choosing the base flow and deriving $G(\psi)$ using (3.3), we make the ansatz that $G(\psi) = 0$ and compute the corresponding base flow. With this assumption (3.3) becomes

$$\psi_{zz}^0 - N^2\left[z + \frac{\beta}{2}(\psi_z^0)^2\right] = 0. \tag{4.6}$$

Introducing $y = \psi_z^0$, we obtain a Ricatti equation for $y(z)$:

$$y' - \alpha^2 y^2 - N^2 z = 0. \tag{4.7}$$

This equation can be linearized by the transformation

$$y(z) = -\frac{1}{\alpha^2}\frac{v(z)'}{v(z)}, \tag{4.8}$$

which leads to

$$v(z)'' + \alpha^2 N^2 z v(z) = 0. \tag{4.9}$$

This can be identified as a Bessel equation whose solution is

$$v(z) = C\sqrt{z} J_{1/3}\left(\frac{2}{3} N\alpha z^{3/2}\right), \tag{4.10}$$

and hence

$$\psi^0(z) = -\frac{1}{\alpha^2}\ln(|v(z)|). \tag{4.11}$$

The resulting Long's equation for this base flow is (3.2) with $G(\psi) = 0$. This equation can be linearized by the transformation $\xi = e^{-\alpha^2\psi}$, and we obtain

$$\nabla^2\xi + N^2\alpha^2 z\xi = 0. \tag{4.12}$$

This equation (for the full flow) is separable, and the nature of the solution will depend on the separation constant. We distinguish three cases:

1. $\xi(x, z) = \xi(z)$.
   In this case ξ satisfies (4.9) and hence ξ is given by (4.10).
2. $\xi(x, z) = e^{-nx}\eta(z)$, $n > 0$.
   This lead to

$$\eta'' + [N^2\alpha^2 z + n^2]\eta = 0. \tag{4.13}$$

3. $\xi(x, z) = \sin(nx)\eta(z)$, $n > 0$.
   The equation for η becomes

$$\eta'' + [N^2\alpha^2 z - n^2]\eta = 0. \tag{4.14}$$

The solution to equations (4.13), (4.14) is given in terms of Airy functions. However, the qualitative nature of the solution of these equations can be deduced from the comparison theorems of Strum and Picone [21]. (In this case the function $e^{-2\alpha^2\psi^0}$ in (3.9) is replaced by 1, and hence $m = M = 1$ in (3.11).) Thus for the second case we have $[N^2\alpha^2 z + n^2] > 0$ and therefore the solution for η will be oscillatory. Moreover, the wavenumber of the oscillations will increase with height. On the other hand, for the third case the solution will be nonoscillatory for small z but may become oscillatory with height, viz. when $[N^2\alpha^2 z - n^2]$ becomes positive.

4.3. $\psi^0 = -\cos(az)$, N = constant. For this base flow $u = a\sin(az)$; that is, u oscillates with the height.

In this case we obtain for $G(\psi)$ the expression

$$G(\psi) = -a^2\psi - N^2\left[\frac{\pi - \arccos(\psi)}{a} + \frac{\beta a^2}{2}(1-\psi^2)\right]. \tag{4.15}$$

The linearized equation for the perturbation $\phi(x,z)$ from the base flow is

$$\nabla^2\phi - 2\alpha^2 a\sin(az)\frac{\partial\phi}{\partial z} + \left[a^2 + 2\alpha^2 a^2\cos(az) - \frac{N^2}{a\sin(az)}\right]\phi = 0. \tag{4.16}$$

For a solution of the form $\phi(x,z) = e^{-kx}\eta(z)$ or $\phi(x,z) = \sin(kx)\eta(z)$ this yields

$$\left[e^{2\alpha^2 cos(az)}\eta'\right]' + e^{2\alpha^2 cos(az)}\left[a^2 \pm k^2 + 2\alpha^2 a^2\cos(az) - \frac{N^2}{a\sin(az)}\right]\eta = 0. \tag{4.17}$$

Oscillations in the solution of this equation will occur whenever the expression in the square brackets of the last term is positive. Since $\sin(az)$ takes both positive and negative values and $\frac{N^2}{asin(az)}$ is dominant when $az \approx 0$ or $az \approx \pi$, the solution will exhibit different qualitative behavior in different regions (i.e., oscillatory in some and nonoscillatory in others).

5. Numerical simulations for shear flow. In previous sections we discussed from an analytical point of view the impact of shear on the generation of gravity waves using first order perturbation expansion. To elicit more insight on this issue we compare numerically in this section the solutions of Long's equation with and without shear over the same topography and with the same values of the geophysical parameters (viz. μ, β, and N^2). Without loss of generality we set $\mu = 1$ in the following (see remark at the beginning of section 3).

The equation for a perturbation from a shearless base flow (without approximations) is given by (2.22). Similarly the exact equation for the perturbation from pure shear flow (see section 4.1) is given by

$$\nabla^2\phi - \alpha^2\left[4z\phi_z + \phi_z^2 + \phi_x^2 - 4\,\phi\right] + N^2[\sqrt{z^2+\phi} - z] = 0. \tag{5.1}$$

To simplify (2.22) and (5.1) we introduce

$$\eta = e^{-\alpha^2\phi}, \quad \alpha \neq 0, \tag{5.2}$$

and observe that

$$-\frac{1}{\alpha^2\eta}\eta_{zz} = \phi_{zz} - \alpha^2\phi_z^2. \tag{5.3}$$

Substituting this result for the second order derivatives of x and z, (2.22) and (5.1) transform, respectively, to

$$\nabla^2\eta - 2\alpha^2\eta_z + N^2\eta\ln\eta = 0, \tag{5.4}$$

$$\nabla^2\eta + \alpha^2\left\{-4z\eta_z + \eta\left[4\ln(\eta) + N^2\left(z - \sqrt{\frac{\alpha^2 z^2 - \ln(\eta)}{\alpha^2}}\right)\right]\right\} = 0. \tag{5.5}$$

These equations are linear in the derivatives of η and nonlinear only in terms which contain η itself. This simplifies the numerical algorithm for their solution.

To solve (5.4), (5.5) we implement Newton's iteration scheme. To this end we define $F(\eta)$ as the left-hand side (of each) of these equations and take its Frechet derivative, i.e., compute

$$F(\eta_0 + \delta) \cong F(\eta_0) + L(\eta_0)\delta + O(\delta^2), \tag{5.6}$$

where $L(\eta_0)$ is a linear operator. A short computation using (5.4) yields

$$L_1(\eta_0) = \nabla^2 + N^2 \left[(1 + \ln \eta_0) - \beta \frac{\partial}{\partial z}\right]. \tag{5.7}$$

Similarly, for (5.5) we obtain

(5.8)

$$L_2(\eta_0) = \nabla^2 + \alpha^2 \left[-4z\frac{\partial}{\partial z} + 4(1 + \ln \eta_0) + N^2 \left(z - \sqrt{z^2 - \frac{\ln(\eta_0)}{\alpha^2}} + \frac{1}{2\alpha^2\sqrt{z^2 - \frac{ln(\eta_0)}{\alpha^2}}}\right)\right].$$

To use Newton's iteration scheme to solve (5.4), (5.5) we now let $F(\eta_0 + \delta) = 0$ in (5.6) with $\delta = \eta_{m+1} - \eta_m$ (where the index m denotes the iteration number). This leads, respectively, to the following iteration schemes for the solution of these equations:

$$L_1(\eta_m)\eta_{m+1} = N^2\eta_m, \tag{5.9}$$

$$L_2(\eta_m)\eta_{m+1} = \left[4\alpha^2 + \frac{N^2}{2\sqrt{z^2 - \frac{ln(\eta_m)}{\alpha^2}}}\right]\eta_m. \tag{5.10}$$

To solve these equations over a finite two dimensional domain $[-a, a] \times [0, b]$ with bottom topography, we used central finite differences with a grid of 151×101 points. The (approximate) boundary conditions which were imposed on η in (5.9), (5.10), respectively, were

$$\eta(-a, z) = 1, \;\; \eta(a, z) = 1, \;\; \eta(x, b) = 1, \;\; \eta(x, 0) = e^{\alpha^2 \epsilon f(x)}, \tag{5.11}$$

$$\eta(-a, z) = 1, \;\; \eta(a, z) = 1, \;\; \eta(x, b) = 1, \;\; \eta(x, 0) = e^{\alpha^2 \epsilon^2 f^2(x)}. \tag{5.12}$$

To mimic radiation boundary conditions and avoid reflection of the outgoing wave we used "sponge boundaries" at $x = a$ and $z = b$ (as is done in the NCAR/MM5 mesoscale model [23] and others). The following values of the parameters were used, respectively, in these simulations:

$$\epsilon = 0.35, \;\; N = 1, \;\; \beta = 4.10^{-3} \tag{5.13}$$

with topography shape function

$$f(x) = \frac{1}{(1 + x^2)^{3/2}}. \tag{5.14}$$

The convergence criterion for the iterations was $\max |\eta_{m+1} - \eta_m| \leq 10^{-10}$. Figures 1 and 2 compare the results obtained for the perturbation $\phi(x, z)$ by using (5.9) and (5.10). We see that Figure 1 (for the shearless base flow) displays a clear pattern of gravity waves. On the other hand, Figure 2 shows that the perturbation from the shear flow feeds on the energy of the base flow and creates a vortex high above the topography.

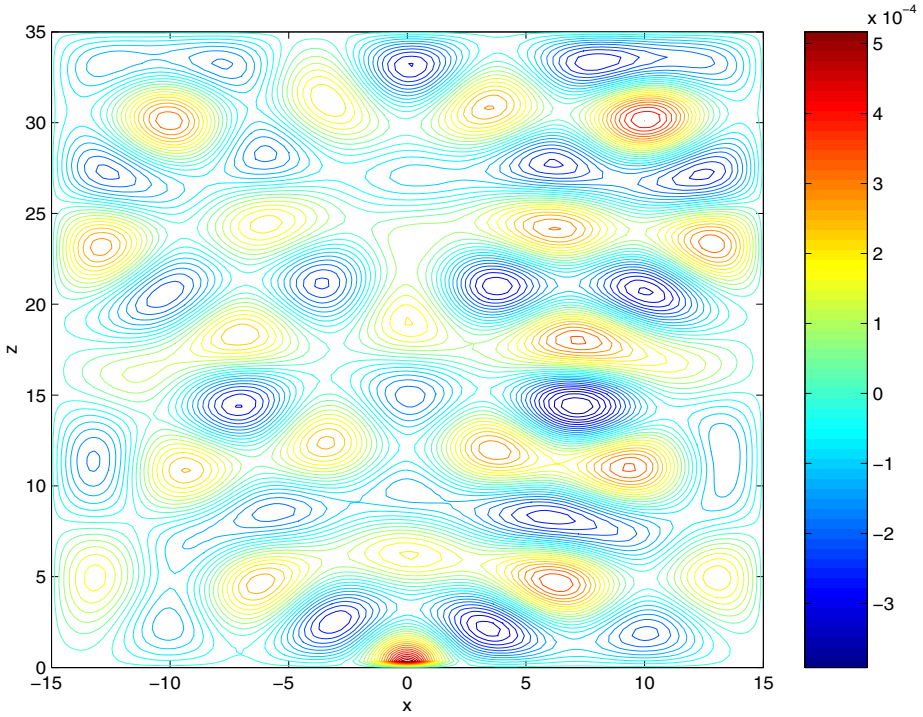


FIG. 1. *Contour plot of* $\alpha^2\phi$ *using* (5.9) *(shearless base flow).*

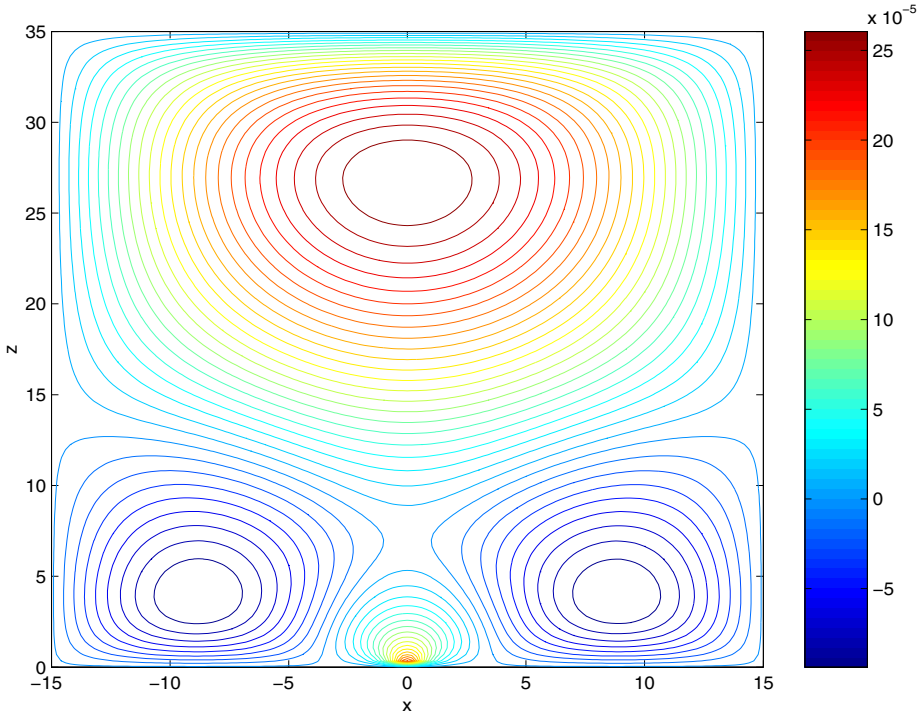


FIG. 2. *Contour plot of* $\alpha^2\phi$ *using* (5.10) *(shear base flow).*

6. Summary and conclusions. We derived in this paper some criteria for the excitation of gravity waves by a flow over topography using Long's equation. These criteria depend on only the nature of the base flow and the variation of N^2 with height. From an operational point of view these criteria will be useful both experimentally and theoretically. Currently the experimental practice is to ignore the shear in the base flow and attempt to deduce the quantitative attributes of the gravity waves using the shearless Long's equation. This procedure can be refined now by taking this important feature into account. Our analysis also shows that no simulation of Long's equation over actual topography is needed to determine the qualitative nature of the perturbation that is generated by the topography.

We also demonstrated that in some cases this perturbation will be damped by the shear, while in other cases the perturbation will grow, feeding on the energy that is present in the base flow.

Appendix: Poincaré inequality.

THEOREM A.1. *Let $u(x)$ be a bounded differentiable function on $[0, a]$ with $u(0) = 0$; then*

$$\int_0^a [u'(x)]^2 dx \geq \frac{\pi^2}{4a^2} \int_0^a u^2(x) dx. \tag{A.1}$$

Proof. To prove this inequality we introduce

$$h(x) = \frac{\pi}{2a} \tan\left(\frac{\pi(x-a)}{2a}\right). \tag{A.2}$$

This function satisfies $h(a) = 0$ and the differential equation

$$h' - h^2 = \frac{\pi^2}{4a^2}. \tag{A.3}$$

We now consider the integral

$$\int_0^a [uh + u']^2 dx \geq 0, \tag{A.4}$$

$$\int_0^a [uh + u']^2 dx = \int_0^a u^2 h^2 dx + \int_0^a (u')^2 dx + 2 \int_0^a uhu' dx \geq 0, \tag{A.5}$$

but

$$\begin{aligned} \int_0^a uu'h dx &= \frac{u^2 h}{2}\bigg|_0^a - \frac{1}{2} \int_0^a u^2 h' dx \\ &= \frac{1}{2}[u^2(a)h(a) - u^2(0)h(0)] - \frac{1}{2} \int_0^a u^2 h' dx \\ &= -\frac{1}{2} \int_0^a u^2 h' dx. \end{aligned} \tag{A.6}$$

Hence from (A.5),

$$\int_0^a (u')^2 dx \geq \int_0^a u^2 h' dx - \int_0^a u^2 h^2 dx = \int_0^a u^2 (h' - h^2) dx = \frac{\pi^2}{4a^2} \int_0^a u^2 dx, \tag{A.7}$$

which proves the theorem. □

REFERENCES

[1] R. R. Long, *Some aspects of the flow of stratified fluids.* I. *Theoretical investigation*, Tellus, 5 (1953), pp. 42–57.

[2] R. R. Long, *Some aspects of the flow of stratified fluids.* II. *Theoretical investigation*, Tellus, 6 (1954), pp. 97–115.

[3] R. R. Long, *Some aspects of the flow of stratified fluids.* III. *Continuous density gradients*, Tellus, 7 (1955), pp. 341–357.

[4] R. R. Long, *The motion of fluids with density stratification*, J. Geophys. Res., 64 (1959), pp. 2151–2163.

[5] P. G. Drazin, *On the steady flow of a fluid of variable density past an obstacle*, Tellus, 13 (1961), pp. 239–251.

[6] P. G. Drazin and D. W. Moore, *Steady two dimensional flow of fluid of variable density over an obstacle*, J. Fluid Mech., 28 (1967), pp. 353–370.

[7] D. R. Durran, *Two-layer solutions to Long's equation for vertically propagating mountain waves*, Quart. J. Roy. Meteor. Soc., 118 (1992), pp. 415–433.

[8] D. K. Lily and J. B. Klemp, *The effect of terrain shape on nonlinear hydrostatic mountain waves*, J. Fluid Mech., 95 (1979), pp. 241–261.

[9] W. R. Peltier and T. L. Clark, *Nonlinear mountain waves in two and three spatial dimensions*, Quart. J. Roy. Meteor. Soc., 109 (1983), pp. 527–548.

[10] R. B. Smith, *Linear theory of stratified hydrostatic flow past an isolated mountain*, Tellus, 32 (1980), pp. 348–364.

[11] R. B. Smith, *Hydrostatic airflow over mountains*, Adv. Geophys., 31 (1989), pp. 1–41.

[12] C.-S. Yih, *Equations governing steady two-dimensional large amplitude motion of a stratified fluid*, J. Fluid Mech., 29 (1967), pp. 539–544.

[13] K. S. Davis, *Flow of Nonuniformly Stratified Fluid of Large Depth over Topography*, M.Sc. thesis in Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, 1999.

[14] M. Humi, *On the solution of Long's equation over terrain*, Il Nuovo Cimento C, 27 (2004), pp. 219–229.

[15] G. J. Shutts, M. Kitchen, and P. H. Hoare, *A large amplitude gravity wave in the lower stratosphere detected by radiosonde*, Quart. J. Roy. Meteor. Soc., 114 (1988), pp. 579–594.

[16] G. Y. Jumper and R. R. Beland, *Progress in the understanding and modeling of atmospheric optical turbulence*, in Proceedings of the AIAA Plasmadynamics and Lasers Conference, 2000, Paper AIAA-2000-2355.

[17] G. Y. Jumper, E. A. Murphy, A. J. Ratkowski, and J. Vernin, *Multisensor campaign to correlate atmospheric optical turbulence to gravity waves*, in Proceedings of the 42nd AIAA Aerospace Sciences Meeting and Exhibit, 2004, Paper AIAA-2004-1077.

[18] P. G. Baines, *Topographic Effects in Stratified Flows*, Cambridge University Press, New York, 1995.

[19] C. J. Nappo, *Atmospheric Gravity Waves*, Academic Press, Boston, 2002.

[20] C.-S. Yih, *Stratified Flows*, Academic Press, New York, 1980.

[21] E. L. Ince, *Ordinary differential equations*, in The Strumian Theory and Its Later Developments, Dover, New York, 1956, pp. 223–228.

[22] A. Ronveaux, ed., *Heun's Differential Equations*, Oxford University Press, Oxford, England, 1995.

[23] P. L. Haagenson, J. Dudhia, G. A. Grell, and D. R. Stauffer, *The Penn State/NCAR Mesoscale Model (MM5) Source Code Documentation*, NCAR Technical Note, NCAR/TN-392+STR, Mesoscale and Microscale Meteorology Division, National Center for Atmospheric Research, Boulder, CO, 1994.

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 1853–1870

EXTINCTION CRITERIA IN STAGE-STRUCTURED POPULATION MODELS WITH IMPULSIVE CULLING*

R. R. L. SIMONS† AND S. A. GOURLEY†

Abstract. We propose stage-structured population models for species whose adult members are subject to culling, with a view to understanding the culling regimes that are likely to result in eradication of the species. A purely time-dependent model is proposed in which culling occurs at particular discrete times, not necessarily equally spaced. Then a reaction-diffusion model is proposed for a situation in which the adults can diffuse; in this model the culling is continuous in time but occurs only at particular discrete points in space. Such a model might be appropriate for pheromone trapping of insects. For both models conditions are obtained that are sufficient for species eradication.

Key words. stage-structure, delay, impulse, pest control

AMS subject classifications. 34K45, 34K20, 35K57, 92D25

DOI. 10.1137/050637777

1. Introduction.

Many species are subject to some form of culling. Often this is for reasons of pest control, and the aim of culling in this case might well be the localized eradication of the pest. In other situations the reason for culling is simply to keep numbers under control for the protection of habitats or other species, and complete eradication is not the aim.

Unlike natural mortality which one might reasonably suppose to occur continuously, the mortality attributable to culling is often more likely to take place only at certain times. Sometimes these times may be prescribed by law, as in the case of game bird and wildfowl shooting in the UK, which takes place in prescribed seasons lasting only a few months. Also, where animals such as deer (which as adults have no natural predators in the UK or Ireland) are culled for habitat protection, culling often occurs only at certain times of the year. In the UK, badgers, which are believed to spread tuberculosis to cattle, are subjected to culling by trapping and shooting, but again there are restrictions on the timing of the culls in an attempt to reduce the problem of badger cubs being orphaned and starving to death. Crop spraying as a way to control insect pests is also a method of control likely to be happening at certain discrete times (sometimes chosen to coincide with critical stages in the insects' development).

One might also envisage situations where some form of culling takes place continuously in time but only at discrete points in space. A good example would be the trapping system used in Australia to control the blowfly *Lucilia cuprina* which is a substantial nuisance to sheep farmers. Female flies lay their eggs in a sheep's fleece. The eggs hatch into larvae which feed on the sheep's damaged skin, creating a wound that can attract other flies. The larval and pupal stages may total around 14 days [6]. One approach to controlling the fly populations is by using pesticides, but this raises concerns regarding pesticide residue on the wool as well as environmental and occupational health and safety. An alternative is to trap the blowflies using specially

*Received by the editors August 8, 2005; accepted for publication (in revised form) May 24, 2006; published electronically August 22, 2006.
http://www.siam.org/journals/siap/66-6/63777.html
†Department of Mathematics, University of Surrey, Guildford, Surrey, GU2 7XH, United Kingdom (s.gourley@surrey.ac.uk).

designed translucent buckets fixed to trees at about the height at which the blowflies work. Entrance cones allow blowflies to enter but not leave the buckets, which contain a chemical attractant which smells like the blowflies' food sources—rotting fleece, carcasses, urine, and feces. Manufacturers of the buckets offer advice regarding where they should be placed. The second model of the present paper, which we study in section 3, proposes a possible model for such trapping of blowflies continuously in time but only at discrete points in a one-dimensional space. The traps in our model do not have to be equally spaced apart, and neither do they all have to be equally effective.

The use of impulsive differential equations as models of pest control seems to be a relatively undeveloped application area. Liu, Zhang, and Chen [8], motivated by the topic of pest control, proposed and studied a Lotka–Volterra predator prey model with impulsive effects (but no delay). Their model exhibits complex dynamics including quasi periodicity and chaos. Models of vaccination are another obvious application area (Hui and Chen [7]). However, impulsive differential equations, as a topic in their own right, have received some attention. See, for example, Wu [10] or the book by Gopalsamy [4]. A number of papers give conditions for existence of periodic solutions and oscillation properties more generally, but this is not our interest in the present paper.

Section 2 of this paper analyzes a purely time-dependent model for culling that occurs only at particular discrete times, while section 3 analyzes a reaction-diffusion model incorporating culling that is continuous in time but discrete in space.

2. Culling at discrete times. In this section we propose a model for a stage structured population with two stages: immature and mature, in which births and naturally occurring deaths occur continuously but culling or trapping occurs only at certain particular times, namely at times t_j with $0 < t_1 < t_2 < \cdots < t_j < \cdots$ and $t_j \to \infty$ as $j \to \infty$. At the cull which occurs at time t_j a proportion b_j of the adult population is culled, causing a sharp decrease in the population and consequently a discontinuity in the evolution at time t_j.

Let $u(t,a)$ be the density of individuals at time t of age a, and assume that an individual becomes mature on reaching the age τ. We will assume that the total number of mature adults $u_m(t)$, defined by

$$u_m(t) = \int_\tau^\infty u(t,a)\,da,$$

obeys an evolution equation of the form

$$u_m'(t) = u(t,\tau) - d(u_m(t)) - \sum_{j=1}^{\infty} b_j u_m(t_j^-)\delta(t-t_j), \tag{2.1}$$

where $u(t,\tau)$ is the number of individuals of age exactly τ and therefore represents adult recruitment, $-d(u_m(t))$ is naturally occurring deaths, and the last term is the culling term. It will be assumed that the immatures are governed by the standard McKendrick–von Foerster model for an age-structured population, namely

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial a} = -\mu u, \qquad t > 0, \quad 0 < a < \tau, \tag{2.2}$$

with $\mu > 0$ constant, the initial condition

$$u(0,a) = u_0(a) \geq 0, \qquad a \geq 0, \tag{2.3}$$

and also the assumption that the birth rate $u(t,0)$ is a function of the total number of adults so that

$$u(t,0) = b(u_m(t)). \tag{2.4}$$

For the present section the mathematical assumptions on the death function $d(u_m)$ and the birth function $b(u_m)$ are listed in (2.7) below. As our results are for the linearized model it is the properties of these functions at low densities that matter in this paper. Two typical birth functions used in much of the literature seem to be $b(u_m) = Pu_m e^{-Au_m}$ and $b(u_m) = Pu_m^2 e^{-Au_m}$, both of which decrease at large densities due to crowding effects. Note that the second of these has $b'(0) = 0$, which is motivated by the fact that in some populations the per capita growth rate at low densities is very small due to lack of group defense and low mating probability. This function does not satisfy (2.7) below.

The solution of (2.2) subject to (2.3) and (2.4) is

$$u(t,a) = \begin{cases} u_0(a-t)\exp(-\mu t), & t < a, \\ b(u_m(t-a))\exp(-\mu a), & t > a. \end{cases} \tag{2.5}$$

From this expression we see that if $t > \tau$, then

$$u(t,\tau) = \exp(-\mu\tau) b(u_m(t-\tau)),$$

whereas if $t < \tau$, then $u(t,\tau) = u_0(\tau - t)\exp(-\mu t)$. Insertion of these expressions for $u(t,\tau)$ into (2.1) yields one nonautonomous evolution equation valid for times $t \in (0,\tau)$ and another autonomous delay equation valid for all times larger than τ. It is common practice in the literature on these types of models to consider only the latter equation, but to consider it for all times $t > 0$ with prescribed initial data on $[-\tau, 0]$. This is what we shall do in the present paper (model (2.6) below). This practice does raise certain issues related to initial data, an issue which is discussed in detail in Bocharov and Hadeler [2]. Strictly speaking, the initial data is prescribed at time $t = 0$ only and is just the function $u_0(a)$. One should proceed by first solving (2.1) with $u(t,\tau) = u_0(\tau - t)\exp(-\mu t)$ for t in the interval $(0,\tau)$, and then by solving the delay equation in (2.6) for times $t > \tau$. One can understand from this procedure that only certain initial data for problem (2.6) is actually related to the original problem. However, since this paper is concerned mainly with the linearized equations, we do not feel this will be too much of a concern.

Our model thus takes the form

$$\begin{aligned} &u_m'(t) = e^{-\mu\tau} b(u_m(t-\tau)) - d(u_m(t)) - \sum_{j=1}^{\infty} b_j u_m(t_j^-)\delta(t - t_j), \quad t > 0, \\ &u_m(t) = \phi(t) \geq 0 \quad \text{for} \quad t \in [-\tau, 0); \quad u_m(0) = u_m^0 > 0, \end{aligned} \tag{2.6}$$

where $\mu > 0$ represents juvenile mortality, $u_m(t)$ is the total number of adults at time t, $u_m(t_j^-)$ is the population just before the impulsive cull at time t_j, τ is the maturation time, b_j is the proportion of the mature species trapped or culled at time t_j, and δ denotes the Dirac delta function. In this model $b(u_m(t))$ is a function representing the birth rate of the species, and $d(u_m(t))$ is the natural death rate of the mature species. The $e^{-\mu\tau} b(u_m(t-\tau))$ term is the rate at which immature individuals become mature, known as the maturation rate. This term incorporates the delay τ and is essentially the birth rate τ time units ago, corrected to allow for juvenile mortality.

Models having the form of (2.6) without impulsive effects have been considered in detail by Cooke, van den Driessche, and Zou [3].

In the present section we will assume the following:

$$
\begin{aligned}
&0 < t_1 < t_2 < \cdots < t_j \to \infty \quad \text{as} \quad j \to \infty,\\
&b_j \in [0,1] \quad \forall j = 1,2,3,\ldots,\\
&b(0) = 0, \quad b'(0) > 0, \quad b(u_m) > 0 \quad \forall u_m > 0,\\
&d(0) = 0, \quad d \in C^1[0,\infty), \quad d(u_m) > 0 \quad \forall u_m > 0.
\end{aligned} \tag{2.7}
$$

Note that if we integrate the delay equation in (2.6) from t_j^- to t_j^+, we obtain

$$u_m(t_j^+) = u_m(t_j^-) - b_j u_m(t_j^-).$$

As a consequence, model (2.6) can be reformulated as

$$
\begin{aligned}
&u_m'(t) = e^{-\mu\tau} b(u_m(t-\tau)) - d(u_m(t)), \quad t \neq t_j,\\
&u_m(t_j^+) = (1-b_j) u_m(t_j^-),\\
&u_m(t) = \phi(t) \geq 0 \quad \text{for} \quad t \in [-\tau, 0); \quad u_m(0) = u_m^0 > 0.
\end{aligned} \tag{2.8}
$$

The two formulations (2.6) and (2.8) of the model are both useful. For most of the analysis in this section we shall be concerned only with linearized versions of these models near the zero solution. The Laplace transform provides a powerful tool for the investigation of these linearized models, but one has to take careful note of the fact that the solution $u_m(t)$ of either (2.6) or the alternative formulation (2.8) will, in general, be discontinuous at the times t_j. The well-known formula

$$\mathcal{L}\{u'(t)\} = sU - u(0) \tag{2.9}$$

for the Laplace transform of the derivative of a function assumes the function $u(t)$ to be continuous for all $t > 0$. Here, U is the Laplace transform of u, and s is the transform variable. For a function $u(t)$ which is continuous except for discontinuous jumps at the times $t = t_j$, the corresponding formula is

$$\mathcal{L}\{u'(t)\} = sU - u(0) + \sum_{j=1}^{\infty} e^{-st_j} \left(u(t_j^-) - u(t_j^+) \right). \tag{2.10}$$

Due care needs to be taken on this issue; otherwise there is a possibility of the discontinuities being taken care of twice over, and if this happens incorrect results are produced by the analysis. Even though the solution of (2.6) will not be continuous, in the treatment of the linearized equation the Laplace transform of the derivative term needs to be calculated using the formula (2.9) which assumes continuity. The discontinuities in the solution are correctly furnished by the Laplace transform analysis because of the presence of the Dirac delta function in (2.6). The alternative approach would be to carry out a Laplace transform analysis of the linearization of (2.8). In this case the derivative term has to be dealt with using (2.10). It can be shown that the two approaches yield the same equation for the transformed state variable and are therefore equivalent. It must be stressed, however, that one has to stick to one approach or the other. The use of (2.10) in a Laplace transform analysis of the linearized version of (2.6) produces incorrect results.

2.1. Positivity. Next, we shall show that solutions of (2.6) or (2.8) enjoy a positivity preserving property.

PROPOSITION 2.1. *Assume* (2.7) *holds; then the solution* $u_m(t)$ *of* (2.6), *or the alternative formulation* (2.8), *satisfies* $u_m(t) \geq 0$ *for all* $t > 0$.

Proof. The proof is by the method of steps and starts by establishing positivity for $t \in (0, \tau]$. First note that positivity (in fact, strict positivity) holds if all the b_j are zero. In this case,

$$u_m'(t) \geq -d(u_m(t)) \qquad \text{when} \qquad t \in (0, \tau].$$

By comparison, $u_m(t) \geq \hat{u}_m(t)$ where $\hat{u}_m(t)$ is the solution of

$$\hat{u}_m'(t) = -d(\hat{u}_m(t)), \quad t \in (0, \tau],$$

satisfying $\hat{u}_m(0) = u_m^0 > 0$. From the assumptions on the function d contained within (2.7), it follows by Taylor's theorem that $d(\hat{u}_m(t)) = \hat{u}_m(t)d'(\theta(t))$ for some function $\theta(t)$. Therefore the above differential equation for $\hat{u}_m(t)$ has zero as one of its solutions and is also of such a form that, given initial data, we are assured of a unique solution. With $\hat{u}_m(0) > 0$ it follows that $\hat{u}_m(t) > 0$ for all $t > 0$; otherwise uniqueness is violated. Therefore $u_m(t) > 0$ for all $t \in (0, \tau]$ in the case when the b_j are all zero. From the method of steps it is clear that if the b_j are zero, then strict positivity of $u_m(t)$ holds for all $t > 0$.

The case when some or all of the b_j are nonzero does not represent a significant complication. They are all in $[0, 1]$, by (2.7), and so by (2.8) the solution is always reset from a nonnegative value to a nonnegative value at one of the times t_j (note, however that if one or more of the b_j is 1, then the solution is reset to zero at the corresponding time t_j, so *strict* positivity of solutions cannot be anticipated in this case). From what we have already shown the solution is certainly strictly positive before the first impulse time t_1, and at time t_1 is reset to some nonnegative value. An argument much like that described in the previous paragraph, but with initial time t_1 rather than 0, then assures us of the nonnegativity of $u_m(t)$ until the next time t_2 at which a resetting occurs, but then the argument just described applies again until the next time t_3 and so on. The proof of Proposition 2.1 is complete. ∎

2.2. Criteria for extinction. Linearizing (2.6) about the steady state $u_m = 0$, we get

$$u_m'(t) = e^{-\mu\tau} b'(0) u_m(t-\tau) - d'(0) u_m(t) - \sum_{j=1}^{\infty} b_j u_m(t_j^-) \delta(t - t_j). \tag{2.11}$$

Integrating from t_j^- to t_j^+ yields the following alternative formulation for the linearized equation:

$$\begin{aligned} &u_m'(t) = e^{-\mu\tau} b'(0) u_m(t-\tau) - d'(0) u_m(t), \quad t \neq t_j, \\ &u_m(t_j^+) = (1 - b_j) u_m(t_j^-). \end{aligned} \tag{2.12}$$

Remark 1. Positivity preservation, Proposition 2.1, also holds for the linearized problem (2.12).

2.2.1. The case when $e^{-\mu\tau}b'(0) < d'(0)$. In this subsection we will prove linear stability of the zero solution of (2.8) under the condition $e^{-\mu\tau}b'(0) < d'(0)$. The ecological interpretation of this condition is that, at low densities, adult recruitment is insufficient to outweigh naturally occurring deaths. Our result confirms that, as we would anticipate, under these circumstances the population will still become extinct when impulsive trapping or culling is introduced whatever the intensity and however frequent or infrequent the culling occurs.

THEOREM 2.2. *Let* (2.7) *hold and assume additionally that*

$$e^{-\mu\tau}b'(0) < d'(0). \tag{2.13}$$

Then the solution $u_m(t)$ *of the linearized problem* (2.12) *satisfies* $u_m(t) \to 0$ *as* $t \to \infty$.

Proof. Applying the Laplace transform

$$\mathcal{L}\{u(t)\} = \int_0^\infty u(t)e^{-st}\,dt$$

to (2.12), using formula (2.10) to take care of the anticipated discontinuities in the solution as explained earlier, and also noting that the Laplace transform of the delay term can be written as

$$\mathcal{L}\{e^{-\mu\tau}b'(0)u_m(t-\tau)\} = e^{-\mu\tau}b'(0)\left(\int_{-\tau}^0 u_m(\eta)e^{-s(\eta+\tau)}d\eta + e^{-s\tau}U\right),$$

where $U = U(s)$ is the Laplace transform of $u_m(t)$, (2.12) gives

$$\begin{aligned}[s - e^{-\mu\tau}b'(0)e^{-s\tau} + d'(0)]U = u_m(0) &- \sum_{j=1}^\infty e^{-st_j}(u_m(t_j^-) - u_m(t_j^+)) \\ &+ e^{-\mu\tau}b'(0)\int_{-\tau}^0 u_m(\eta)e^{-s(\eta+\tau)}\,d\eta.\end{aligned}$$

Using the impulse condition from (2.12) to replace $u_m(t_j^+)$, we get

$$\begin{aligned}[s - e^{-\mu\tau}b'(0)e^{-s\tau} + d'(0)]U = u_m(0) &+ e^{-\mu\tau}b'(0)\int_{-\tau}^0 u_m(\eta)e^{-s(\eta+\tau)}d\eta \\ &- \sum_{j=1}^\infty e^{-st_j}b_j u_m(t_j^-).\end{aligned} \tag{2.14}$$

Now define $y(t)$ by

$$\begin{aligned} y'(t) &= e^{-\mu\tau}b'(0)y(t-\tau) - d'(0)y(t), \qquad t > 0, \\ y(t) &= 0 \quad \text{for} \quad t \in [-\tau, 0); \quad y(0) = 1, \end{aligned} \tag{2.15}$$

the continuous analogy of (2.12) without impulses. It is easy to show (similarly to the proof of Proposition 2.1) that $y(t) > 0$ for all $t > 0$.

Applying the Laplace transform to (2.15), and letting $Y = Y(s) = \mathcal{L}\{y(t)\}$, gives

$$sY - 1 = e^{-\mu\tau}b'(0)\left[\int_{-\tau}^0 y(\xi)e^{-s(\xi+\tau)}d\xi + e^{-s\tau}Y\right] - d'(0)Y$$

so that, since $y(t) = 0$ for $t \in [-\tau, 0)$,

$$Y = \frac{1}{s + d'(0) - e^{-\mu\tau} b'(0) e^{-s\tau}} \tag{2.16}$$

and so

$$y(t) = \mathcal{L}^{-1}\left\{ \frac{1}{s + d'(0) - e^{-\mu\tau} e^{-s\tau} b'(0)} \right\}. \tag{2.17}$$

From this it is easy to see that $y(t) \to 0$ as $t \to \infty$. To deduce this conclusion it suffices (by the inversion formula for Laplace transforms) to show that all the poles of the function Y (i.e., the zeros of the denominator of (2.16)) are strictly in the left half of the complex plane. For a contradiction, assume a zero $\hat{s}$ exists satisfying $\operatorname{Re} \hat{s} \geq 0$. Then

$$|\hat{s} + d'(0)| = e^{-\mu\tau} b'(0) |e^{-\hat{s}\tau}| = e^{-\mu\tau} b'(0) e^{-\tau \operatorname{Re} \hat{s}} \leq e^{-\mu\tau} b'(0)$$

so that $\hat{s}$ lies in the closed disk in the complex plane centered at $-d'(0)$ and of radius $e^{-\mu\tau} b'(0)$. But condition (2.13) implies that this disk is entirely within the open left half of the complex plane, and this contradicts $\operatorname{Re} \hat{s} \geq 0$. Thus $y(t) \to 0$ as $t \to \infty$.

The denominator of the right-hand side of (2.16) appears on the left-hand side of (2.14). Dividing by this quantity and taking inverse Laplace transforms gives

$$\begin{aligned}
u_m(t) &= u_m(0) y(t) + \mathcal{L}^{-1}\left\{ \frac{e^{-\mu\tau} b'(0) \int_{-\tau}^{0} u_m(\eta) e^{-s(\eta+\tau)} d\eta}{s + d'(0) - e^{-\mu\tau} b'(0) e^{-s\tau}} \right\} \\
&\quad - \sum_{j=1}^{\infty} b_j u_m(t_j^-) \mathcal{L}^{-1}\left\{ \frac{e^{-s t_j}}{s + d'(0) - e^{-\mu\tau} b'(0) e^{-s\tau}} \right\} \\
&= u_m(0) y(t) + \mathcal{L}^{-1}\left\{ \frac{e^{-\mu\tau} b'(0) \int_{-\tau}^{0} u_m(\eta) e^{-s(\eta+\tau)} d\eta}{s + d'(0) - e^{-\mu\tau} b'(0) e^{-s\tau}} \right\} \\
&\quad - \sum_{j=1}^{\infty} b_j u_m(t_j^-) \int_0^t y(t-s) \delta(s - t_j)\, ds \\
&= u_m(0) y(t) + \mathcal{L}^{-1}\left\{ \frac{e^{-\mu\tau} b'(0) \int_{-\tau}^{0} u_m(\eta) e^{-s(\eta+\tau)} d\eta}{s + d'(0) - e^{-\mu\tau} b'(0) e^{-s\tau}} \right\} \\
&\quad - \sum_{j=1}^{\infty} b_j u_m(t_j^-) H(t - t_j) y(t - t_j),
\end{aligned} \tag{2.18}$$

where $H(t - t_j)$ is the Heaviside function. In this calculation we have used the convolution theorem for the Laplace transform.

Our intention is to deduce from this that $u_m(t) \to 0$ as $t \to \infty$ under condition (2.13). We already know that $y(t) \to 0$ under this condition. The second term in the expression (2.18) for $u_m(t)$ also tends to zero as $t \to \infty$. This is because it is the inverse Laplace transform of a ratio in which the numerator is an analytic function of s while the denominator has all of its zeros in $\operatorname{Re} s < 0$ as has already been shown.

From nonnegativity of $u_m(t)$ for $t > 0$, and strict positivity of $y(t)$, we know the sign of the last term in the expression (2.18) for $u_m(t)$ and so we can write

$$0 \leq u_m(t) \leq u_m(0) y(t) + \mathcal{L}^{-1}\left\{ \frac{e^{-\mu\tau} b'(0) \int_{-\tau}^{0} u_m(\eta) e^{-s(\eta+\tau)} d\eta}{s + d'(0) - e^{-\mu\tau} b'(0) e^{-s\tau}} \right\}.$$

Hence $u_m(t) \to 0$ as $t \to \infty$. □

2.2.2. The case when $e^{-\mu\tau}b'(0) > d'(0)$. In this subsection we shall show that the zero solution of (2.8) can also be asymptotically linearly stable (i.e., the population will be driven to extinction) in the case when adult recruitment outweighs deaths at low densities if culling occurs in sufficient measure and with sufficient frequency in the sense to be described below. Note that from the alternative formulation of the original model (2.8) if the b_j's are close to 1, then it means that aggressive culling is taking place and a large majority of the mature species population is wiped out at each time t_j. We can also see that even if the b_j's were exactly equal to 1 and all the mature species were wiped out, this would not necessarily cause extinction, because immatures conceived at a previous time may mature at a later date. However, it is reasonable to speculate that if the b_j's are close enough to 1 and culling takes place sufficiently frequently in some sense, then the population would be driven to extinction.

For reasons that will become clear later, we need to understand the properties of the function $\phi(t)$ defined by

$$\begin{aligned} &\phi'(t) = e^{-\mu\tau}b'(0)\phi(t-\tau) - e^{-\mu\tau}b'(0)\phi(t), \\ &\phi(t) = 0, \quad t \in [-\tau, 0), \quad \phi(0) = 1. \end{aligned} \tag{2.19}$$

PROPOSITION 2.3. *The solution $\phi(t)$ of (2.19) is strictly positive for all $t > 0$ and satisfies*

$$\lim_{t\to\infty} \phi(t) = \frac{1}{1 + e^{-\mu\tau}b'(0)\tau}. \tag{2.20}$$

Consequently, the quantity $\phi^* := \inf_{t\geq 0} \phi(t)$ satisfies $\phi^* > 0$.

Proof. Strict positivity of $\phi(t)$ for $t > 0$ follows from arguments similar to those in the first part of the proof of Proposition 2.1. Strict positivity together with (2.20) immediately yields the last statement in the proposition, that $\phi^* > 0$. Therefore, it remains to prove only (2.20). Taking the Laplace transform of (2.19) and letting $\Phi = \Phi(s)$ denote the Laplace transform of ϕ, we obtain

$$s\Phi - 1 = e^{-\mu\tau}b'(0)\left[\int_{-\tau}^{0} \phi(\xi)e^{-s(\xi+\tau)}d\xi + e^{-s\tau}\Phi\right] - e^{-\mu\tau}b'(0)\Phi.$$

Since $\phi(t) = 0$ for $t \in [-\tau, 0)$,

$$\Phi = \frac{1}{s - e^{-s\tau}e^{-\mu\tau}b'(0) + e^{-\mu\tau}b'(0)}$$

so that

$$\begin{aligned} \phi(t) &= \mathcal{L}^{-1}\left\{\frac{1}{s - e^{-s\tau}e^{-\mu\tau}b'(0) + e^{-\mu\tau}b'(0)}\right\} \\ &= \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} \frac{e^{st}\,ds}{s - e^{-s\tau}e^{-\mu\tau}b'(0) + e^{-\mu\tau}b'(0)} \end{aligned} \tag{2.21}$$

in which the integral is the standard Bromwich integral. The quantity σ in (2.21) can be taken as any real number which strictly exceeds the supremum of the real parts of the zeros of the denominator in the integrand. In this case we can take any real $\sigma > 0$ as we now explain. Evaluation of the integral (2.21) will be via Cauchy's

residue theorem, which requires us to identify the poles of the integrand, i.e., the zeros of its denominator. By inspection, one of these is clearly $s = 0$. We claim that the equation $s - e^{-s\tau}e^{-\mu\tau}b'(0) + e^{-\mu\tau}b'(0) = 0$ has no roots satisfying $\operatorname{Re} s \geq 0$ other than the root $s = 0$. Indeed, if $\operatorname{Re} s \geq 0$, then

$$|s + e^{-\mu\tau}b'(0)| = e^{-\mu\tau}b'(0)|e^{-s\tau}| \leq e^{-\mu\tau}b'(0)e^{-\tau \operatorname{Re} s} \leq e^{-\mu\tau}b'(0)$$

so that s is in the closed disk in $\mathbf{C}$ with center $-e^{-\mu\tau}b'(0)$ and radius $e^{-\mu\tau}b'(0)$. But this disk contains no points s with $\operatorname{Re} s \geq 0$ apart from $s = 0$. Therefore the poles of the integrand in (2.21) consist of the pole at $s = 0$ (which is easily checked to be simple) together with the remaining zeros of the integrand's denominator, all of which satisfy $\operatorname{Re} s < 0$. Evaluation of (2.21) by Cauchy's residue theorem gives an expression of the form

$$\phi(t) = \sum \operatorname{res}\left\{ \frac{e^{st}}{s - e^{-s\tau}e^{-\mu\tau}b'(0) + e^{-\mu\tau}b'(0)},\ \ s \in \mathbf{P} \right\}, \tag{2.22}$$

where $\mathbf{P}$ is the set of all roots of $s - e^{-s\tau}e^{-\mu\tau}b'(0) + e^{-\mu\tau}b'(0) = 0$. But we know that the roots of this equation are $s = 0$ together with other roots, all of which satisfy $\operatorname{Re} s < 0$. It is well known that for a function $f(s)$ of the form $f(s) = h(s)/k(s)$ with $h(s)$ and $k(s)$ analytic functions of s, $h(a) \neq 0$, $k(a) = 0$, and $k'(a) \neq 0$, that the residue of $f(s)$ at the simple pole $s = a$ is given by $\operatorname{res}\{f(s);\ s = a\} = h(a)/k'(a)$. Applying this formula to the calculation of the residue at any $s \in \mathbf{P}$ with $\operatorname{Re} s < 0$ yields that the residue is an exponentially decaying function of t. Therefore

$$\begin{aligned}\phi(t) &= \operatorname{res}\left\{ \frac{e^{st}}{s - e^{-s\tau}e^{-\mu\tau}b'(0) + e^{-\mu\tau}b'(0)},\ s = 0 \right\} \\ &\quad + \text{exponentially decaying terms in } t \\ &= \frac{1}{1 + e^{-\mu\tau}b'(0)\tau} + \text{exponentially decaying terms in } t,\end{aligned}$$

and thus (2.20) holds. The proof of Proposition 2.3 is complete. □

Remark 2. Although we are assured of the strict positivity of the quantity ϕ^* defined in the statement of Proposition 2.3, we point out that ϕ^* is not necessarily equal to the limit in (2.20). It can be shown that the convergence to the limit in (2.20) will be nonmonotone if $e^{-\mu\tau}b'(0)\tau$ is sufficiently large,

Our next main result, Theorem 2.4 below, presents some conditions under which extinction of the population is predicted. Even though the problem under consideration is the linearized problem (2.12), analysis thereof is difficult. Our method of analysis involves the use of the Euler–Maclaurin summation formula [1], a technique for converting sums to integrals or vice versa. We can only retain certain terms in the use of this formula (those that do not involve the Bernoulli numbers), and as a consequence the following theorem must be interpreted in an approximate sense. Nevertheless, it is quite insightful as we will discuss later. We draw the reader's attention to the function $t(\cdot)$ referred to in the statement of Theorem 2.4 below. This function is not uniquely defined, but a sensible choice would be one that is piecewise linear but smoothed at the integers so as to be differentiable. The function $t(\cdot)$ tells us something about the spacing of the impulse times t_j (for example, if its derivative t' is very small, then the impulse times are rather close together; under these circumstances we might expect that extinction would be more likely, and this is what Theorem 2.4 indeed predicts). Condition (2.23) in the theorem essentially states that

the impulses must occur sufficiently close together in some sense depending on the proportion of the species that is removed at each impulse and also, not surprisingly, on the per capita natural death rate and adult recruitment rate at low densities.

THEOREM 2.4. *Let* (2.7) *hold, and let* $t(\xi) : [0,\infty) \to [0,\infty)$ *be a strictly monotonically increasing differentiable function with the property that* $t(i) = t_i$, $i = 1, 2, 3, \ldots$, *and* $t(0) = 0$. *If*

$$e^{-\mu\tau} b'(0) > d'(0)$$

and

$$\inf_{j \in \mathbf{N}} \{ b_j - (e^{-\mu\tau} b'(0) - d'(0)) t'(j) \} > 0, \tag{2.23}$$

then the solution $u_m(t)$ *of the linearized problem* (2.12) *satisfies* $u_m(t) \to 0$ *as* $t \to \infty$ *according to an analysis based on the Euler–Maclaurin summation formula.*

Proof. It will be convenient to rewrite (2.12) in the form

$$\begin{aligned} &u_m'(t) = e^{-\mu\tau} b'(0) u_m(t-\tau) - e^{-\mu\tau} b'(0) u_m(t) + (e^{-\mu\tau} b'(0) - d'(0)) u_m(t), \quad t \neq t_j, \\ &u_m(t_j^+) = (1 - b_j) u_m(t_j^-). \end{aligned} \tag{2.24}$$

Taking Laplace transforms of (2.24) and using formula (2.10) gives

$$\begin{aligned} (s - e^{-\mu\tau} e^{-s\tau} b'(0) + e^{-\mu\tau} b'(0)) U = u_m(0) + e^{-\mu\tau} e^{-s\tau} b'(0) \int_{-\tau}^{0} e^{-s\xi} u_m(\xi)\, d\xi \\ + (e^{-\mu\tau} b'(0) - d'(0)) U - \sum_{j=1}^{\infty} b_j u_m(t_j^-) e^{-s t_j}. \end{aligned} \tag{2.25}$$

Using (2.21) and taking inverse Laplace transforms of (2.25), we get

$$\begin{aligned} u_m(t) &= f(t) + \mathcal{L}^{-1}\left\{ \frac{(e^{-\mu\tau} b'(0) - d'(0)) U}{s - e^{-\mu\tau} e^{-s\tau} b'(0) + e^{-\mu\tau} b'(0)} \right\} \\ &\quad - \sum_{j=1}^{\infty} b_j u_m(t_j^-) \mathcal{L}^{-1}\left\{ \frac{e^{-s t_j}}{s - e^{-\mu\tau} e^{-s\tau} b'(0) + e^{-\mu\tau} b'(0)} \right\} \\ &= f(t) + (e^{-\mu\tau} b'(0) - d'(0)) \int_0^t \phi(t-s) u_m(s)\, ds \\ &\quad - \sum_{j=1}^{\infty} b_j u_m(t_j^-) \int_0^t \phi(t-s) \delta(s - t_j)\, ds \\ &= f(t) + (e^{-\mu\tau} b'(0) - d'(0)) \int_0^t \phi(t-s) u_m(s)\, ds \\ &\quad - \sum_{j=1}^{\infty} b_j u_m(t_j^-) \phi(t - t_j) H(t - t_j), \end{aligned} \tag{2.26}$$

where we recall that $\phi(t)$ is defined by (2.19), and where

$$f(t) = u_m(0)\phi(t) + \mathcal{L}^{-1}\left\{ \frac{e^{-\mu\tau} b'(0) e^{-s\tau} \int_{-\tau}^{0} e^{-s\xi} u_m(\xi) d\xi}{s - e^{-\mu\tau} e^{-s\tau} b'(0) + e^{-\mu\tau} b'(0)} \right\}. \tag{2.27}$$

If we substitute $t = t_i^-$ into (2.26) and let

$$u_i = u_m(t_i^-), \qquad f_i = f(t_i^-),$$

we obtain, noting that $\phi(t)$ is continuous,

$$\begin{aligned} u_i &= f_i + (e^{-\mu\tau}b'(0) - d'(0)) \int_0^{t_i} \phi(t_i - s)u_m(s)ds - \sum_{j=1}^{i-1} b_j u_j \phi(t_i - t_j) \\ &= f_i + (e^{-\mu\tau}b'(0) - d'(0)) \int_0^{i} \phi(t(i) - t(\xi))u_m(t(\xi))t'(\xi)\, d\xi - \sum_{j=1}^{i-1} b_j u_j \phi(t_i - t_j), \end{aligned}$$

having made the substitution $s = t(\xi)$ in the integral term.

We now convert the integral in the above expression into a sum. This will be achieved by using a first approximation of the Euler–Maclaurin formula:

$$\int_0^n h(k)\, dk \approx \sum_{k=1}^{n-1} h_k + \frac{h(0) + h(n)}{2}. \tag{2.28}$$

Applying this, we get

(2.29)

$$\begin{aligned} u_i = f_i + (e^{-\mu\tau}b'(0) - d'(0)) &\left(\sum_{j=1}^{i-1} \phi(t_i - t_j)u_j t'(j) + \frac{\phi(t_i)u_m(0)t'(0) + \phi(0)u_i t'(i)}{2} \right) \\ &- \sum_{j=1}^{i-1} b_j u_j \phi(t_i - t_j). \end{aligned}$$

We now claim that the function $f(t)$ defined by (2.27) above tends to a strictly positive limit $C > 0$ as $t \to \infty$ (so that also $f_i \to C$ as $i \to \infty$). By Proposition 2.3, $\phi(t)$ certainly approaches a strictly positive limit. The second term in the expression for $f(t)$ does so as well, as can be shown similarly to a contour integral argument discussed earlier where the singularities were the same: a simple pole at the origin and various other poles all with strictly negative real part. By the inversion formula for Laplace transforms and Cauchy's residue theorem,

$$\begin{aligned} &\mathcal{L}^{-1}\left\{ \frac{e^{-\mu\tau}b'(0)e^{-s\tau} \int_{-\tau}^0 e^{-s\xi}u_m(\xi)\, d\xi}{s - e^{-\mu\tau}e^{-s\tau}b'(0) + e^{-\mu\tau}b'(0)} \right\} \\ &= \operatorname{res}\left\{ \frac{e^{-\mu\tau}b'(0)e^{st}\, e^{-s\tau} \int_{-\tau}^0 e^{-s\xi}u_m(\xi)\, d\xi}{s - e^{-\mu\tau}e^{-s\tau}b'(0) + e^{-\mu\tau}b'(0)},\ s = 0 \right\} \\ &\quad + \text{exponentially decreasing terms in } t. \end{aligned}$$

Thus

$$\lim_{t\to\infty} \mathcal{L}^{-1}\left\{ \frac{e^{-\mu\tau}b'(0)e^{-s\tau} \int_{-\tau}^0 e^{-s\xi}u_m(\xi)\, d\xi}{s - e^{-\mu\tau}e^{-s\tau}b'(0) + e^{-\mu\tau}b'(0)} \right\} = \frac{e^{-\mu\tau}b'(0) \int_{-\tau}^0 u_m(\xi)\, d\xi}{1 + \tau e^{-\mu\tau}b'(0)}.$$

Hence $f(t)$ tends to a limit as $t \to \infty$. Writing (2.29) a different way, and recalling that $\phi(0) = 1$,

$$\begin{aligned} & u_i \left(1 - \frac{(e^{-\mu\tau}b'(0) - d'(0))t'(i)}{2}\right) \\ &= \sum_{j=1}^{i-1} u_j \phi(t_i - t_j) \left[\left(e^{-\mu\tau}b'(0) - d'(0)\right) t'(j) - b_j\right] \\ &\quad + f_i + \tfrac{1}{2}(e^{-\mu\tau}b'(0) - d'(0))\phi(t_i)u_m(0)t'(0). \end{aligned}$$

Since f_i and $\phi(t_i)$ both approach limits as $i \to \infty$, there exists C^* such that the totality of the last two terms in the above expression is bounded above by C^* for all i. Using this fact, and also adding $\lambda \sum_{j=1}^{i-1} u_j$ to both sides,

$$\begin{aligned} & u_i \left(1 - \frac{(e^{-\mu\tau}b'(0) - d'(0))t'(i)}{2}\right) + \lambda \sum_{j=1}^{i-1} u_j \\ &\leq \sum_{j=1}^{i-1} u_j \left\{\lambda + \phi(t_i - t_j) \left[\left(e^{-\mu\tau}b'(0) - d'(0)\right) t'(j) - b_j\right]\right\} + C^*, \end{aligned}$$

with $\lambda > 0$ to be chosen. Recall that $\phi(t) \geq \phi^* > 0$, where ϕ^* is defined in the statement of Proposition 2.3, and note also that the hypotheses of the theorem imply that $(e^{-\mu\tau}b'(0) - d'(0))\, t'(j) - b_j < 0$ for each j. Hence

$$\lambda + \phi(t_i - t_j) \left[\left(e^{-\mu\tau}b'(0) - d'(0)\right) t'(j) - b_j\right] \leq \lambda + \phi^* \left[\left(e^{-\mu\tau}b'(0) - d'(0)\right) t'(j) - b_j\right]$$

which we should like to be negative for all j. Therefore we choose any $\lambda > 0$ such that

$$\lambda \leq \phi^* \inf_{j \in \mathbf{N}} \left\{b_j - \left(e^{-\mu\tau}b'(0) - d'(0)\right) t'(j)\right\},$$

which is possible because the infimum is strictly positive by hypothesis. With this choice of λ we have

$$u_i \left(1 - \frac{(e^{-\mu\tau}b'(0) - d'(0))t'(i)}{2}\right) + \lambda \sum_{j=1}^{i-1} u_j \leq C^*.$$

Finally note that $(e^{-\mu\tau}b'(0) - d'(0))t'(i) < b_i \leq 1$ for each i. Hence

$$\frac{1}{2}u_i + \lambda \sum_{j=1}^{i-1} u_j \leq u_i \left(1 - \frac{(e^{-\mu\tau}b'(0) - d'(0))t'(i)}{2}\right) + \lambda \sum_{j=1}^{i-1} u_j \leq C^*.$$

This is true for all i, and furthermore $u_i \geq 0$ for each i. Hence $\sum_{j=1}^{\infty} u_j < \infty$, and so $u_i \to 0$ as $i \to \infty$. The proof is complete. □

3. Culling at discrete points in space. Up to now we have examined a purely time-dependent model in which the culling occurs only at specific times. The present section will examine a reaction-diffusion model for the situation in which the adults (but not the juveniles) can move around in a random way and where culling occurs

continuously in time but only at specific points x_j in a one-dimensional infinite spatial domain $x \in (-\infty, \infty)$. The equation we will analyze is

$$\begin{aligned} \frac{\partial u_m}{\partial t}(x,t) &= D\frac{\partial^2 u_m}{\partial x^2}(x,t) + e^{-\mu\tau} b(u_m(x,t-\tau)) - d(u_m(x,t)) \\ &\quad - \sum_{j=-\infty}^{\infty} B_j u_m(x_j,t)\delta(x-x_j), \\ u_m(x,t) &= \phi(x,t) \geq 0 \quad \text{for} \quad (x,t) \in (-\infty,\infty)\times[-\tau,0] \\ &\quad \text{with } \phi(\cdot,t) \in L^2 \text{ for each } t \in [-\tau,0] \text{ and } u_m(x,0) \not\equiv 0. \end{aligned} \tag{3.1}$$

Model (3.1) is only appropriate if the juvenile members do not diffuse. This is because we are using the same derivation for the adult recruitment term $e^{-\mu\tau}b(u_m(x,t-\tau))$ as was used to derive model (2.6). However, if the juveniles diffuse, then a diffusion term would have to be added to (2.2) with the consequence that the solution of the latter would no longer be (2.5). Thus, our model (3.1) is for the case when only the adults diffuse. Fortunately, this assumption is quite realistic in many species. For example, in many insect species the juveniles are larvae and move very little or not at all. Locust larvae attach themselves to tree roots and do not move at all, whereas adult locusts can move great distances. The blowfly *Lucilia cuprina* larvae live in sheep and might move a little in the sense of being carried about by their host sheep within a farm, but it is only the adults that can move great distances and thereby transfer infestations from farm to farm.

Situations in which the juveniles do move appreciably can be studied too. As previously noted, one would need to add to (2.2) a term representing the mobility of the juveniles, with the consequence that instead of (3.1) we would have an equation containing a spatial nonlocality caused by the mobility of the juveniles. Such equations have been studied extensively in recent years; see, for example, So, Wu, and Zou [9] or the recent survey article by Gourley and Wu [5].

The quantities B_j, $j = 0, \pm 1, \pm 2, \ldots$, in (3.1) have a somewhat different ecological interpretation to the corresponding quantities b_j in model (2.6). The quantity B_j is not the proportion removed at x_j but rather is a measure of the culling effort at that location (as will become clear in the next paragraph) and can be any nonnegative number. It is reasonable to anticipate that if the B_j's are large, then the population would become extinct if either the x_j's are sufficiently close together or the diffusivity D is sufficiently large. This is because in the limiting case when the B_j's are all infinite, one can imagine that the problem effectively would decompose into infinitely many uncoupled problems each consisting of the partial differential equation in (3.1) on the finite domain consisting of the interval between two adjacent culling locations, subject to homogeneous Dirichlet boundary conditions.

The positioning of the delta function in (3.1) is such that the solution $u_m(x,t)$ will be continuous in x, but its derivative $\partial u_m/\partial x$ will not. If we integrate (3.1) from x_j^- to x_j^+, the result is

$$D\left[\left(\frac{\partial u_m}{\partial x}\right)_{x_j^+} - \left(\frac{\partial u_m}{\partial x}\right)_{x_j^-}\right] = B_j u_m(x_j,t). \tag{3.2}$$

Keeping in mind that the Laplacian representation for diffusion comes about from using the formula $J = -D\partial u_m/\partial x$ for the flux $J(x,t)$ (defined as the net rate at which individuals cross x in the positive x direction), then if we imagine the domain to be broken up into subdomains defined by the culling locations, (3.2) has the

interpretation that individuals that leave the subdomain $[x_j, x_{j+1}]$ at x_j do so either by being culled at x_j, or by entering the adjacent subdomain $[x_{j-1}, x_j]$. The culling effort at x_j is B_j, and the culling yield at this location is $B_j u_m(x_j, t)$ per unit time, i.e., proportional to the density at x_j. This leads us to expect that (3.1) should have a positivity preserving property, which is what we shall prove next. For the analysis of the present section, assumption (2.7) will be replaced by the following:

$$
\begin{aligned}
&\cdots < x_{-2} < x_{-1} < x_0 < x_1 < x_2 < \cdots \\
&\text{with } x_n \to \infty \text{ and } x_{-n} \to -\infty \text{ as } n \to \infty, \\
&B_j \geq 0 \quad \forall j = 0, \pm 1, \pm 2, \ldots, \\
&b(0) = 0, \quad b'(0) > 0, \quad b(u_m) > 0 \quad \forall u_m > 0, \\
&d(0) = 0, \quad d \in C^1[0, \infty), \quad d'(0) > 0, \quad d(u_m) > 0 \quad \forall u_m > 0.
\end{aligned} \tag{3.3}
$$

PROPOSITION 3.1. *Let* (3.3) *hold. Then all solutions of* (3.1) *which decay to zero as* $|x| \to \infty$ *for all* $t \geq 0$ *remain nonnegative for all* $t > 0$.

Proof. Let us make a C^1 extension to the definition of the death function to $u_m < 0$ by defining $d(u_m) = d'(0)u_m$ when $u_m < 0$. Then $d \in C^1(\mathbf{R})$. Let us first prove nonnegativity of $u_m(x,t)$ for $t \in (0, \tau]$ only. The proof is by contradiction. Suppose u_m goes negative on this time interval. Since $u_m(\pm\infty, t) = 0$, $u_m(x,t)$ must then attain a negative global minimum on the set $(x,t) \in (-\infty, \infty) \times (0, \tau]$. Let us first consider the possibility that the minimum is attained at a point (x^*, t^*) where x^* is not one of the culling sites x_j. Then x^* is in some open interval throughout which the delta function in (3.1) is inactive. Thus, $u_m(x^*, t^*) < 0$, $u_{m,xx}(x^*, t^*) \geq 0$, and $u_{m,t}(x^*, t^*) \leq 0$ (noting that the minimum could be at a point with $t^* = \tau$). Since $t^* - \tau \leq 0$, the adult recruitment term in (3.1) is nonnegative at (x^*, t^*). Using our extension of the death function to $u_m < 0$, it follows that

$$
\underbrace{\frac{\partial u_m}{\partial t}(x^*, t^*)}_{\leq 0} = \underbrace{D \frac{\partial^2 u_m}{\partial x^2}(x^*, t^*)}_{\geq 0} + e^{-\mu\tau} \underbrace{b(u_m(x^*, t^* - \tau))}_{\geq 0} - d'(0) \underbrace{u_m(x^*, t^*)}_{< 0},
$$

which is a contradiction. Now suppose that the negative global minimum is attained at a point (x^*, t^*) where x^* is one of the x_j. The delta function is active, and the above argument fails. As a function of x, the function $u_m(x,t)$ must now show cusp-like behavior, with $u_m(x^*, t^*) < 0$, $u_{m,x}(x^{*-}, t^*) \leq 0$, and $u_{m,x}(x^{*+}, t^*) \geq 0$ (if, for example, the second of these were violated, then, for x just larger than x^*, $u_m(x, t^*)$ would be below $u_m(x^*, t^*)$, contradicting (x^*, t^*) being the global minimum). Using this information in (3.2) at time t^* gives

$$
D \underbrace{\left(\frac{\partial u_m}{\partial x} \right)_{x^{*+}}}_{\geq 0} - D \underbrace{\left(\frac{\partial u_m}{\partial x} \right)_{x^{*-}}}_{\leq 0} = B_j \underbrace{u_m(x^*, t^*)}_{< 0},
$$

a contradiction. Thus $u_m(x,t) \geq 0$ for times $t \in (0, \tau]$. By the method of steps, $u_m(x,t) \geq 0$ for all $t > 0$, and the proof is complete. ∎

The linearization of (3.1) about the zero solution is

$$
\begin{aligned}
\frac{\partial u_m}{\partial t}(x,t) = D \frac{\partial^2 u_m}{\partial x^2}(x,t) &+ e^{-\mu\tau} b'(0) u_m(x, t-\tau) - d'(0) u_m(x,t) \\
&- \sum_{j=-\infty}^{\infty} B_j u_m(x_j, t) \delta(x - x_j).
\end{aligned} \tag{3.4}
$$

We will prove the following theorem giving conditions under which it is predicted that extinction will result. The quantity $B_{\inf}$ defined below embodies information on the spacing of the culling locations. The analysis uses the Euler–Maclaurin summation formula and therefore has to be interpreted in an approximate sense.

THEOREM 3.2. *Let* (3.3) *hold. Let* $X(\xi) : \mathbf{R} \to \mathbf{R}$ *be a strictly monotonically increasing differentiable function with the property that* $X(j) = x_j$ *for each* $j \in \mathbf{Z}$, *and let* $B(\xi) : \mathbf{R} \to [0, \infty)$ *be the piecewise linear function such that* $B(j) = B_j$ *for all* $j \in \mathbf{Z}$. *If*

$$e^{-\mu\tau} b'(0) < d'(0) + B_{\inf}, \tag{3.5}$$

where

$$B_{\inf} = \inf_{y \in \mathbf{R}} \left\{ \frac{B(y)}{X'(y)} \right\},$$

then, provided the derivative of the function $\xi \to B(\xi) u_m^2(X(\xi), t)$ *is not too high, the solution* $u_m(x, t)$ *of the linearized problem* (3.4) *satisfies* $u_m(x, t) \to 0$ *in* L^2 *as* $t \to \infty$, *according to an analysis based on the Euler–Maclaurin summation formula.*

Proof. First note the following alternative formula for $B_{\inf}$:

$$B_{\inf} = \inf_{y \in \mathbf{R}} \left\{ B(X^{-1}(y))(X^{-1})'(y) \right\}. \tag{3.6}$$

We multiply (3.4) by $u_m(x, t)$ and then integrate with respect to x over $(-\infty, \infty)$. As in the previous section, care needs to be taken to ensure that the effect of the delta function is not taken care of twice over. One approach (the approach we shall adopt) is to remove the last term in (3.4) and account for its presence in the way we treat the Laplacian term, using (3.2). The Laplacian term will be dealt with via integration by parts, and (3.2) will be used to take account of the effect of the discontinuities in the spatial derivative of u_m, thereby fully accounting for the effect of the delta function in (3.4). In fact,

$$\begin{aligned}
&D \int_{-\infty}^{\infty} u_m \frac{\partial^2 u_m}{\partial x^2}\, dx \\
&= D \sum_{j=-\infty}^{\infty} \int_{x_{j-1}}^{x_j} u_m \frac{\partial^2 u_m}{\partial x^2}\, dx \\
&= D \sum_{j=-\infty}^{\infty} \left(u_m(x_j, t) \frac{\partial u_m}{\partial x}(x_j^-, t) - u_m(x_{j-1}, t) \frac{\partial u_m}{\partial x}(x_{j-1}^+, t) - \int_{x_{j-1}}^{x_j} \left(\frac{\partial u_m}{\partial x} \right)^2 dx \right) \\
&= \sum_{j=-\infty}^{\infty} \left(- B_j u_m^2(x_j, t) + D u_m(x_j, t) \frac{\partial u_m}{\partial x}(x_j^+, t) - D u_m(x_{j-1}, t) \frac{\partial u_m}{\partial x}(x_{j-1}^+, t) \right. \\
&\quad \left. - D \int_{x_{j-1}}^{x_j} \left(\frac{\partial u_m}{\partial x} \right)^2 dx \right) \quad \text{using (3.2)} \\
&= - \sum_{j=-\infty}^{\infty} B_j u_m^2(x_j, t) - D \int_{-\infty}^{\infty} \left(\frac{\partial u_m}{\partial x} \right)^2 dx
\end{aligned}$$

since a telescoping series is involved. Therefore (3.4) becomes

$$\frac{1}{2}\frac{d}{dt}\|u_m(t)\|^2 = -D\int_{-\infty}^{\infty}\left(\frac{\partial u_m}{\partial x}\right)^2 dx + e^{-\mu\tau}b'(0)\int_{-\infty}^{\infty} u_m(x,t)u_m(x,t-\tau)\,dx$$

$$-d'(0)\|u_m(t)\|^2 - \sum_{j=-\infty}^{\infty} B_j u_m^2(x_j,t), \tag{3.7}$$

where

$$\|u_m(t)\| = \|u_m(\cdot,t)\| = \left(\int_{-\infty}^{\infty} u_m^2(x,t)\,dx\right)^{\frac{1}{2}}.$$

For compactness of notation, where $u_m(x,t)$ appears under a norm we shall write it simply as $u_m(t)$. Our aim is to show convergence of $u_m(x,t)$ to zero in L^2, i.e., that $\|u_m(t)\| \to 0$ as $t \to \infty$. From (3.7) it follows that

$$\|u_m(t)\|\frac{d}{dt}\|u_m(t)\| \le e^{-\mu\tau}b'(0)\|u_m(t)\|\|u_m(t-\tau)\| - d'(0)\|u_m(t)\|^2$$

$$-\sum_{j=-\infty}^{\infty} B_j u_m^2(x_j,t), \tag{3.8}$$

where we used the Cauchy–Schwarz inequality on the delay term.

Euler–Maclaurin summation can be used to approximate the last term in (3.8) as

$$\sum_{j=-\infty}^{\infty} B_j u_m^2(x_j,t) \approx \int_{-\infty}^{\infty} B(\xi)u_m^2(X(\xi),t)\,d\xi$$

which, on making the substitution $y = X(\xi)$, becomes

$$\int_{-\infty}^{\infty} B(X^{-1}(y))u_m^2(y,t)(X^{-1})'(y)\,dy \tag{3.9}$$

$$\ge \inf_{y\in\mathbf{R}}\left\{B(X^{-1}(y))(X^{-1})'(y)\right\}\|u_m(t)\|^2 \tag{3.10}$$

$$= B_{\inf}\|u_m(t)\|^2$$

by the alternative formula (3.6) for $B_{\inf}$. Using this estimate in (3.8) and dividing through by $\|u_m(t)\|$, we get

$$\frac{d}{dt}\|u_m(t)\| \le e^{-\mu\tau}b'(0)\|u_m(t-\tau)\| - (d'(0) + B_{\inf})\,\|u_m(t)\|. \tag{3.11}$$

From this, we can conclude (using similar methods to those discussed earlier) that $\|u_m(t)\| \to 0$ as $t \to \infty$ if

$$d'(0) + B_{\inf} > e^{-\mu\tau}b'(0), \tag{3.12}$$

which holds by hypothesis. The proof is complete. □

The quantity $B_{\inf}$ has the interpretation of being an infimum culling rate per unit density per unit length, and (3.5) states that it must exceed the adult recruitment rate minus the natural death rate, per unit density per unit length, at low densities.

Let us discuss the situations in which the Euler–Maclaurin summation as used here might lose its ability to predict accurate results. Essentially, we are assuming that the derivative of the function $\xi \to B(\xi)u_m^2(X(\xi),t)$ is not too high, and one situation in which this assumption might lose its validity is if the culling is aggressive but the culling sites are spaced far apart. Very aggressive culling would result in the population being effectively zero at the actual culling sites, but if these are far apart (or if there is very low diffusion), there is no reason why the species should not survive within at least some of the (now decoupled) subdomains $[x_j, x_{j+1}]$, essentially since individuals would be unlikely to wander into a culling site. This can be investigated by solving (3.4) (without the summation term) on the domain $x \in (x_j, x_{j+1})$ subject to homogeneous Dirichlet boundary conditions. Trial solutions of the form

$$u_m(x,t) = e^{\lambda t} \sin\left\{\frac{n\pi(x-x_j)}{x_{j+1}-x_j}\right\}, \qquad n = 1,2,3,\ldots,$$

exist whenever

$$\lambda + \frac{Dn^2\pi^2}{(x_{j+1}-x_j)^2} + d'(0) = b'(0)e^{-\mu\tau}e^{-\lambda\tau}, \tag{3.13}$$

which is another transcendental equation for λ that can be tackled using ideas similar to those presented earlier. Specifically it is possible to show that if

$$e^{-\mu\tau}b'(0) < d'(0) + \frac{D\pi^2}{(x_{j+1}-x_j)^2},$$

then all roots λ of (3.13) satisfy $\operatorname{Re}\lambda < 0$ for every $n = 1,2,3,\ldots$, giving a condition for extinction of the species inhabiting $[x_j, x_{j+1}]$, in this case of intensive culling at sites spaced far apart. This condition says that, at low densities, adult recruitment is not sufficient to offset deaths together with losses at the ends of the domain where culling is occurring. If the above condition is reversed, then one can show that (3.13) (with $n = 1$) has a real positive root λ, so that the species can survive in the subdomain $[x_j, x_{j+1}]$.

4. Discussion. For the purely time-dependent model the most important result we have proved concerning (2.8) is Theorem 2.4, which addresses the situation when, at low densities, adult recruitment outweighs natural mortality. In this situation condition (2.23) essentially describes culling regimes that will result in extinction. The condition involves the proportions b_j removed at the cull times t_j, and a function $t(\xi)$, the derivative of which can be viewed as a measure of the spacing of the cull times t_j.

From condition (2.23) one can make several inferences. If the culling effort is very small, i.e., at each cull only a small proportion b_j of the individuals are removed (which could still vary from cull to cull), then no matter how small this effort is, provided $\inf_{j\in\mathbf{N}} b_j > 0$, extinction can still result if the culling occurs sufficiently frequently in the sense that $t'(j)$ is sufficiently small for each j. A period of more aggressive culling (i.e., larger b_j for several consecutive j) can result in extinction even when the culls are less frequent. An obvious particular case is that in which the culls are equally spaced in time; i.e., $t_j = jT$ for $j = 1,2,3,\ldots$ and some constant $T > 0$, and the same proportion b^* is removed at each cull. In this case the only obvious choice for the function $t(\xi)$ is $t(\xi) = T\xi$, and thus condition (2.23) can be put in the form

$$e^{-\mu\tau}b'(0) < d'(0) + \frac{b^*}{T}$$

which says that, at low densities, the per capita death rate plus the proportion culled per unit time is too high to be compensated for by adult recruitment. Thus, the condition makes sense and is what we would expect in this particular case of a fixed proportion being culled at equally spaced culling times.

Condition (2.23) fails if even just one of the b_j is zero; i.e., there is a "cull," which we might call a zero cull, at which no animals are killed. However, provided only a finite number of the b_j are zero, there will exist a time beyond which all culls are "proper" culls (i.e., culls with $b_j > 0$), and one could shift the origin of time appropriately so that in condition (2.23) the infimum would be taken starting at the first proper cull having no subsequent zero culls. More interesting is the possibility of infinitely many zero culls. Mathematically, the most obvious solution is to remove them by relabelling the sequence t_j (i.e., passing to a subsequence of the original). This would, however, have the effect of changing the interpolating function $t(\xi)$ and in particular of increasing its derivative so that (2.23) would be less likely to hold. The outcome is that the population is less likely to be driven to extinction as expected.

For the model of section 3, which attempts to study culling continuously in time but at discrete points in space, one can draw inferences analogous to those above for the time-dependent model. The condition in Theorem 3.2 predicts extinction if the culling effort as described by the function $B(y)$ is sufficiently large in a sense that also involves the spacing apart of the culling sites (as described by the function $X(y)$) as we would anticipate. If the culling sites are close together, then X will have a small derivative and so B_{inf} is more likely to be large enough to satisfy (3.5).

Acknowledgment. We would like to thank the referees for their valuable suggestions which are much appreciated.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, Eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th printing, Dover, New York, 1972.

[2] G. Bocharov and K. P. Hadeler, *Structured population models, conservation laws, and delay equations*, J. Differential Equations, 168 (2000), pp. 212–237.

[3] K. Cooke, P. van den Driessche, and X. Zou, *Interaction of maturation delay and nonlinear birth in population and epidemic models*, J. Math. Biol., 39 (1999), pp. 332–352.

[4] K. Gopalsamy, *Stability and Oscillations in Delay Differential Equations of Population Dynamics*, Math. Appl. 74, Kluwer, Dordrecht, The Netherlands, 1992.

[5] S. A. Gourley and J. Wu, *Delayed non-local diffusion systems in biological invasion and disease spread*, in Nonlinear Dynamics and Evolution Equations, Fields Inst. Commun. 48, AMS, Providence, RI, 2006, pp. 137–200.

[6] W. S. C. Gurney, S. P. Blythe, and R. M. Nisbet, *Nicholson's blowflies revisited*, Nature, 287 (1980), pp. 17–21.

[7] J. Hui and L. Chen, *Impulsive vaccination of SIR epidemic models with nonlinear incidence rates*, Discrete Contin. Dyn. Syst. Ser. B., 4 (2004), pp. 595–605.

[8] B. Liu, Y. Zhang, and L. Chen, *Dynamic complexities in a Lotka-Volterra predator-prey model concerning impulsive control strategy*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 517–531.

[9] J. W.-H. So, J. Wu, and X. Zou, *A reaction-diffusion model for a single species with age structure. I. Travelling wavefronts on unbounded domains*, Proc. R. Soc. Lond. Ser. A. Math. Phys. Eng. Sci., 457 (2001), pp. 1841–1853.

[10] J. Wu, *A survey of impulsive systems: Theory and applications*, Appl. Math. Notes, 14 (1989), pp. 43–61.

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 1871–1894

RAY SOLUTION OF A SINGULARLY PERTURBED ELLIPTIC PDE WITH APPLICATIONS TO COMMUNICATIONS NETWORKS*

DIEGO DOMINICI† AND CHARLES KNESSL‡

Abstract. We analyze a second order, linear, elliptic PDE with mixed boundary conditions. This problem arose as a limiting case of a Markov-modulated queueing model for data handling switches in communications networks. We use singular perturbation methods to analyze the problem. In particular we use the ray method to solve the PDE in the limit where convection dominates diffusion. We show that there are both interior and boundary caustics, as well as a cusp point where two caustics meet, an internal layer, boundary layers, and a corner layer. Our analysis leads to approximate formulas for the queue length (or buffer content) distribution at the switch.

Key words. asymptotics, elliptic PDE, ray method, probability distribution

AMS subject classifications. 34E20, 60J20

DOI. 10.1137/050632683

1. Introduction.

In a model proposed by Anick, Mitra, and Sondhi [1], a buffer receives messages from N statistically independent and identical information sources, which asynchronously alternate between exponentially distributed periods in the "on" and "off" states. While "on," a source transmits data at unit rate. The buffer depletes through an output channel, with a given maximum rate of transmission C. The rate at which a source turns "on" is equal to λ, and the "off" rate is μ. If $C < N$, the buffer may be nonempty, and the condition $\frac{\lambda}{\lambda+\mu}N < C$ is needed for stability. This simply says that the mean number of "on" sources (each transmitting data at unit rate) must be less than the total transmission capacity of the channel. This model is analyzed exactly in [1], and the asymptotic limit $N \to \infty$, with $\frac{C}{N} = \frac{\lambda}{\lambda+\mu} + O\left(N^{-\frac{1}{2}}\right)$, is studied in [21]. This limit is referred to as "heavy traffic."

Analyzing the steady state joint probability distribution of the number of active sources and the buffer content involves solving a system of N linear ODEs. In heavy traffic this can be simplified to a backward-forward parabolic PDE of the type in (1.2). This model has the disadvantage of treating the buffer content as a deterministic fluid.

These types of fluid models have received much recent attention in the literature. They have been used as models of production lines with multiple stages [29], statistical multiplexers in asynchronous transfer mode (ATM) networks [25, 26, 30], packet speech multiplexers [31], buffer storage in manufacturing models [32], buffer memory in store-and-forward systems [15], and voice packet communications systems [8].

Among the main quantities of interest in fluid models are tail probabilities, i.e., the probability that the fluid or buffer level exceeds some prescribed large value. These were studied in [10, 27, 22]. The tail probabilities can be used to estimate loss

*Received by the editors May 30, 2005; accepted for publication (in revised form) June 1, 2006; published electronically August 25, 2006.
http://www.siam.org/journals/siap/66-6/63268.html

†Department of Mathematics, State University of New York at New Paltz, 75 S. Manheim Blvd., Suite 9, New Paltz, NY 12561-2443 (dominicd@newpaltz.edu). This author's research was supported by NSF grant DMS 99-73231, provided by Professor Floyd Hanson.

‡Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago (m/c 249), 851 South Morgan Street, Chicago, IL 60607-7045 (knessl@uic.edu). This author's research was partially supported by NSF grants DMS 99-71656, DMS 02-02815, DMS 05-03746 and NSA grant MDA 904-03-1-0036.

of information (of say, voice) on packet switched models. The tail behavior may be computed via the type of asymptotic analysis presented here.

A modification of the basic model in [1], which allows for service variability, is as follows. Again there are N independent and identical sources. When a source is "on" it generates a Poisson arrival stream to a queue. In the "off" state no arrivals are generated. The service time distribution is allowed to be general. The model just described may be called a Markov-modulated M/G/1 queue.

In [20] it is shown that the joint steady state distribution of the number of active sources, the queue length, and the elapsed service time of the customer currently being served satisfies a complicated system of integro-differential equations. In the heavy traffic limit, where $N \to \infty$ and the average arrival rate is close to the mean service rate, this system may be approximated by the following BVP:

$$\begin{aligned} Df_{yy} + (c-\xi)f_y + f_{\xi\xi} + (\xi f)_\xi &= 0, \qquad 0 < y < \infty,\ -\infty < \xi < \infty, \\ Df_y(0,\xi) + (c-\xi)f(0,\xi) &= 0, \qquad -\infty < \xi < \infty, \\ \int_{-\infty}^{\infty}\int_0^{\infty} f(y,\xi)\,dy\,d\xi &= 1. \end{aligned} \tag{1.1}$$

Here the variable y is related to the queue length, ξ corresponds to a scaled measure of the number of "on" sources above their mean value, $c > 0$ is the normalized excess of the service rate over the mean arrival rate, and $D > 0$ measures variability effects in the service time distribution.

The exact solution to (1.1) was analyzed in [20]. It is not completely explicit and involves finding one eigenvector of an infinite matrix, whose elements are complicated expressions involving Laguerre functions. This (infinite!) eigenvector must be computed numerically. In the same paper the limit $D \to \infty$ was considered. Now the matrix becomes diagonally dominant, and much more explicit results can be obtained.

The (highly singular) limit $D \to 0$ was studied in [9], resulting in a very complicated asymptotic solution involving contour integrals of parabolic cylinder and Airy functions. When $D = 0$ we see that the problem (1.1) degenerates into a parabolic one that is forward parabolic for $\xi > c$ and backward parabolic for $\xi < c$:

$$\begin{aligned} (c-\xi)\Im_y + \Im_{\xi\xi} + (\xi\Im)_\xi &= 0, \qquad 0 < y < \infty,\ -\infty < \xi < \infty, \\ \Im(0,\xi) &= 0, \qquad c < \xi, \\ \int_{-\infty}^{\infty} \Im(\infty,\xi)\,d\xi &= 1. \end{aligned} \tag{1.2}$$

Now $\Im$ is a density in ξ and a distribution in y. The problem (1.2) corresponds to the heavy traffic limit of the fluid model in [1]. Knessl and Morrison [21] derived the exact solution of (1.2). The limit $c \to \infty$ was studied in [22] by using the saddle point method and in [23] by using the ray method [18].

The study of backward-forward parabolic PDEs goes back to [12], and more recent analyses appear in [3, 4, 5, 11, 19]. Such problems arise in a wide variety of applications, such as counter-current separators [14], mean exit times [13], the Milne problem of statistical physics [6], neutron transport theory [16], and diffusion in spatially varying convection fields [17]. The interesting mathematical feature of these problems is that the initial (or boundary) conditions can be imposed only where the PDE is forward parabolic. This "half-boundary condition" makes these problems difficult to analyze.

In this paper we will solve (1.1) asymptotically in the limit $c \to \infty$ by using the ray method, the boundary layer method, and asymptotic matching [7]. In doing so, we shall analyze no fewer than seven different scales, and one more will be briefly discussed in the conclusion section. The asymptotic structure of (1.1) proves much more complicated than that of (1.2) in the same limit [23].

To analyze (1.1) for large c, it is convenient to introduce the new variables $\eta = \xi/c$ and $x = y/c$ and the small parameter $\varepsilon = c^{-2}$. Then (1.1) becomes the following problem for $F(x,\eta) = \varepsilon^{-1} f(y,\xi)$:

$$\begin{aligned} \varepsilon(DF_{xx} + F_{\eta\eta}) + (1-\eta)F_x + \eta F_\eta + F = 0, \quad & x \geq 0, \ -\infty < \eta < \infty, \\ D\varepsilon F_x(0,\eta) + (1-\eta)F(0,\eta) = 0, \quad & -\infty < \eta < \infty, \\ \int_{-\infty}^{\infty}\int_{0}^{\infty} F(x,\eta)dxd\eta = 1. & \end{aligned} \tag{1.3}$$

The boundary condition together with the normalization condition implies that the marginal distribution in η is the Gaussian

$$\int_0^\infty F(x,\eta)dx = \frac{1}{\sqrt{2\pi\varepsilon}} \exp\left(-\frac{\eta^2}{2\varepsilon}\right). \tag{1.4}$$

An important quantity to compute is the marginal distribution in the x variable, i.e.,

$$M(x) = \int_{-\infty}^{\infty} F(x,\eta)d\eta. \tag{1.5}$$

In section 2 we consider the case when x is close to 0 and $\eta < 1$; this will be very useful to match with other asymptotic solutions. Section 3 is dedicated to using the ray method to analyze (1.3) for $\varepsilon \to 0$ with x, η fixed. This yields asymptotic solutions in two main regions separated by the curve $x = \eta - \ln(\eta) - 1$, $\eta > 1$. We also derive boundary layer solutions for $x = O(\varepsilon^{\frac{2}{3}})$ and $\eta > 1$, $x = O(\varepsilon)$ and $\eta > 1$, a corner layer solution in the neighborhood of the point $(0,1)$, and in section 4 a transition layer solution along $x = \eta - \ln(\eta) - 1$. We show that all the solutions asymptotically match to each other in the appropriate limits and also agree with the approximation found in section 2. In section 5 we summarize and discuss the main results. In section 6 we check the identity (1.4) for $F(x,\eta)$ and compute the marginal distribution in x.

2. An expansion for small x. To solve (1.3) for ε small, we will first consider the scaling $x = O(\varepsilon)$. Thus we introduce the variable $v = x/\varepsilon$ and convert (1.3) into the problem

$$\begin{aligned} DF_{vv} + (1-\eta)F_v + \varepsilon(\eta F_\eta + F) + \varepsilon^2 F_{\eta\eta} = 0, \quad & v \geq 0, \ -\infty < \eta < \infty, \\ DF_v(0,\eta) + (1-\eta)F(0,\eta) = 0, \quad & -\infty < \eta < \infty, \\ \int_{-\infty}^{\infty}\int_{0}^{\infty} F(v,\eta)dvd\eta = \frac{1}{\varepsilon}. & \end{aligned} \tag{2.1}$$

On this scale (1.4) transforms to

$$\int_0^\infty F(v,\eta)dv = \frac{\varepsilon^{-\frac{3}{2}}}{\sqrt{2\pi}} \exp\left(-\frac{\eta^2}{2\varepsilon}\right). \tag{2.2}$$

We consider solutions to (2.1) which have the asymptotic form

$$F(v,\eta) \sim \frac{\varepsilon^{-\frac{3}{2}}}{\sqrt{2\pi}} \exp\left(-\frac{\eta^2}{2\varepsilon}\right)\left[F^{(0)}(v,\eta) + \sqrt{\varepsilon}F^{(1)}(v,\eta) + O(\varepsilon)\right]. \tag{2.3}$$

Substituting (2.3) into (2.1) and equating the coefficients of like powers of ε we get to leading order the equation

$$DF^{(0)}_{vv} + (1-\eta)F^{(0)}_v = 0 \tag{2.4}$$

with boundary condition

$$DF^{(0)}_v(0,\eta) + (1-\eta)F^{(0)}(0,\eta) = 0, \quad -\infty < \eta < \infty. \tag{2.5}$$

Solving for $F^{(0)}(v,\eta)$ and taking into account (2.2) we have obtained the following proposition.

PROPOSITION 2.1. *For $x = v\varepsilon = O(\varepsilon)$, (1.3) has the asymptotic solution to leading order*

$$F(v,\eta) \sim \varepsilon^{-\frac{3}{2}} \frac{1-\eta}{D\sqrt{2\pi}} \exp\left[-\frac{\eta^2}{2\varepsilon} - \frac{(1-\eta)v}{D}\right], \quad \eta < 1. \tag{2.6}$$

We see that for $\eta = \overline{\eta}\sqrt{\varepsilon}$, $\overline{\eta} = O(1)$, the solution decouples into a Gaussian in $\overline{\eta}$ times an exponential function of v,

$$F(v,\overline{\eta}) \sim \varepsilon^{-\frac{3}{2}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\overline{\eta}^2}{2}\right\} \times \frac{1}{D} \exp\left\{-\frac{v}{D}\right\}. \tag{2.7}$$

Such a decoupling was also observed in [20], where (1.3) was analyzed in the limit $D \to \infty$ with c fixed. The cases where x is small and $\eta > 1$ or $\eta \approx 1$ are treated in subsections 3.6 and 3.7.

3. The ray expansion. Now we consider solutions of (1.3) which have the asymptotic form

$$F(x,\eta) \sim \varepsilon^{\nu_1} \exp\left[\frac{1}{\varepsilon}\Psi(x,\eta)\right] K(x,\eta). \tag{3.1}$$

We substitute (3.1) into (1.3) and equate the coefficients of the lowest power of ε to get the eikonal equation for Ψ,

$$D\left(\Psi_x\right)^2 + \left(\Psi_\eta\right)^2 + \eta\left(\Psi_\eta - \Psi_x\right) + \Psi_x = 0, \quad \Psi_x(0,\eta) = \frac{\eta - 1}{D}. \tag{3.2}$$

Equating the coefficients of the next power of ε yields the transport equation for K,

$$DK\Psi_{xx} + K_x + 2DK_x\Psi_x + K\Psi_{\eta\eta} + \eta K_\eta + 2K_\eta\Psi_\eta - \eta K_x + K = 0, \quad K_x(0,\eta) = 0.$$

3.1. The rays. We solve (3.2) by introducing the characteristic curves or rays $[x(t),\eta(t)]$, written in terms of a parameter t. We first consider rays starting from the η-axis, and impose the initial conditions $[x(0),\eta(0)] = [0,s]$. The characteristic ODEs for (3.2) are

$$\begin{aligned}
\frac{dx}{dt} &= -2D\Psi_x + \eta - 1, \quad x(0) = 0,\\
\frac{d\eta}{dt} &= -2\Psi_\eta - \eta, \quad \eta(0) = s,\\
\frac{d\Psi_x}{dt} &= 0, \quad \frac{d\Psi_\eta}{dt} = \Psi_\eta - \Psi_x,\\
\frac{d\Psi}{dt} &= \Psi_x\frac{dx}{dt} + \Psi_\eta\frac{d\eta}{dt} = -D\left(\Psi_x\right)^2 - \left(\Psi_\eta\right)^2.
\end{aligned} \tag{3.3}$$

From (2.6) we note that $\Psi(0,\eta) = -\eta^2/2$, which implies that $\Psi\,[x(0),\eta(0)] = \Psi(0,s) = -s^2/2$.

Setting $\Psi_x(0,s) = A$, $\Psi_\eta(0,s) = B$ and solving (3.3) yields

$$\begin{aligned} x &= (A-B)e^t - (A+B+s)e^{-t} - (2DA+2A+1)t + 2B + s,\\ \eta &= (A-B)e^t + (A+B+s)e^{-t} - 2A,\\ \Psi_x &= A, \quad \Psi_\eta = (B-A)e^t + A,\\ \Psi &= -\frac{1}{2}(A-B)^2e^{2t} + 2A(A-B)e^t - A^2(D+1)t + AB - \frac{3}{2}A^2 + \frac{1}{2}B^2 - \frac{s^2}{2}. \end{aligned} \tag{3.4}$$

The constants A, B can be determined by evaluating the eikonal equation (3.2) at $x = 0$ (corresponding to $t = 0$), and also using the boundary condition from (1.3). This yields $A = \frac{s-1}{D}$ and $B = -s$ or $B = 0$. To decide which value of B is the right one, we take the derivative of Ψ with respect to s at $t = 0$,

$$-s = \frac{d}{ds}\Psi(0,s) = A\frac{d}{ds}x(0,s) + B\frac{d}{ds}\eta(0,s) = B. \tag{3.5}$$

Replacing A, B in (3.4) we get

$$\begin{aligned} x &= e^t - 1 - t - \frac{(D+1)(2t-e^t) + D + e^{-t}}{D}(s-1),\\ \eta &= e^t + \frac{e^{-t} + (D+1)e^t - 2}{D}(s-1),\\ \Psi &= -\frac{1}{2}e^{2t} + \frac{2e^t - (D+1)e^{2t} - 1}{D}(s-1)\\ &\quad + \frac{-1 + [4e^t - 2(t+1)](D+1) - e^{2t}(D+1)^2}{2D^2}(s-1)^2. \end{aligned} \tag{3.6}$$

For $t \geq 0$ and each value of s, the first two equations in (3.6) determine a ray in the (x,η)-plane, which starts from $(0,s)$ at $t = 0$. For $s = 1$ and $s = \frac{1}{D+1}$, we can eliminate t from (3.6) and obtain the explicit expressions

$$\begin{aligned} &x = X_0(\eta) = \eta - \ln(\eta) - 1, \quad s = 1, \quad \eta \geq 1,\\ &x = \frac{1}{D+1} - \eta - \ln(2 - \eta - D\eta), \quad s = \frac{1}{D+1}, \quad \frac{1}{D+1} \leq \eta < \frac{2}{D+1}. \end{aligned} \tag{3.7}$$

For $s > \frac{1}{D+1}$, we have both $x(t)$ and $\eta(t)$ increasing for $t > 0$. For $s = \frac{1}{D+1}$, $x(t)$ increases and $\eta(t)$ is asymptotic to $\frac{2}{D+1}$.

For $s < \frac{1}{D+1}$ the rays "turn around" and return to $x = 0$ for some $t^* > 0$, with $x(t^*) = 0$, $\eta(t^*) < s$. The maximum value in x reached by the ray occurs at $t = t_{x\,\max}$:

$$t_{x\,\max} = \ln\left[\frac{-2sD + D + 2 - 2s + \sqrt{D\,(4s^2D - 4sD - 8s + 4s^2 + D + 4)}}{2(1 - s - Ds)}\right].$$

For $0 < s < \frac{1}{D+1}$ the ray reaches its maximum in η at $t = t_{\eta\,\max}$:

$$t_{\eta\,\max} = \frac{1}{2}\ln\left[\frac{1-s}{1-s-Ds}\right], \quad \eta\,(t_{\eta\,\max}) = 2\frac{1-D}{s} + \frac{2s + 2D - 2 - s^2D - sD^2}{s\sqrt{(1-s)\,(1-s-Ds)}}.$$

For $s \leq 0$, $\eta(t)$ decreases for $0 < t < t^*$.

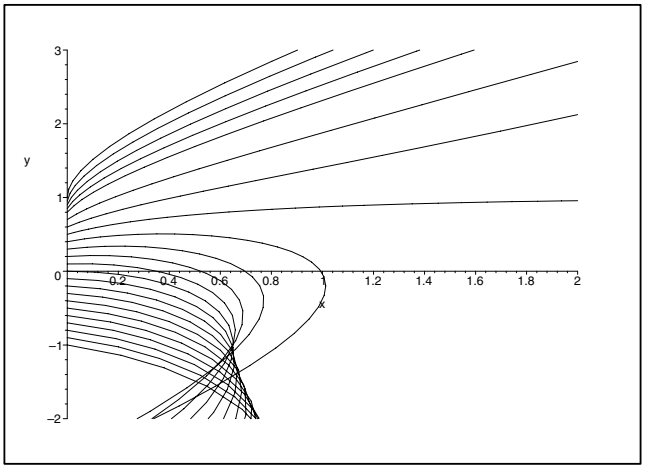


FIG. 3.1. *A sketch of the rays in Region* I *for* $D = 1$.

Solving for s in the η-equation (3.6) yields

$$s = \frac{e^{-t} + e^{t} - 2 + D\eta}{e^{-t} + (D+1)e^{t} - 2}, \tag{3.8}$$

and solving in the x-equation gives

$$s = \frac{-e^{t} + e^{-t} + Dt + 2t - Dx}{-De^{t} - e^{t} + e^{-t} + 2Dt + 2t + D}. \tag{3.9}$$

Equating (3.8) and (3.9) we get the implicit equation $R \equiv 0$ for the rays, where

$$\begin{aligned} R(x,\eta,t) = {} & \left[e^{-t} + (D+1)e^{t} - 2\right] x + (3 - D\eta - t - Dt - \eta)e^{t} + (1 + t + \eta)e^{-t} \\ & - 4 - 2t + D\eta + 2t\eta + 2D\eta t. \end{aligned}$$

We sketch several of the rays in Figure 3.1. They fill Region I, defined as

$$\text{Region I} \equiv \{x > X_0 = \eta - \ln(\eta) - 1, \quad \eta > 1\} \cup \{x > 0, \quad \eta \le 1\}. \tag{3.10}$$

3.2. Caustics and cusps. The Jacobian of the transformation in (3.6) from Cartesian to ray coordinates is

$$\begin{aligned} J = \frac{dx}{dt}\frac{d\eta}{ds} - \frac{dx}{ds}\frac{d\eta}{dt} = {} & \left[2(t-2)(s-1)D^{-2} + (-2t - 5s + 4ts + 2)D^{-1} - s + 2ts + 1\right] e^{t} \\ & + \left[-2(t+2)(s-1)D^{-2} + (2t - 2ts + 2 - 3s)D^{-1}\right] e^{-t} + 8(s-1)D^{-2} + 4(2s-1)D^{-1}. \end{aligned}$$

When $J = 0$ we can solve for s as a function of t, $S_0 = s\,|_{J=0}$:

$$S_0 = \frac{(-2D - D^2 - 4 + 2Dt + 2t)e^{2t} + 4(D+2)e^{t} - 2(2 + D + Dt + t)}{(-D^2 - 5D - 4 + 2t + 4Dt + 2tD^2)e^{2t} + 8(D+1)e^{t} - 3D - 4 - 2t - 2Dt}. \tag{3.11}$$

The equation for the *caustic(s)*, i.e., the points in the (x,η)-plane at which the Jacobian is zero, can be given in parametric form. We replace s by S_0 in the equation

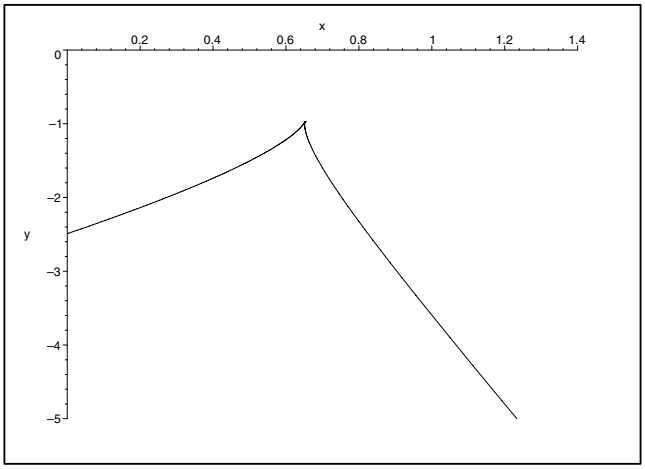


FIG. 3.2. *A sketch of the caustic curves for* $D = 1$.

of the rays, and let $x_{ca} = x(t, S_0)$, $\eta_{ca} = \eta(t, S_0)$:

(3.12)
$$x_{ca} = \left[-(D+1)^2 e^{3t} + (2D^2t^2 - 3tD + D^2t + 2t^2 - 4t + D^2 + 4t^2D + 6D + 8)e^{2t} - 2(3D+7)e^t - e^{-t} + 2(D+1)t^2 + (3D+4)t + 2(D+4)\right] / \left[(2D^2t + 4Dt - 4 + 2t - D^2 - 5D)e^{2t} + 8(D+1)e^t - (3D+4) - 2(D+1)t\right],$$

(3.13)
$$\eta_{ca} = \frac{-(D+1)^2 e^{3t} + 2(2tD + 2t + 2D - 1)e^{2t} + 2(4 - 2t - 2tD - D)e^t + e^{-t} - 6}{(2D^2t + 4Dt - 4 + 2t - D^2 - 5D)e^{2t} + 8(D+1)e^t - (3D+4) - 2(D+1)t}.$$

In Figure 3.2 we sketch the caustic curves for $D = 1$. There is also a cusp where the two caustics meet. Our numerical studies show that the basic structure (i.e., the two caustics coming together as a cusp) occurs for all $D > 0$.

Outside the caustic region, the correspondence between (t, s) and (x, η) is one-to-one. When we are exactly on the caustic curves, the correspondence is two-to-one, and inside the region bounded by the two caustics it is three-to-one. In Figure 3.3 we sketch more densely the rays for $D = 1$ to indicate this correspondence. The evaluation of (3.1) near caustics and cusps is discussed in more detail in section 5.

3.3. The transport equation. Now we shall solve the transport equation by using (3.3) to write it as an ODE along a ray:

$$\frac{dK}{dt} = (D\Psi_{xx} + \Psi_{\eta\eta} + 1)K. \tag{3.14}$$

After some algebra, we can show that

$$D\Psi_{xx} + \Psi_{\eta\eta} + 1 = \frac{1}{2} - \frac{1}{2J}\frac{dJ}{dt}, \tag{3.15}$$

and hence

$$K(x, \eta) = k(s)\frac{e^{\frac{t}{2}}}{\sqrt{J}}. \tag{3.16}$$

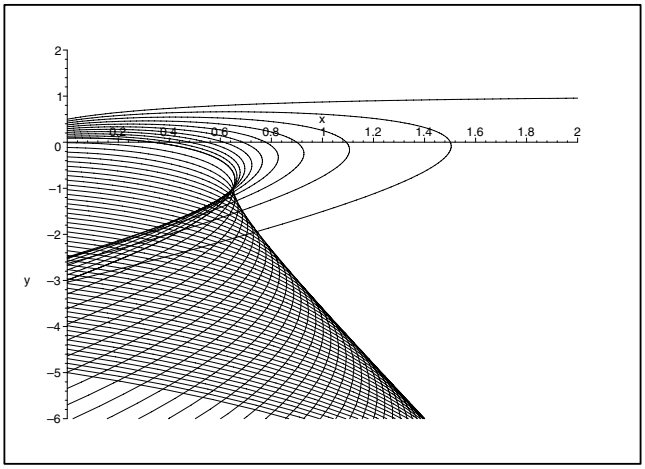


FIG. 3.3. *A sketch of the rays in Region* I *for* $D = 1$.

To determine $k(s)$ we evaluate the previous result at $t = 0$, $K(0, s) = k(s)\frac{1}{\sqrt{1-s}}$. Using the approximation (2.6) and the fact that $s = \eta$ at $t = 0$, we get

$$k(s) = \frac{1}{\sqrt{2\pi D}}(1-s)^{\frac{3}{2}}, \quad s < 1, \quad \text{and} \quad \nu_1 = -\frac{3}{2}. \tag{3.17}$$

The same result can be obtained by using the BC $K_x(0, \eta) = 0$ and fixing the multiplicative constant by normalization. So far we have determined Ψ and K only for $s < 1$. Thus we divide the half-plane $x \geq 0$, $-\infty < \eta < \infty$, into two parts. The portion filled by the rays for $s < 1$ we call Region I, and the remainder of the half-plane we call Region II. The latter is a shadow of the rays (see also Figure 3.1).

To summarize, we have established the following proposition.

PROPOSITION 3.1. *The solution of* (1.3) *in Region* I *is asymptotically given by*

$$F(x, \eta) \sim \varepsilon^{-\frac{3}{2}} K(x, \eta) \exp\left[\frac{1}{\varepsilon}\Psi(x, \eta)\right], \tag{3.18}$$

where

$$\begin{aligned} K(x, \eta) &= \frac{1}{\sqrt{2\pi}}(1-s)^{\frac{3}{2}} \frac{e^{\frac{t}{2}}}{\sqrt{J(t, s)}}, \\ \Psi(x, \eta) &= -\frac{1}{2}e^{2t} + \frac{2e^t - (D+1)e^{2t} - 1}{D}(s-1) \\ &\quad + \frac{-1 + [4e^t - 2(t+1)](D+1) - e^{2t}(D+1)^2}{2D^2}(s-1)^2, \end{aligned} \tag{3.19}$$

(x, η) is related to (t, s) by (3.6)*, and $J(t, s)$ is defined by* (3.11).

3.4. Region II. For this region, we consider solutions of (1.3) which have the asymptotic form

$$F(x, \eta) \sim \varepsilon^{\nu_2} \exp\left[\frac{1}{\varepsilon}\Phi(x, \eta) + \frac{1}{\varepsilon^{\frac{1}{3}}}\Gamma(x, \eta)\right] L(x, \eta). \tag{3.20}$$

The term $\varepsilon^{-\frac{1}{3}}\Gamma(x,\eta)$ in the exponent must be included in order for the expansion to asymptotically match those valid for small x and $\eta > 1$, which we construct later.

It follows that Φ satisfies (3.2), L satisfies the transport equation, and for Γ we get the following PDE:

$$(\eta - 1 - 2D\Phi_x)\Gamma_x - (2\Phi_\eta + \eta)\Gamma_\eta = 0, \tag{3.21}$$

which is equivalent to $\frac{d\Gamma}{d\tau} = 0$. Thus we conclude that Γ is a function of σ only and write $\Gamma(x,\eta) = \Gamma(\sigma)$. Here (τ,σ) are the new parameters for the ray which apply in Region II. Thus a ray starts at $\tau = 0$ from $\eta = \sigma > 1$ and enters the domain for $\tau > 0$.

The solutions of the characteristic equations are

$$\begin{aligned} x &= (b-a)e^{\tau} + (a+b-\sigma)e^{-\tau} + [2a(D+1)-1]\tau - 2b + \sigma, \\ \eta &= (b-a)e^{\tau} - (a+b-\sigma)e^{-\tau} + 2a, \\ \Phi_x &= -a, \quad \Phi_\eta = (a-b)e^{\tau} - a, \\ \Phi &= -a^2(D+1)\tau + 2a\,(a-b)\,(e^{\tau}-1) - \frac{1}{2}\,(a-b)^2\left(e^{2\tau}-1\right) + \Phi_0(\sigma). \end{aligned} \tag{3.22}$$

Here $\Phi_0(\sigma)$ is the value of Φ at $\tau = 0$, which corresponds to the η-axis for $\eta > 1$.

Since from the result for Region I $\frac{dx}{dt} = \frac{(s-1)}{D} = 0$ for $s = 1$, we impose the condition $\frac{dx}{d\tau}(0,\sigma) = 0$ for all $\sigma > 1$. This means that the boundary $x = 0$ will be a caustic curve for $\eta > 1$. Then a has the value

$$a(\sigma) = \frac{1-\sigma}{2D}. \tag{3.23}$$

Evaluating (3.2) at $x = 0$ we get

$$Da^2 + b^2 + \sigma(b-a) + a. \tag{3.24}$$

Using (3.23) in (3.24) and solving for b we find that

$$b = \frac{\sigma}{2} \pm \frac{\sqrt{\beta(\sigma)}}{2\sqrt{D}}, \quad \beta(\sigma) = D\sigma^2 + (\sigma-1)^2\,. \tag{3.25}$$

For small τ we get from (3.22) and (3.23) that $x \sim \left(b - \frac{\sigma}{2}\right)\tau^2$, $\tau \to 0$, and this implies that the solution $b = \frac{\sigma}{2} - \frac{\sqrt{\beta(\sigma)}}{2\sqrt{D}}$ must be rejected, in order that the rays enter the domain $x \geq 0$ as τ increases. Hence,

$$b(\sigma) = \frac{\sigma}{2} + \frac{\sqrt{\beta(\sigma)}}{2\sqrt{D}}. \tag{3.26}$$

To find $\Phi_0(\sigma)$ we impose the continuity condition $\Phi_0(1) = \Psi(0,1) = -\frac{1}{2}$. Since

$$\frac{d}{d\sigma}\Phi(0,\sigma) = -a\frac{d}{d\sigma}x(0,\sigma) - b\frac{d}{d\sigma}\eta(0,\sigma) = -b, \tag{3.27}$$

we conclude that

$$\begin{aligned} \Phi_0(\sigma) = -\frac{1}{2} - \int_1^{\sigma} b\,(u)\,du = -\frac{1}{4} - \frac{\sigma^2}{4} - \frac{1}{4\sqrt{D}}\left\{\left(\sigma - \frac{1}{D+1}\right)\sqrt{\beta(\sigma)}\right. \\ \left. + \frac{D}{(D+1)^{\frac{3}{2}}}\,\mathrm{arcsinh}\left[\frac{(D+1)\sigma - 1}{\sqrt{D}}\right] - \frac{D^{\frac{3}{2}}\sigma}{(D+1)} - \frac{D}{(D+1)^{\frac{3}{2}}}\,\mathrm{arcsinh}\left(\sqrt{D}\right)\right\}. \end{aligned} \tag{3.28}$$

As before, the transport equation can be solved to obtain

$$L(\tau,\sigma) = L_0(\sigma)\frac{e^{\frac{\tau}{2}}}{\sqrt{\widetilde{J}}}, \tag{3.29}$$

where

$$\begin{aligned}\widetilde{J} = \frac{dx}{d\tau}\frac{d\eta}{d\sigma} - \frac{dx}{d\sigma}\frac{d\eta}{d\tau} = &\left\{\left[-\sigma+1+\frac{1}{2}\tau(\sigma-1)\right]D^{-2} + \left[\frac{1}{2}\sqrt{\beta(\sigma)}(\tau-1)\right]D^{-\frac{3}{2}}\right.\\ &\left. + \left(-\sigma-\frac{1}{2}\tau+\tau\sigma\right)D^{-1} + \frac{1}{2}\tau\sqrt{\beta(\sigma)}D^{-\frac{1}{2}} + \frac{1}{2}\tau\sigma\right\}e^{\tau}\\ + &\left\{\left(\frac{1}{2}\tau+1\right)(1-\sigma)D^{-2} + \left[\frac{1}{2}\sqrt{\beta(\sigma)}(\tau+1)\right]D^{-\frac{3}{2}}\right.\\ &\left. + \left(-\sigma+\frac{1}{2}\tau-\tau\sigma\right)D^{-1} + \frac{1}{2}\tau\sqrt{\beta(\sigma)}D^{-\frac{1}{2}} - \frac{1}{2}\tau\sigma\right\}e^{-\tau} + 2(\sigma-1)D^{-2} + 2\sigma D^{-1}.\end{aligned}$$

In Region II, $\widetilde{J}=0$ only for $\tau=0$. To determine $L_0(\sigma)$ and $\Gamma(\sigma)$ we shall analyze the problem for small x, and we will find that not one but two boundary layer expansions are needed to satisfy the boundary conditions (1.3) in this region.

3.5. Approximation for $x = O(\varepsilon^{\frac{2}{3}})$, $\eta > 1$ (inner solution). We introduce the stretched variable $\mu = \varepsilon^{-\frac{2}{3}}x$ and transform (1.3) into

$$(1-\eta)F_\mu + \varepsilon^{\frac{1}{3}}DF_{\mu\mu} + \varepsilon^{\frac{2}{3}}(\eta F_\eta + F) + \varepsilon^{\frac{5}{3}}F_{\eta\eta} = 0. \tag{3.30}$$

We represent F in the asymptotic form

$$F \sim \varepsilon^{\nu_3}\exp\left\{\varepsilon^{-1}\Phi_0(\eta) + \varepsilon^{-\frac{1}{3}}\left[\frac{\eta-1}{2D}\mu + \Gamma(\eta)\right]\right\}\left[R_0(\mu,\eta) + \varepsilon^{\frac{1}{3}}R_1(\mu,\eta)\right], \tag{3.31}$$

which, when inserted into (3.30), gives the following PDEs for R_0, R_1:

$$\begin{aligned}&2D^2\frac{\partial^2 R_0}{\partial\mu^2} + [2\Phi_0'(\eta)+\eta][2D\Gamma'(\eta)+\mu]R_0 = 0,\\ &2D^2\frac{\partial^2 R_1}{\partial\mu^2} + [2\Phi_0'(\eta)+\eta][2D\Gamma'(\eta)+\mu]R_1\\ &\qquad + 2D\left\{[2\Phi_0'(\eta)+\eta]\frac{\partial R_0}{\partial\eta} + [\Phi_0''(\eta)+1]R_0\right\} = 0.\end{aligned} \tag{3.32}$$

Solving (3.32), we get

$$R_0 = C_1(\eta)\mathrm{Ai}\left\{2^{-\frac{1}{3}}D^{-\frac{5}{6}}\beta(\eta)^{\frac{1}{6}}[\mu + 2D\Gamma'(\eta)]\right\}, \tag{3.33}$$

where $\mathrm{Ai}(\cdot)$ denotes the Airy function and $\beta(\eta)$ is given by (3.25). Using (3.33) and solving for R_1, we obtain

$$\begin{aligned}R_1 = &\frac{1}{24}2^{\frac{2}{3}}D^{\frac{1}{6}}\beta(\eta)^{-\frac{5}{6}}C_1(\eta)\beta'(\eta)\overline{\mu}^2\mathrm{Ai}(\overline{\mu}) + [2D\beta(\eta)]^{\frac{1}{3}}C_1(\eta)\Gamma''(\eta)\overline{\mu}\mathrm{Ai}(\overline{\mu}) + C_2(\eta)\mathrm{Ai}(\overline{\mu})\\ &+ \left\{2^{\frac{2}{3}}[D\beta(\eta)]^{\frac{1}{6}}C_1'(\eta) - 2^{-\frac{1}{3}}D^{\frac{2}{3}}\beta(\eta)^{-\frac{1}{3}}C_1(\eta) + \frac{1}{3}2^{\frac{2}{3}}D^{\frac{1}{6}}\beta(\eta)^{-\frac{5}{6}}\alpha(\eta)C_1(\eta)\right\}\mathrm{Ai}'(\overline{\mu})\end{aligned} \tag{3.34}$$

with

$$\overline{\mu} = 2^{-\frac{1}{3}} D^{-\frac{5}{6}} \beta(\eta)^{\frac{1}{6}} \left[\mu + 2D\Gamma'(\eta)\right], \quad \alpha(\eta) = (D+1)\,\eta - 1. \tag{3.35}$$

The function $C_1(\eta)$ will be determined below. This solution can't satisfy the boundary condition (1.3), and thus we require another boundary layer expansion, where $x = o(\varepsilon^{\frac{2}{3}})$.

3.6. Approximation for $x = O(\varepsilon)$, $\eta > 1$ (inner-inner solution). We introduce the variable $v = x/\varepsilon$ and transform (1.3) to

$$DF_{vv} + (1-\eta)F_v + \varepsilon(\eta F_\eta + F) + \varepsilon^2 F_{\eta\eta} = 0, \quad DF_v(0,\eta) + (1-\eta)F(0,\eta) = 0. \tag{3.36}$$

We seek solutions of the form

$$F \sim \varepsilon^{\nu_4} \exp\left\{\frac{1}{\varepsilon}\Phi_0(\eta) + \frac{1}{\varepsilon^{\frac{1}{3}}}\Gamma(\eta) + \frac{1}{2}\frac{\eta-1}{D}v\right\} W(v,\eta). \tag{3.37}$$

Using (3.37) in (3.36) and taking into account that

$$\Phi_0'(\eta) = -b(\eta) = -\left(\frac{\eta}{2} + \frac{\sqrt{D\eta^2 + (\eta-1)^2}}{2\sqrt{D}}\right) \tag{3.38}$$

yields

$$DW_{vv} + (1-\eta)W_v + \Phi_0'(\eta)\left[\Phi_0'(\eta) + \eta\right] W = 0, \quad 2DW_v(0,\eta) + (1-\eta)W(0,\eta) = 0,$$

whose general solution is

$$W(v,\eta) = w(\eta)\left[\frac{1}{2D}(\eta-1)v + 1\right]. \tag{3.39}$$

The next step will be finding a corner layer solution valid in a neighborhood of the point $(0,1)$ that matches to both the approximation (2.6) and the inner-inner solution. This will allow us to determine ν_4 and $w(\eta)$ explicitly.

3.7. Corner layer. Let us first write $F(x,\eta) = \varepsilon^{\nu_5} \exp\left(-\frac{\eta^2}{2\varepsilon}\right) \overline{G}(x,\eta)$, which transforms (1.3) into

$$D\varepsilon\overline{G}_{xx} - \eta\overline{G}_\eta + \varepsilon\overline{G}_{\eta\eta} + (1-\eta)\overline{G}_x = 0, \quad D\varepsilon\overline{G}_x(0,\eta) + (1-\eta)\overline{G}(0,\eta) = 0. \tag{3.40}$$

Then we introduce the stretched variables $\mu = \varepsilon^{-\frac{2}{3}}x$ and $\gamma = \varepsilon^{-\frac{1}{3}}(\eta-1)$, and (3.40) becomes

$$\varepsilon^{\frac{2}{3}}\overline{G}_{\gamma\gamma} - \varepsilon^{\frac{1}{3}}\gamma\overline{G}_\gamma + D\overline{G}_{\mu\mu} - \gamma\overline{G}_\mu - \overline{G}_\gamma = 0, \quad D\overline{G}_\mu(0,\gamma) - \gamma\overline{G}(0,\gamma) = 0. \tag{3.41}$$

To leading order $\overline{G}(\mu,\gamma) \sim G(\mu,\gamma)$, where

$$DG_{\mu\mu} - \gamma G_\mu - G_\gamma = 0, \quad DG_\mu(0,\gamma) - \gamma G(0,\gamma) = 0. \tag{3.42}$$

The solution to (3.42) matches to (2.6) (with $\mu = 0$) if

$$\varepsilon^{\nu_5} G(0,\gamma) \sim \frac{1-\eta}{\sqrt{2\pi}D}\varepsilon^{-\frac{3}{2}} = -\frac{\gamma}{\sqrt{2\pi}D}\varepsilon^{-\frac{7}{6}}, \quad \gamma \to -\infty, \tag{3.43}$$

so that $\nu_5 = -\frac{7}{6}$. In [24] an explicit solution to (3.42) and (3.43) was obtained, with

$$G(\mu,\gamma) = \frac{\exp\left\{\frac{1}{\varepsilon}\left[\frac{\mu\gamma}{2D} - \frac{\gamma^3}{12D}\right]\right\}}{\sqrt{2\pi}2^{\frac{1}{3}}D^{\frac{2}{3}}}\frac{1}{2\pi i}\int_{Br}\exp\left\{2^{-\frac{2}{3}}D^{-\frac{1}{3}}\gamma\lambda\right\}\frac{\mathrm{Ai}\left(\lambda + 2^{-\frac{1}{3}}D^{-\frac{2}{3}}\mu\right)}{\left[\mathrm{Ai}\,(\lambda)\right]^2}d\lambda,$$

where Br is a vertical contour in the complex λ-plane on which $\mathrm{Re}(\lambda) \geq 0$ and $\mathrm{Ai}(\cdot)$ is the Airy function.

By combining the preceding results we have, on the corner scale,

$$\begin{aligned} F(x,\eta) &\sim \varepsilon^{-\frac{7}{6}}\exp\left\{\Psi_C(\mu,\gamma)\right\}L_C(\mu,\gamma) \equiv \widetilde{F}(\mu,\gamma), \\ \Psi_C(\mu,\gamma) &= -\frac{\eta^2}{2\varepsilon} + \frac{\mu\gamma}{2D} - \frac{\gamma^3}{12D}, \\ L_C(\mu,\gamma) &= \frac{1}{\sqrt{2\pi}2^{\frac{1}{3}}D^{\frac{2}{3}}}\frac{1}{2\pi i}\int_{Br}\exp\left\{2^{-\frac{2}{3}}D^{-\frac{1}{3}}\gamma\lambda\right\}\frac{\mathrm{Ai}\left(\lambda + 2^{-\frac{1}{3}}D^{-\frac{2}{3}}\mu\right)}{\left[\mathrm{Ai}\,(\lambda)\right]^2}d\lambda, \end{aligned} \tag{3.44}$$

where $\mu = \varepsilon^{-\frac{2}{3}}x$ and $\gamma = \varepsilon^{-\frac{1}{3}}(\eta - 1)$.

In [24, Theorem 4] several asymptotic expansions for a function closely related to (3.44) were obtained. We use these results in the following sections in order to match the different solutions that we have found so far and determine the unknown functions and constants.

3.8. Matching the solution in Region II and the inner solution. From (3.22) we get the local inversion between (τ,σ) and (x,η) for $x \to 0$:

$$\begin{aligned} \tau &\sim \sqrt{2}D^{\frac{1}{4}}\beta(\eta)^{-\frac{1}{4}}x^{\frac{1}{2}} + \frac{2}{3}\frac{\alpha(\eta)}{\beta(\eta)}x + \frac{\sqrt{2}}{36}\beta(\eta)^{-\frac{7}{4}}D^{-\frac{1}{4}}\left[14\beta(\eta) + 11D\beta(\eta) - 20D\right]x^{\frac{3}{2}}, \\ \sigma &\sim \eta - \sqrt{2}D^{\frac{1}{4}}\beta(\eta)^{-\frac{1}{4}}x^{\frac{1}{2}} + \frac{1}{3}\frac{\alpha(\eta)}{\sqrt{D\beta(\eta)}}x + \frac{\sqrt{2}}{36}\beta(\eta)^{-\frac{5}{4}}D^{-\frac{3}{4}}\left[10\beta(\eta) + D\beta(\eta) - 4D\right]x^{\frac{3}{2}}. \end{aligned} \tag{3.45}$$

Using (3.45) in (3.29), we obtain

$$L \sim L_0(\eta)2^{-\frac{1}{4}}\left[\frac{D}{\beta(\eta)}\right]^{\frac{1}{8}}x^{-\frac{1}{4}} + O(x^{\frac{1}{4}}), \quad x \to 0. \tag{3.46}$$

Expanding (3.33) for $\mu \to \infty$ yields

$$R_0 \sim C_1(\eta)\beta(\eta)^{-\frac{1}{24}}2^{-\frac{11}{12}}D^{\frac{5}{24}}\frac{1}{\sqrt{\pi}}\mu^{-\frac{1}{4}}\exp\left[-\frac{\sqrt{2}}{3}\beta(\eta)^{\frac{1}{4}}D^{-\frac{5}{4}}\mu^{\frac{3}{2}} - \sqrt{2}\beta(\eta)^{\frac{1}{4}}D^{-\frac{1}{4}}\Gamma'(\eta)\mu^{\frac{1}{2}}\right].$$

Using the above in (3.31) yields the expansion of F in (3.31) as $\mu \to \infty$. By expanding $\Phi(x,\eta)$ for small x we see that the exponential parts match automatically, and the matching of the algebraic factors implies that

$$\varepsilon^{\nu_2}L_0(\eta)2^{-\frac{1}{4}}\left[\frac{D}{\beta(\eta)}\right]^{\frac{1}{8}}x^{-\frac{1}{4}} = \varepsilon^{\nu_3}C_1(\eta)D^{\frac{5}{24}}\beta(\eta)^{-\frac{1}{24}}2^{-\frac{11}{12}}\frac{1}{\sqrt{\pi}}\mu^{-\frac{1}{4}}. \tag{3.47}$$

Hence, we have $\nu_2 = \nu_3 + \frac{1}{6}$ and

$$L_0(\eta) = C_1(\eta)D^{\frac{1}{12}}\beta(\eta)^{\frac{1}{12}}2^{-\frac{2}{3}}\frac{1}{\sqrt{\pi}}. \tag{3.48}$$

3.9. Matching the inner and inner-inner solutions. We take the limit $\mu \to 0$ in (3.33) and (3.34) to get

(3.49)
$$R_0 \sim C_1(\eta)\left\{\mathrm{Ai}\left[2^{\frac{2}{3}}D^{\frac{1}{6}}\beta(\eta)^{\frac{1}{6}}\Gamma'(\eta)\right] + \frac{1}{2}2^{\frac{2}{3}}D^{-\frac{5}{6}}\beta(\eta)^{\frac{1}{6}}\,\mathrm{Ai}'\left[2^{\frac{2}{3}}D^{\frac{1}{6}}\beta(\eta)^{\frac{1}{6}}\Gamma'(\eta)\right]\mu\right\}$$

and

$$\begin{aligned} R_1 \sim & \left[\frac{1}{6}\sqrt{\frac{D}{\beta(\eta)}}C_1(\eta)\beta'(\eta)\left(\Gamma'(\eta)\right)^2 + C_2(\eta) + 2\sqrt{D\beta(\eta)}C_1(\eta)\Gamma''(\eta)\Gamma'(\eta)\right] \\ & \times \mathrm{Ai}\left[2^{\frac{2}{3}}D^{\frac{1}{6}}\beta(\eta)^{\frac{1}{6}}\Gamma'(\eta)\right] \\ & + 2^{\frac{2}{3}}\left[D^{\frac{1}{6}}\beta(\eta)^{\frac{1}{6}}C_1'(\eta) - \frac{1}{2}D^{\frac{2}{3}}\beta(\eta)^{-\frac{1}{3}}C_1(\eta) + \frac{1}{3}D^{\frac{1}{6}}C_1(\eta)\alpha\beta^{-\frac{5}{6}}\right] \\ & \times \mathrm{Ai}'\left[2^{\frac{2}{3}}D^{\frac{1}{6}}\beta(\eta)^{\frac{1}{6}}\Gamma'(\eta)\right]. \end{aligned} \tag{3.50}$$

In order to complete the matching with the inner-inner solution, we must have

$$\varepsilon^{\nu_4}\, W(v,\eta)|_{v\to\infty} \sim \varepsilon^{\nu_3}\left[R_0(\mu,\eta) + \varepsilon^{\frac{1}{3}}R_1(\mu,\eta)\right]_{\mu\to 0}. \tag{3.51}$$

From (3.39), (3.49), and (3.50), we conclude that $\nu_3 + \frac{1}{3} = \nu_4$, and

$$\mathrm{Ai}\left[2^{\frac{2}{3}}D^{\frac{1}{6}}\beta(\eta)^{\frac{1}{6}}\Gamma'(\eta)\right] = 0, \tag{3.52}$$

$$C_1(\eta)\frac{1}{2}2^{\frac{2}{3}}D^{-\frac{5}{6}}\beta(\eta)^{\frac{1}{6}}\,\mathrm{Ai}'\left[2^{\frac{2}{3}}D^{\frac{1}{6}}\beta(\eta)^{\frac{1}{6}}\Gamma'(\eta)\right] = w(\eta)\frac{1}{2D}(\eta-1), \tag{3.53}$$

$$\begin{aligned} w(\eta) = & 2^{\frac{2}{3}}\left[D^{\frac{1}{6}}\beta(\eta)^{\frac{1}{6}}C_1'(\eta) - \frac{1}{2}D^{\frac{2}{3}}\beta(\eta)^{-\frac{1}{3}}C_1(\eta) + \frac{1}{3}D^{\frac{1}{6}}C_1(\eta)\alpha\beta^{-\frac{5}{6}}\right] \\ & \times \mathrm{Ai}'\left[2^{\frac{2}{3}}D^{\frac{1}{6}}\beta(\eta)^{\frac{1}{6}}\Gamma'(\eta)\right]. \end{aligned}$$

If we denote by r_0 the smallest (in absolute value) of the roots of Ai, i.e.,

$$r_0 = \max\{z : \mathrm{Ai}(z) = 0\} \simeq -2.33810741, \tag{3.54}$$

then we have from (3.52) that

$$\Gamma'(\eta) = 2^{-\frac{2}{3}}D^{-\frac{1}{6}}\beta(\eta)^{-\frac{1}{6}}r_0. \tag{3.55}$$

From (3.53)–(3.55) we obtain an ODE for $C_1(\eta)$,

$$C_1'(\eta) + \left[\frac{1}{4}\frac{\alpha(\eta)}{D\beta(\eta)} - \frac{1}{4\sqrt{D\beta(\eta)}} - \frac{1}{6}\frac{\alpha(\eta)}{\beta(\eta)} - \frac{1}{\eta-1}\right]C_1(\eta) = 0, \tag{3.56}$$

and a relation between $C_1(\eta)$ and $w(\eta)$,

$$w(\eta) = \frac{1}{\eta-1}C_1(\eta)2^{\frac{2}{3}}D^{\frac{1}{6}}\beta(\eta)^{\frac{1}{6}}\,\mathrm{Ai}'(r_0), \quad \eta > 1. \tag{3.57}$$

The solution of (3.56) is

$$C_1(\eta) = k_0(\eta-1)\beta(\eta)^{-\frac{1}{6}}\left[\frac{\alpha(\eta)}{\sqrt{D+1}} + \sqrt{\beta(\eta)}\right]^{\frac{\sqrt{D}}{2\sqrt{D+1}}} \tag{3.58}$$

with k_0 a constant to be determined.

3.10. Matching the corner and Region II solutions. From [24, Theorem 4(iv)] we have

(3.59)
$$\widetilde{F}(\mu,\gamma)\sim\varepsilon^{-\frac{7}{6}}L_{II}(\mu,\gamma)\exp\left\{\Psi_{II}(\mu,\gamma)\right\},\quad L_{II}(\mu,\gamma)=D^{-\frac{5}{6}}\frac{1}{\pi}\frac{1}{\left[\mathrm{Ai}'(r_0)\right]^2}2^{-\frac{29}{12}}\gamma\mu^{-\frac{1}{4}},$$
$$\Psi_{II}(\mu,\gamma)=-\frac{1}{2\varepsilon}\eta^2-\frac{1}{12D}\gamma^3+\frac{1}{2D}\mu\gamma-\frac{1}{3D}\sqrt{2}\mu^{\frac{3}{2}}+\frac{1}{2}2^{\frac{1}{3}}D^{-\frac{1}{3}}r_0\gamma-2^{-\frac{1}{6}}D^{-\frac{1}{3}}r_0\sqrt{\mu}$$
(3.60)

which are valid when μ and $\gamma\to\infty$, with $\gamma-\sqrt{\mu}\to\infty$.

Combining (3.46), (3.48), and (3.58), we have

$$L\sim k_0 2^{-\frac{11}{12}}\frac{1}{\sqrt{\pi}}\left[\frac{D}{\sqrt{D+1}}+\sqrt{D}\right]^{\frac{\sqrt{D}}{2\sqrt{D+1}}}(\eta-1)x^{-\frac{1}{4}},\quad x\to 0,\tag{3.61}$$

which agrees with (3.59) if

$$k_0=D^{-\frac{5}{6}}\frac{1}{\sqrt{\pi}}\frac{1}{\left[\mathrm{Ai}'(r_0)\right]^2}2^{-\frac{3}{2}}\left[\frac{D}{\sqrt{D+1}}+\sqrt{D}\right]^{-\frac{\sqrt{D}}{2\sqrt{D+1}}}.\tag{3.62}$$

Since in Region II $F(x,\eta)\sim\varepsilon^{\nu_2}\exp\left[\varepsilon^{-1}\Phi(x,\eta)+\varepsilon^{-\frac{1}{3}}\Gamma(x,\eta)\right]L(x,\eta)$, we must have $\nu_2=-\frac{4}{3}$. We use (3.45) in (3.22), (3.28), and (3.55) and find that, as $(x,\eta)\to(0,1)$,

$$\Phi(x,\eta)\sim-\frac{1}{2}-(\eta-1)-\frac{1}{2}(\eta-1)^2-\frac{1}{12D}(\eta-1)^3+\frac{1}{2D}x(\eta-1)-\frac{1}{3D}\sqrt{2}x^{\frac{3}{2}},$$
$$\Gamma(\sigma)\sim\Gamma(1)+\frac{1}{2}2^{\frac{1}{3}}D^{-\frac{1}{3}}r_0(\eta-1)-2^{-\frac{1}{6}}D^{-\frac{1}{3}}r_0\sqrt{x},$$

and from (3.60) we conclude that $\Gamma(1)=0$.

We have now determined all the unknown functions from the previous sections and summarize them below:

$$L(x,\eta)=D^{-\frac{3}{4}}(\sigma-1)\frac{1}{\pi}2^{-\frac{5}{2}}\beta(\sigma)^{-\frac{1}{12}}\left[\frac{\alpha(\sigma)+\sqrt{\beta(\sigma)(D+1)}}{D+\sqrt{D(D+1)}}\right]^{\frac{\sqrt{D}}{2\sqrt{D+1}}}\times\frac{1}{\left[\mathrm{Ai}'(r_0)\right]^2}\frac{e^{\frac{\tau}{2}}}{\sqrt{\widetilde{J}}},\tag{3.63}$$

$$\Gamma(\sigma)=2^{-\frac{2}{3}}D^{-\frac{1}{6}}r_0\int_1^\sigma\beta(u)^{-\frac{1}{6}}\,du,\tag{3.64}$$

$$R_0(\mu,\eta)=(\eta-1)D^{-\frac{5}{6}}\frac{1}{\sqrt{\pi}}2^{-\frac{3}{2}}\beta(\eta)^{-\frac{1}{6}}\left[\frac{\alpha(\eta)+\sqrt{\beta(\eta)(D+1)}}{D+\sqrt{D(D+1)}}\right]^{\frac{\sqrt{D}}{2\sqrt{D+1}}}\times\frac{\mathrm{Ai}\left[2^{-\frac{1}{3}}D^{-\frac{5}{6}}\beta(\eta)^{\frac{1}{6}}\mu+r_0\right]}{\left[\mathrm{Ai}'(r_0)\right]^2},\tag{3.65}$$

$$W(v,\eta) = 2^{-\frac{5}{6}}\frac{1}{\sqrt{\pi}}D^{-\frac{2}{3}}\left[\frac{\alpha(\eta)+\sqrt{\beta(\eta)(D+1)}}{D+\sqrt{D(D+1)}}\right]^{\frac{\sqrt{D}}{2\sqrt{D+1}}} \times \frac{1}{\mathrm{Ai}'(r_0)}\left[\frac{1}{2D}(\eta-1)v+1\right]. \tag{3.66}$$

With (3.63) and (3.64) we have completely determined the ray expansion in Region II, with (3.65) we have the inner solution (for $x = O(\varepsilon^{\frac{2}{3}})$ and $\eta > 1$), and with (3.66) we have the inner-inner solution (for $x = O(\varepsilon)$ and $\eta > 1$). We have also shown that $\nu_2 = -\frac{4}{3}$, $\nu_3 = -\frac{3}{2}$, and $\nu_4 = -\frac{7}{6}$.

4. Transition layer. Finally, we shall find the boundary layer solution near the curve $x = X_0(\eta)$ defined by (3.7), which separates Regions I and II. We introduce the stretched variable $\omega = (x - X_0)\varepsilon^{-\frac{1}{3}}$, and (1.3) becomes

$$-2\eta^2(\eta-1)F_\omega + \eta^2(\eta F_\eta + F)\varepsilon^{\frac{1}{3}} + \beta F_{\omega\omega}\varepsilon^{\frac{2}{3}} - \left[2\eta(\eta-1)F_{\omega\eta} + F_\omega\right]\varepsilon + \eta^2 F_{\eta\eta}\varepsilon^{\frac{4}{3}} = 0. \tag{4.1}$$

When $s = 1$ ($\sigma = 1$), $t = \ln(\eta)$ ($\tau = \ln(\eta)$), and we have

$$j = J\left[\ln(\eta),1\right] = 2\left(1+\frac{1}{D}\right)\ln(\eta)\,\eta + \frac{1}{D}\left(4-3\eta-\frac{1}{\eta}\right) = 2\widetilde{J}\left[\ln(\eta),1\right] = 2j_1. \tag{4.2}$$

Since

$$\Psi \sim -\frac{1}{2}\eta^2 - \frac{\eta}{2Dj}(x-X_0)^2, \quad x \to X_0, \tag{4.3}$$

we should look for solutions of the form

$$F \sim \varepsilon^{\nu_6}\exp\left\{-\frac{1}{2\varepsilon}\eta^2 - \frac{\eta}{2Dj}\omega^2\varepsilon^{-\frac{1}{3}}\right\}\Upsilon(\omega,\eta). \tag{4.4}$$

Using (4.4) in (4.1) yields for Υ the equation

$$2D^2j^2\omega\beta(\eta)\Upsilon_w + \eta^2D^3j^3\Upsilon_\eta + \beta(\eta)\left[D^2j^2 - 2\omega^3(\eta-1)\right]\Upsilon = 0, \tag{4.5}$$

whose general solution is

$$\Upsilon(\omega,\eta) = g\left(\frac{\eta\omega}{Dj}\right)\sqrt{\frac{\eta}{Dj}}\exp\left\{-\frac{\omega^3}{2\eta D^3j^3}\left[(2\eta-1)\left(2D\eta^2+2\eta^2-2\eta+1\right)\right]\right\}, \tag{4.6}$$

where g is a function still unknown. It will be determined in the next section by matching with the corner solution.

4.1. Matching the corner and transition layer solutions. Let us first introduce the new variable Ω defined by

$$\Omega = \frac{1}{(2D)^{\frac{1}{3}}}\left(\mu - \frac{1}{2}\gamma^2\right)\frac{1}{\gamma}. \tag{4.7}$$

From [24, Theorem 4(ii)] we have the following result for $\mu, \gamma \to \infty$, Ω fixed:

$$\varepsilon^{-\frac{7}{6}}e^{\frac{\eta^2}{2\varepsilon}}F(\mu,\gamma) \sim \varepsilon^{-\frac{7}{6}}\frac{2^{\frac{5}{6}}}{4\pi\sqrt{D\gamma}}\wp(\Omega)\exp\left\{\frac{\Omega^3}{6} - \frac{1}{4}\gamma\Omega^2 2^{\frac{2}{3}}D^{-\frac{1}{3}}\right\}, \tag{4.8}$$

where

$$\wp(\Omega) = \frac{1}{2\pi i}\int_{Br} \frac{e^{-\lambda\Omega}}{\left[\mathrm{Ai}\left(2^{\frac{1}{3}}\lambda\right)\right]^2} d\lambda. \tag{4.9}$$

The following properties of $\wp(\Omega)$ are established in [24]:

$$\wp(0) = 2^{-\frac{1}{3}}, \quad \wp(\Omega) \sim \Omega^{\frac{3}{2}}\sqrt{\pi}2^{-\frac{5}{6}}\exp\left\{-\frac{\Omega^3}{24}\right\}, \quad \Omega \to \infty,$$

$$\wp(\Omega) \sim -\frac{\Omega 2^{-\frac{2}{3}}}{\left[\mathrm{Ai}'(r_0)\right]^2}\exp\left\{-2^{-\frac{1}{3}}r_0\Omega\right\}, \quad \Omega \to -\infty.$$

In order to match with (4.6), we first note that

$$\omega \sim (2D)^{\frac{1}{3}}(\eta-1)\Omega, \quad \eta \to 1. \tag{4.10}$$

Thus, the right-hand side of (4.4) behaves as

$$\begin{aligned}&\varepsilon^{\nu_6} g\left[(2D)^{-\frac{2}{3}}\Omega\right]\frac{1}{\sqrt{2D(\eta-1)}}\exp\left\{-\frac{1}{8}\frac{2D+1}{D^2}\Omega^3 - \frac{1}{4}\Omega^2 2^{\frac{2}{3}}D^{-\frac{1}{3}}(\eta-1)\varepsilon^{-\frac{1}{3}}\right\}\\ &= \varepsilon^{\nu_6} g\left[(2D)^{-\frac{2}{3}}\Omega\right]\frac{1}{\varepsilon^{\frac{1}{6}}\sqrt{2D\gamma}}\exp\left\{-\frac{1}{8}\frac{2D+1}{D^2}\Omega^3 - \frac{1}{4}\Omega^2 2^{\frac{2}{3}}D^{-\frac{1}{3}}\gamma\right\}.\end{aligned}$$

Comparing the above with (4.8), we must have $\nu_6 = -1$ and

$$g\left[(2D)^{-\frac{2}{3}}\Omega\right] = \exp\left\{\frac{\Omega^3}{6} + \frac{1}{8}\frac{2D+1}{D^2}\Omega^3\right\}\frac{1}{\pi}2^{-\frac{2}{3}}\wp(\Omega), \tag{4.11}$$

which implies that

$$g(Z) = \exp\left\{\frac{Z^3}{6}\left(4D^2+6D+3\right)\right\}\frac{1}{\pi}2^{-\frac{2}{3}}\wp\left[(2D)^{\frac{2}{3}}Z\right]. \tag{4.12}$$

We conclude by writing the complete transition layer solution in (4.4):

$$\begin{aligned}F \sim \frac{1}{\varepsilon\pi}2^{-\frac{2}{3}}\sqrt{\frac{\eta}{Dj}}\wp\left[\frac{2^{\frac{2}{3}}}{D^{\frac{1}{3}}}\frac{\eta\omega}{j}\right]\exp\left\{-\frac{\eta^2}{2\varepsilon} - \frac{\eta}{2Dj\varepsilon^{\frac{1}{3}}}\omega^2 + \frac{1}{6}\left(4D^2+6D+3\right)\left(\frac{\eta\omega}{Dj}\right)^3\right.\\ \left. - \frac{\omega^3}{2\eta D^3 j^3}\left[(2\eta-1)\left(2D\eta^2+2\eta^2-2\eta+1\right)\right]\right\} \equiv \varepsilon^{-1}L_{X_0}(\omega,\eta)\exp\left\{\Psi_{X_0}(\omega,\eta;\varepsilon)\right\}.\end{aligned} \tag{4.13}$$

We can show that (4.13) matches to both of the solutions in Regions I and II.

5. Multivaluedness of the ray expansion. In that part of Region I outside the caustic region (cf. Figure 3.2) the mapping between (t,s) and (x,η) is one-to-one, and K and Ψ are unambiguously determined by the formulas in Proposition 3.1. Inside the caustic region the mapping is three-to-one, and we should rewrite (3.1) as

$$\varepsilon^{-\frac{3}{2}}\left[K_1\exp\left(\frac{1}{\varepsilon}\Psi_1\right) + K_2\exp\left(\frac{1}{\varepsilon}\Psi_2\right) + K_3\exp\left(\frac{1}{\varepsilon}\Psi_3\right)\right],$$

where Ψ_j and K_j correspond to the three different values of (t, s) leading to the same (x, η). When $t = 0$ let us define the starting points on the η-axis of these three rays by the ordering $s_1 < s_2 < s_3$, where s_j corresponds to Ψ_j and K_j. We denote the two caustics by C_+ and C_- and the cusp where they meet as (x_c, η_c). Note that the cusp location depends only on D.

The curve C_+ has $\eta \to -\infty$ as $x \to \infty$, while C_- reaches the η-axis at some critical point $(0, \eta_*)$ where again $\eta_* = \eta_*(D)$. We have verified numerically that along C_+ we have $s_1 = s_2$, $\Psi_1 = \Psi_2$, and K_1, K_2 develop singularities. However, here $\Psi_3 > \Psi_1 = \Psi_2$ and K_3 remains finite. Thus we have $F \sim \varepsilon^{-\frac{3}{2}} K_1 \exp\left(\frac{1}{\varepsilon}\Psi_1\right)$ on and near C_+. Similarly, along C_- we have $s_2 = s_3$, $\Psi_2 = \Psi_3$, and K_2, K_3 develop singularities. But $\Psi_1 > \Psi_2 = \Psi_3$, and K_1 remains finite. Thus the result in Proposition 3.1 remains valid near the caustics, except near the cusp point where all three Ψ_j are approximately equal. Here the expansion in Proposition 3.1 breaks down.

Our preliminary results suggest that a new expansion must be constructed near the cusp with the scaling

$$x - x_c = O(\sqrt{\varepsilon}), \quad \eta - \eta_c - A_c(x - x_c) = O(\varepsilon^{\frac{3}{4}}).$$

Here A_c is the slope at which both C_+ and C_- hit the cusp. We have thus far not been able to complete this analysis. We also note that while the expansion near the cusp presents an interesting problem in asymptotics, it is not needed for computing the marginal distribution $M(x)$ (1.5), which is the most important quantity from the point of view of applications, and which we calculate in the next section.

6. Marginal distributions. The last "piece of the puzzle" is to verify that (1.4) is satisfied, and also to compute the marginal distribution $M(x)$ in (1.5).

We evaluate the integral in (1.4) for $\varepsilon \to 0$. For $\eta < 1$, $F(x, \eta)$ is concentrated near $x = 0$, and the result follows from the approximation (2.6). The cases $\eta > 1$ and $\eta \approx 1$ will be considered below.

6.1. $\eta > 1$. In this region $F(x, \eta)$ is concentrated near $x = X_0$, and using (4.13) and (4.10) we have

$$\begin{aligned} F &\sim \exp\left\{-\frac{\eta^2}{2\varepsilon} - \frac{\eta}{2Dj\varepsilon^{\frac{1}{3}}}\omega^2\right\} \frac{1}{\varepsilon\pi} 2^{-\frac{2}{3}} \sqrt{\frac{\eta}{Dj}}\wp(0) \\ &= \exp\left\{-\frac{\eta^2}{2\varepsilon} - \frac{\eta}{2Dj\varepsilon}(x - X_0)^2\right\} \frac{1}{2\pi\varepsilon}\sqrt{\frac{\eta}{Dj}}, \quad x \to X_0, \end{aligned}$$

and hence, by Laplace's method,

$$\int_0^\infty F(x, \eta)dx \sim \int_{-\infty}^\infty \exp\left\{-\frac{\eta^2}{2\varepsilon} - \frac{\eta}{2Dj\varepsilon}(x - X_0)^2\right\} \frac{1}{2\pi\varepsilon}\sqrt{\frac{\eta}{Dj}}dx = \frac{1}{\sqrt{2\pi\varepsilon}} \exp\left(-\frac{\eta^2}{2\varepsilon}\right).$$

This verifies (1.4) (at least asymptotically as $\varepsilon \to 0$) for $\eta > 1$.

6.2. $\eta \approx 1$. For $\eta \to 1$ and x small we use the corner layer expansion, i.e.,

$$\begin{aligned} F(x, \eta) \sim{}& \varepsilon^{-\frac{7}{6}} \frac{1}{\sqrt{2\pi} 2^{\frac{1}{3}} D^{\frac{2}{3}}} \exp\left\{-\frac{\eta^2}{2\varepsilon} + \frac{\gamma\mu}{2D} - \frac{\gamma^3}{12D}\right\} \\ &\times \frac{1}{2\pi i}\int_{Br} \exp\left\{2^{-\frac{2}{3}} D^{-\frac{1}{3}}\gamma\lambda\right\} \frac{\operatorname{Ai}\left(\lambda + 2^{-\frac{1}{3}} D^{-\frac{2}{3}}\mu\right)}{[\operatorname{Ai}(\lambda)]^2} d\lambda, \end{aligned}$$

where $x = \mu\varepsilon^{\frac{2}{3}}$ and $\eta - 1 = \gamma\varepsilon^{\frac{1}{3}}$. In the local variable μ, (1.4) becomes

$$\int_0^\infty F(x,\eta)d\mu = \varepsilon^{-\frac{7}{6}}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{\eta^2}{2\varepsilon}\right\}; \tag{6.1}$$

thus we have to show that

$$\Lambda(\gamma) = 2^{\frac{1}{3}}D^{\frac{2}{3}}\exp\left\{\frac{\gamma^3}{12D}\right\}, \tag{6.2}$$

where

$$\begin{aligned}\Lambda(\gamma) &= \int_0^\infty \frac{1}{2\pi i}\int_{Br}\exp\left\{\left(\frac{\mu}{2D}+2^{-\frac{2}{3}}D^{-\frac{1}{3}}\lambda\right)\gamma\right\}\frac{\mathrm{Ai}\left(\lambda+2^{-\frac{1}{3}}D^{-\frac{2}{3}}\mu\right)}{\left[\mathrm{Ai}\,(\lambda)\right]^2}d\lambda d\mu \\ &= 2^{\frac{1}{3}}D^{\frac{2}{3}}\frac{1}{2\pi i}\int_{Br}\int_\lambda^{\infty+i\mathrm{Im}(\lambda)}\exp\left\{2^{-\frac{2}{3}}D^{-\frac{1}{3}}\gamma\rho\right\}\frac{\mathrm{Ai}\,(\rho)}{\left[\mathrm{Ai}\,(\lambda)\right]^2}d\rho d\lambda. \end{aligned} \tag{6.3}$$

Taking the derivative of Λ and using [2] $\mathrm{Ai}''(\rho) = \rho\mathrm{Ai}\,(\rho)$ yields

$$\begin{aligned}\Lambda'(\gamma) &= 2^{\frac{1}{3}}D^{\frac{2}{3}}\frac{1}{2\pi i}\int_{Br}\int_\lambda^{\infty+i\mathrm{Im}(\lambda)}2^{-\frac{2}{3}}D^{-\frac{1}{3}}\rho\exp\left\{2^{-\frac{2}{3}}D^{-\frac{1}{3}}\gamma\rho\right\}\frac{\mathrm{Ai}\,(\rho)}{\left[\mathrm{Ai}\,(\lambda)\right]^2}d\rho d\lambda \\ &= 2^{-\frac{1}{3}}D^{\frac{1}{3}}\frac{1}{2\pi i}\int_{Br}\int_\lambda^{\infty+i\mathrm{Im}(\lambda)}\exp\left\{2^{-\frac{2}{3}}D^{-\frac{1}{3}}\gamma\rho\right\}\frac{\mathrm{Ai}''(\rho)}{\left[\mathrm{Ai}\,(\lambda)\right]^2}d\rho d\lambda. \end{aligned} \tag{6.4}$$

Two integrations by parts give

$$\begin{aligned}&\int_\lambda^{\infty+i\mathrm{Im}(\lambda)}\exp\left\{2^{-\frac{2}{3}}D^{-\frac{1}{3}}\gamma\rho\right\}\mathrm{Ai}''(\rho)\,d\rho \\ &= \left[\mathrm{Ai}\,(\lambda)\right]^2\frac{d}{d\lambda}\left[\exp\left\{2^{-\frac{2}{3}}D^{-\frac{1}{3}}\gamma\lambda\right\}\frac{1}{\mathrm{Ai}\,(\lambda)}\right] \\ &\quad+\left(2^{-\frac{2}{3}}D^{-\frac{1}{3}}\gamma\right)^2\int_\lambda^{\infty+i\mathrm{Im}(\lambda)}\exp\left\{2^{-\frac{2}{3}}D^{-\frac{1}{3}}\gamma\rho\right\}\mathrm{Ai}\,(\rho)\,d\rho,\end{aligned}$$

which, when used in (6.4), leads to the differential equation

$$\Lambda'(\gamma) = \frac{1}{4D}\gamma^2\Lambda(\gamma). \tag{6.5}$$

Solving (6.5) yields

$$\Lambda(\gamma) = \Lambda_0\exp\left\{\frac{\gamma^3}{12D}\right\}, \tag{6.6}$$

where Λ_0 is a constant. To determine Λ_0, we let $\gamma \to -\infty$ in (6.3). Expanding the double integral by a combination of the Laplace and saddle point methods leads to

$$\Lambda(\gamma) \sim D^{\frac{2}{3}}2^{\frac{1}{3}}\exp\left\{\frac{\gamma^3}{12D}\right\}, \quad \gamma \to -\infty. \tag{6.7}$$

Comparing this to (6.6) we obtain $\Lambda_0 = D^{\frac{2}{3}}2^{\frac{1}{3}}$, which verifies (6.2).

6.3. The marginal distribution $M(x)$. To evaluate (1.5) by Laplace's method, we find where Ψ and Φ are maximal as functions of η. We thus examine the equations $\Psi_\eta = 0$ and $\Phi_\eta = 0$. We recall from (3.22) that $\Phi_\eta = (a-b)e^\tau - a$. The equation $\Phi_\eta = 0$ then reads

$$e^\tau = \frac{\sqrt{D}(\sigma-1)}{\sqrt{D}(\sigma-1) + D^{\frac{3}{2}}\sigma + D\sqrt{D\sigma^2 + (\sigma-1)^2}} < 1 \quad \text{for all} \quad D>0,\ \sigma>1.$$

We conclude that there is no solution to $\Phi_\eta = 0$ for $\tau > 0$, and hence $\Phi_\eta < 0$ in Region II.

From (3.4), $\Psi_\eta = (A-B)e^t - A$, and consequently

$$\Psi_\eta = 0 \quad \Leftrightarrow \quad t = \ln\left[\frac{1-s}{1-(D+1)s}\right], \tag{6.8}$$

which, when used in (3.6), yields

$$\begin{aligned} &\Psi_\eta = 0 \quad \Leftrightarrow \quad x = X_1(\eta), \quad 0 \le \eta < \frac{1}{D+1}, \\ &X_1(\eta) = -2\eta - \frac{1}{D}\left(2D\eta - D + 2\eta - 2\right)\ln\left[\frac{1-\eta}{1-(D+1)\eta}\right]. \end{aligned} \tag{6.9}$$

The equation $x = X_1(\eta)$ implicitly defines η as a function of x, $\eta = E(x)$. We introduce the function $\Psi_1(x) \equiv \Psi\left[x, E(x)\right]$ and from (3.6) we get

$$\Psi_1(x) = \frac{E(x)\left[1-E(x)\right]}{D} + \frac{D+1}{D^2}\left[1-E(x)\right]^2 \ln\left[\frac{1-(D+1)E(x)}{1-E(x)}\right]. \tag{6.10}$$

From the defining equation

$$-2E(x) + \frac{1}{D}\left[2(D+1)E(x) - D - 2\right]\ln\left[\frac{1-(D+1)E(x)}{1-E(x)}\right] = x, \tag{6.11}$$

we obtain the asymptotic results

$$\begin{aligned} &E(x) \sim \frac{x}{D} - \frac{1}{2}\frac{x^2}{D} + \frac{1}{6}\frac{D-4}{D^2}x^3, \quad x \to 0, \\ &E(x) \sim \frac{1}{D+1} - \frac{D}{(D+1)^2}\exp\left(-x - \frac{2}{D+1}\right), \quad x \to \infty. \end{aligned} \tag{6.12}$$

Use of Laplace's method to evaluate the integral in (1.5) as $\varepsilon \to 0$ yields

$$M(x) \sim \varepsilon^{-\frac{3}{2}} K\left[x, E(x)\right]\sqrt{2\pi}\frac{1}{\sqrt{-\varepsilon^{-1}\Psi_{\eta\eta}\left[x, E(x)\right]}}\exp\left\{\frac{1}{\varepsilon}\Psi_1(x)\right\}, \tag{6.13}$$

and from (3.6), after some algebra, we have

$$\begin{aligned} M(x) &\sim \varepsilon^{-1}\frac{\left[1-E(x)\right]^2}{\sqrt{\Delta}}\exp\left\{\frac{1}{\varepsilon}\Psi_1(x)\right\}, \\ \Delta &= \frac{2\left[1-(D+1)E(x)\right]\left[1-E(x)\right]\left[x+2E(x)\right](D+1)D}{2(D+1)E(x) - D - 2} \\ &\quad + D\left[D + 2E(x) - 2(D+1)E(x)^2\right]. \end{aligned}$$

We can get more explicit results if x is either small or large, using (6.12). We obtain

$$M(x) \sim \varepsilon^{-1} \frac{1}{D}\left(1-\frac{x}{D}\right)\exp\left\{\frac{1}{\varepsilon}\left(-\frac{x}{D}+\frac{x^2}{2D^2}\right)\right\}, \quad x \to 0,$$

$$M(x) \sim \varepsilon^{-1}\left[\frac{D}{(1+D)^2}+\frac{2D+1}{D(1+D)^2}\exp\left(-x-\frac{2}{D+1}\right)\right] \tag{6.14}$$

$$\times \exp\left[-\frac{1}{\varepsilon}\left(\frac{x}{1+D}+\frac{1}{(1+D)^2}\right)\right], \quad x \to \infty.$$

The first result in (6.14) shows that $M(x)$ is concentrated in the range $x = O(\varepsilon)$, and the second result is consistent with the spectral solution to (1.3) obtained in [20].

7. Discussion and numerical studies. As discussed in the introduction, a numerical method for solving (1.1) (or (1.3)) appears in [20]. However, it becomes very difficult to apply when c is large. Furthermore, the present asymptotic analysis yields a great deal of information about the model, as described below.

We recall that x represents a (scaled) queue length or buffer level, while η is a measure of the input to the queue above some mean value. As long as $\eta < 1$, the server has sufficient time to process the input, and no queue develops.

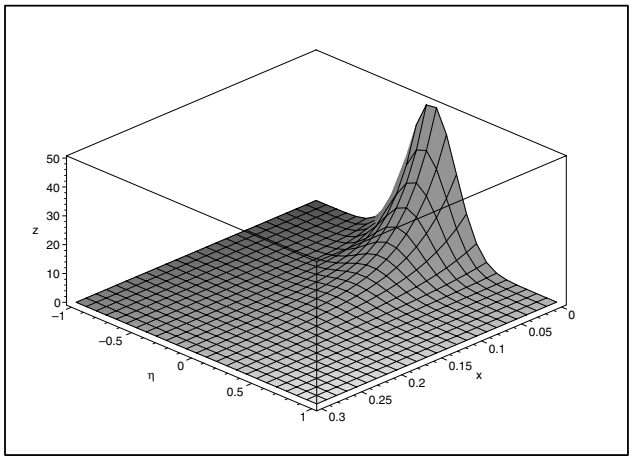
Our analysis first showed that if $|\eta|$ is small (i.e., the input flow is near its average value) and the buffer is also scaled to be small, the buffer and input processes decouple, and the distribution becomes Gaussian in η and exponential in x. But if η and/or x are not small, we obtain very different behaviors.

For example, given a fixed $\eta < 1$, we showed that the conditional buffer content is exponentially distributed, with mean $D/(1-\eta)$ (cf. (2.6)). This shows that if the input is below the maximum processing capacity, the buffer will tend to be small. But if $\eta > 1$, our analysis shows that the buffer content is approximately $x = X_0(\eta) = \eta - \ln(\eta) - 1$, with a Gaussian spread about this value. The transition from small buffers with exponential spread to large buffers with Gaussian spread occurs for $(x, \eta) \approx (0, 1)$ (more precisely, $x = O(\varepsilon^{2/3})$, $\eta - 1 = O(\varepsilon^{1/3})$). Thus, given a fixed $\eta \approx 1$ we obtain intermediate sized buffers, with the conditional buffer distribution now involving a contour integral of the Airy function (cf. (3.44)).

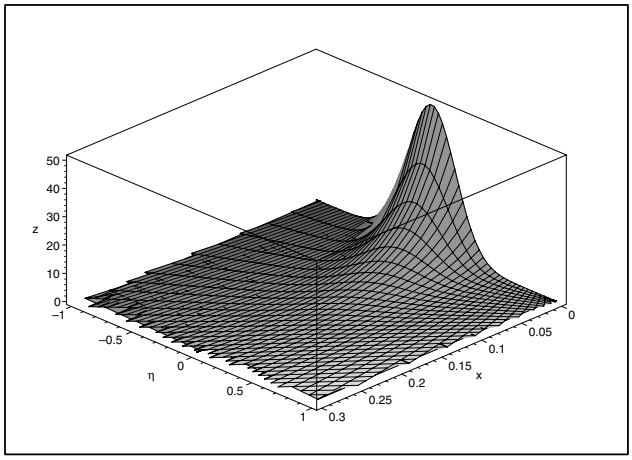
Our results also show that given a buffer level $x > 0$, the level of the input process is most likely to be $\eta = E(x)$, with $E(x)$ defined in (6.11). We also obtained the marginal distribution of the buffer level asymptotically, and our approximation is valid both for x small (where most mass concentrates) as well as $x = O(1)$ (the tail). Getting an accurate approximation to the latter is important in applications, since this can be used to estimate loss rates.

Next, we discuss the numerical accuracy of our asymptotic approximations. We use the numerical method in [20] to calculate $F(x, \eta)$ in (1.3) as a spectral expansion. We take $D = 1$ and $c = 5$, so that $\varepsilon = .04$. In Figure 7.1 we plot the "surface" $F(x, \eta)$ obtained by the numerical method and compare this to our approximation in (3.18). We take the range $-1 < \eta < 1$ and $0 < x < 0.3$. The figure shows that the surfaces are quite similar, and that our analysis correctly predicts the "ridge" at $\eta = E(x)$, along which F is maximal in x, for fixed η.

In Figure 7.2 we plot the numerical and asymptotic values of $F(0, \eta)$ for $-0.3 < \eta < 0.2$. This corresponds to the distribution of the input level if we are given an empty buffer. The asymptotic and numerical results lead to nearly indistinguishable curves. Also, our analysis correctly predicts that the maximum of $F(0, \eta)$ occurs at a value $\eta < 0$, and correctly predicts the skewness, or deviation from Gaussian



(a) Numerical



(b) Asymptotic

FIG. 7.1. *A sketch of the "surface" $F(x, \eta)$.*

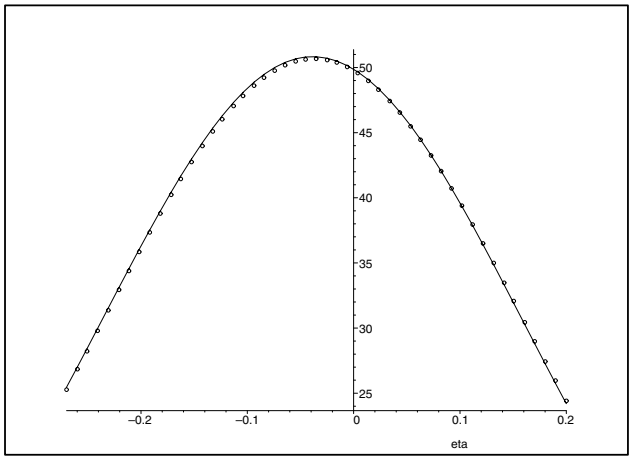


FIG. 7.2. *A comparison of the numerical (solid curve) and asymptotic* (ooo) *values of $F(0, \eta)$.*

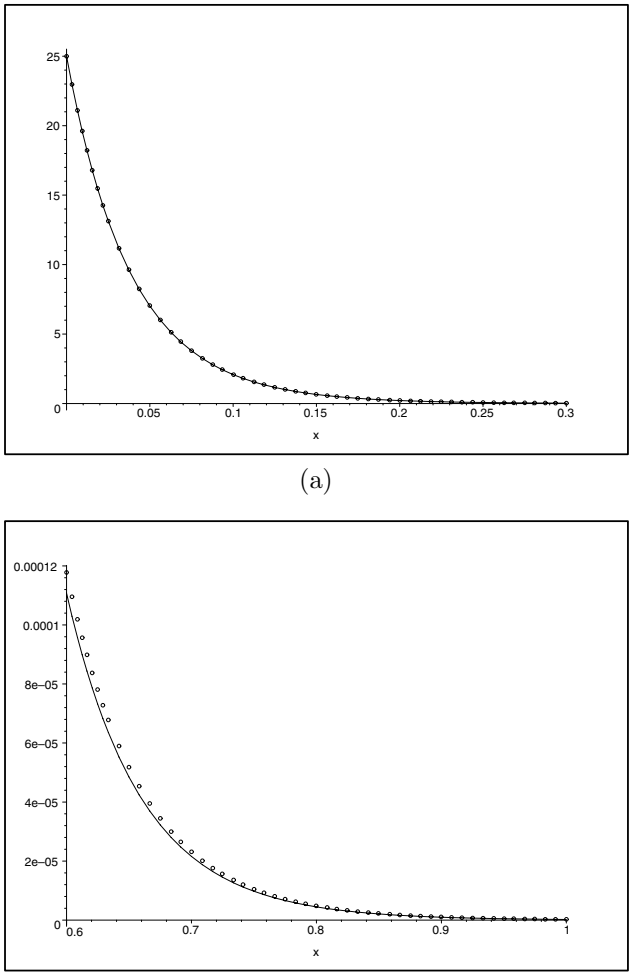


FIG. 7.3. *A comparison of the numerical (solid curve) and asymptotic* (ooo) *values of* $M(x)$.

behavior. Indeed, (2.6) shows that the maximum should be at $\eta \sim -\varepsilon = -.04$, while the numerics yields $-.036782659$.

In Figure 7.3(a) we plot the exact and numerical marginal buffer distributions in the range $x \in [0, 0.3]$. Note that our approximation is $M(x)$ in (6.14), that $M(0) = c^2 = \varepsilon^{-1}$ exactly, and that $M(x)$ is quite concentrated near $x = 0$. Again we see that the asymptotic and numerical curves nearly coincide. Finally, we plot in Figure 7.3(b) asymptotic and numerical values of $M(x)$ for x in the range $[0.6, 1]$. This shows that the asymptotic formula also accurately estimates the tail of the distribution.

Thus, we have shown that the asymptotic analysis yields both highly accurate approximations and qualitative insights about the joint, marginal, and conditional distributions of the input level and buffer processes.

Acknowledgments. D. Dominici wishes to thank Professor Floyd Hanson for his generous sponsorship. We also want to thank the anonymous referees for their valuable suggestions and comments.

REFERENCES

[1] D. Anick, D. Mitra, and M. M. Sondhi, *Stochastic theory of a data-handling system with multiple sources*, Bell System Tech. J., 61 (1982), pp. 1871–1894.

[2] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, 9th ed., Dover, New York, 1972.

[3] A. K. Aziz, D. A. French, S. Jensen, and R. B. Kellogg, *Origins, analysis, numerical analysis, and numerical approximation of a forward-backward parabolic problem*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 895–922.

[4] A. K. Aziz and J.-L. Liu, *A Galerkin method for the forward-backward heat equation*, Math. Comp., 56 (1991), pp. 35–44.

[5] M. S. Baouendi and P. Grisvard, *Sur une équation d'évolution changeant de type*, J. Funct. Anal., 2 (1968), pp. 352–367.

[6] C. Bardos, R. E. Caflisch, and B. Nicolaenko, *The Milne and Kramers problems for the Boltzmann equation of a hard sphere gas*, Comm. Pure Appl. Math., 39 (1986), pp. 323–352.

[7] C. M. Bender and S. A. Orzag, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, New York, 1978.

[8] J. N. Daigle and J. D. Langford, *Models for analysis of packet voice communication systems*, IEEE J. Selected Areas Commun., 4 (1986), pp. 847–855.

[9] D. Dominici and C. Knessl, *A small elliptic perturbation of a backward-forward parabolic problem with applications to stochastic models*, Appl. Math. Lett., 17 (2004), pp. 535–542.

[10] N. G. Duffield, *Conditioned asymptotics for tail probabilities in large multiplexers*, Perform. Eval., 31 (1998), pp. 281–300.

[11] M. Freidlin and H. Weinberger, *On a backward-forward parabolic equation and its regularization*, J. Differential Equations, 105 (1993), pp. 264–295.

[12] M. Gevrey, *Œuvres de Maurice Gevrey*, Éditions du Centre National de la Recherche Scientifique, Paris, 1970.

[13] P. S. Hagan, C. R. Doering, and C. D. Levermore, *Mean exit times for particles driven by weakly colored noise*, SIAM J. Appl. Math., 49 (1989), pp. 1480–1513.

[14] P. S. Hagan and J. R. Ockendon, *Half-range analysis of a counter-current separator*, J. Math. Anal. Appl., 160 (1991), pp. 358–378.

[15] O. Hashida and M. Fujika, *Queueing models for buffer memory in store-and-forward systems*, in Proceedings of the Seventh International Teletraffic Congress, Stockholm, Sweden, 1973, pp. 323/1–323/7.

[16] E. Hopf, *Mathematical Problems of Radiative Equilibrium*, Cambridge Tracts in Mathematics and Mathematical Physics 31, Stechert-Hafner, New York, 1964.

[17] D. Jung, M. Lücke, and A. Szprynger, *Influence of inlet and bulk noise on Rayleigh-Bénard convection with lateral flow*, Phys. Rev. E (3), 63 (2001), pp. 1–20.

[18] J. B. Keller, *Rays, waves and asymptotics*, Bull. Amer. Math. Soc., 84 (1978), pp. 727–750.

[19] J. B. Keller and H. F. Weinberger, *Boundary and initial-boundary value problems for separable backward-forward parabolic problems*, J. Math. Phys., 38 (1997), pp. 4343–4353.

[20] C. Knessl and C. Tier, *Heavy traffic analysis of a Markov-modulated queue with finite capacity and general service times*, SIAM J. Appl. Math., 58 (1998), pp. 257–323.

[21] C. Knessl and J. A. Morrison, *Heavy-traffic analysis of a data-handling system with many sources*, SIAM J. Appl. Math., 51 (1991), pp. 187–213.

[22] C. Knessl, *Asymptotic analysis of a backward-forward parabolic problem for data-handling systems*, SIAM J. Appl. Math., 61 (2000), pp. 914–933.

[23] C. Knessl and J. B. Keller, *Ray solution of a backward-forward parabolic problem for data handling systems*, European J. Appl. Math., 11 (2000), pp. 1–12.

[24] C. Knessl, *Exact and asymptotic solutions to a PDE that arises in time-dependent queues*, Adv. in Appl. Probab., 32 (2000), pp. 256–283.

[25] Q. Ren and H. Kobayashi, *A mathematical theory for transient analysis of communications networks*, IEICE Trans. Commun., E75-B (1992), pp. 1266–1276.

[26] Q. Ren and H. Kobayashi, *Transient solutions for the buffer behavior in statistical multiplexing*, Perform. Eval., 23 (1995), pp. 65–87.

[27] S. Q. Li, *Study of information loss in packet switched voice systems*, IEEE Trans. Commun., 37 (1989), pp. 1192–1202.

[28] M. Mandjes, *Overflow asymptotics for large communications systems with general Markov fluid sources*, Probab. Engrg. Inform. Sci., 10 (1996), pp. 501–518.

[29] B. A. Sevast′yanov, *Influence of storage bin capacity on the average standstill time of a production line*, Teor. Veroyatnost. i Primenen., 7 (1962), pp. 438–446 (in Russian); Theory Probab. Appl., 7 (1962), pp. 429–438 (in English).

[30] H. O. Tanaka, T. and Y. Takahashi, *Transient analysis of fluid model for ATM statistical multiplexer*, Perform. Eval., 23 (1995), pp. 145–162.
[31] R. C. F. Tucker, *Accurate method for analysis of a packet-speech multiplexer with limited delay*, IEEE Trans. Commun., 36 (1988), pp. 479–483.
[32] J. Wijngaard, *The effect of interstage buffer storage on the output of two unreliable production units in series with different production rates*, AIIE Trans., 11 (1979), pp. 42–47.

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 1895–1916

SUPPRESSION OF THE DIRICHLET EIGENVALUES OF A COATED BODY*

STEVE ROSENCRANS† AND XUEFENG WANG†

Abstract. We consider the problem of protecting from overheating the interiors of anisotropically heat-conducting bodies whose boundaries are maintained at a high temperature. The bodies are composites consisting of a thin anisotropic insulating coating surrounding an isotropically conducting interior (e.g., a space shuttle painted with an insulator). This anisotropy is a common feature of the *nanocomposite* materials used as insulators. Denote by A the thermal tensor (matrix) of the coated body and consider the Dirichlet eigenvalues of the elliptic operator $u \mapsto -\nabla \cdot (A\nabla u)$ on the coated body. The eigenfuction expansion of the interior temperature shows that small eigenvalues favor insulation of the interior. This is the motivation for studying the idealized mathematical problem of suppression of the Dirichlet eigenvalues. Suppose A is a constant matrix $\overline{A}$ on the coating. The focus of this paper is estimation of the elliptic eigenvalues and qualitative description of the eigenfunctions using *only* the eigenvalues of $\overline{A}$, the scalar conductivity of the uncoated body, and certain scalar characteristics of the geometry of the uncoated body. We study the effect of small matrix eigenvalues, small thickness of the coating, and their interplay. If the thermal tensor of the coating is spatially varying and optimally configured so that the minimum eigenvalue has eigenvector normal to the body at all boundary points of the body (and remains equal to that normal vector at each point in the coating on the straight line in that normal direction), only that minimum eigenvalue need be small. A by-product is a new characterization of the first positive Neumann eigenvalue in terms of a sequence of *second* Dirichlet eigenvalues.

Key words. nanocomposite, Dirichlet eigenvalue, anisotropic heat conduction, thermal tensor, thermal management, insulation, reinforcement

AMS subject classifications. 35J05, 35J20, 80A20, 80M30, 80M40

DOI. 10.1137/040621181

1. Introduction: Suppression of Dirichlet eigenvalues.

In this paper we consider the problem of protecting interior subregions of anisotropically heat-conducting bodies whose boundaries are maintained at a high temperature. If $A = (a_{ij})$ is the thermal tensor of the conducting medium (i.e., the matrix of thermal diffusion coefficients, assumed symmetric, positive-definite), then heat flow is *isotropic* (the same in all directions) if and only if A is invariant under all coordinate rotations, which is equivalent to $A = kI$ for a scalar coefficient k (I is the identity matrix). Every thermal tensor not of this form is associated with *anisotropic* heat conduction. Currently, good insulators can be created by proper designing of nanocomposite materials,[1] which, at the macroscale, commonly exhibit such anisotropy. When isotropic nanoscale (for example, periodic) structure is averaged out by the process of homogenization (see [2] for the periodic case), the resulting effective diffusion is generally anisotropic. See the paper [6] of Zheng et al. for a calculation of the thermal tensor explicitly showing its dependence on nanostructural parameters (volume fraction of inclusion, orientation, and the like).

*Received by the editors December 20, 2004; accepted for publication (in revised form) February 21, 2006; published electronically August 29, 2006.
http://www.siam.org/journals/siap/66-6/62118.html
†Mathematics Department, Tulane University, New Orleans, LA 70118 (sir@math.tulane.edu, xdw@math.tulane.edu).
[1]For example, nanoclay platelets are added to traditional polymers to enhance thermal, mechanical, or gas permeability properties.

We consider composites consisting of a (thin) nanocoating surrounding an isotropically conducting interior requiring protection (e.g., a space shuttle painted with a nanoinsulator).

The temperature $T(t, \mathbf{x})$ in the body Ω satisfies the heat equation

$$\begin{cases} T_t = \nabla \cdot (A \nabla T), & \mathbf{x} \in \Omega,\ t > 0, \\ T = T_0(\mathbf{x}), & \mathbf{x} \in \Omega,\ t = 0, \\ T = H, & \mathbf{x} \in \partial\Omega,\ t > 0, \end{cases} \tag{1}$$

where T_0 is the initial temperature distribution and H is some positive constant (temperature) that is large compared to the values of T_0. The evolution of the temperature T is obtained from the its expansion in Dirichlet eigenfunctions of the elliptic operator $u \to -\nabla \cdot (A \nabla u)$,

$$\begin{cases} \nabla \cdot (A \nabla u) + \lambda u = 0 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{2}$$

Let λ_1 be the principal (i.e., smallest) Dirichlet eigenvalue and ϕ_1 the positive (in the interior) normalized ($\int \phi_1^2 = 1$) principal eigenfunction. Let (λ_m, ϕ_m), $m = 2, 3, \dots$ be the higher eigenvalues and normalized eigenfunctions, $\lambda_1 < \lambda_2 \le \lambda_3 \le \cdots$, where the eigenvalues are repeated according to their geometric multiplicity. Then

$$T - H = \sum_{m \ge 1} e^{-\lambda_m t} \phi_n(\mathbf{x}) \int_\Omega \phi_n(\mathbf{x}') \left(T_0(\mathbf{x}') - H\right) \, d\mathbf{x}'. \tag{3}$$

Thus $T(\mathbf{x}, t)$ eventually converges to H as $t \to \infty$. To protect the interior from overheating, we want to slow down this process; i.e., we desire that for a long period of time T not diverge too far from its initial value T_0. To achieve this, the following considerations are important:

(i) Even though $\lambda_i \to \infty$ as $i \to \infty$, we can make $T(\mathbf{x}, t)$ remain close to $T_0(\mathbf{x})$ for a long period of time by making λ_i as small as possible for as many i as possible.

(ii) Thus the first thing to do is to make the principal eigenvalue λ_1 small. Indeed, for t large enough ($t \gg 1/\lambda_2$), $T - H$ is well approximated by the first term,

$$T - H \approx e^{-\lambda_1 t} \phi_1(\mathbf{x}) \int_\Omega \phi_1(\mathbf{x}') \left(T_0(\mathbf{x}') - H\right) \, d\mathbf{x}'; \tag{4}$$

therefore small values of λ_1 favor insulation of the interior.

(iii) The shape of the principal eigenfunction $\phi_1(\mathbf{x})$ indicates what part of Ω is well protected. Large values of ϕ_1 in some subdomain can mitigate the overheating, by compensating to some extent for insufficiently small λ_1, because the small exponential $e^{-\lambda_1 t}$ is multiplied by ϕ_1. Small values of ϕ_1 tend to encourage overheating.

(iv) The shapes of the higher eigenmodes $\phi_i(\mathbf{x})$ help us to understand the evolution more completely: For example, if many higher eigenmodes are small in a subdomain Ω' of Ω, then (4) is valid in Ω' for t not necessarily large.

We emphasize that having small λ_i's is not enough to prevent overheating. This is elucidated by an example in section 2.2.

In most of this paper we suppose that the thermal tensor A is a constant matrix $\overline{A}$ on the coating. One certainly expects the coating to be a better insulator if all

eigenvalues of the *matrix* $\overline{A}$ are small. This is confirmed and quantified by estimation of the elliptic eigenvalues and qualitative description of the eigenfunctions using *only* the eigenvalues of the thermal tensor (matrix) $\overline{A}$ of the coating, the scalar conductivity of the coated body, and certain scalar characteristics of the geometry of the coated body. We study in detail the effect of small matrix eigenvalues, small thickness of the coating, and their interplay.

If the thermal tensor of the coating varies spatially, i.e., $\overline{A} = \overline{A}(\mathbf{x})$, its eigenvalues and their eigenvectors vary from point to point in the coating. The optimum configuration of coating relative to body would be that in which the smallest eigenvalue has an eigenvector normal to the body boundary. In this case we prove some of the same results as for a constant thermal tensor, assuming only that the smallest eigenvalue is small. The magnitude of the other eigenvalues is not significant. See section 3.3.

For discussions of the issue of *how* to design the nanostructure (of the coating material) so that its thermal tensor has small eigenvalues, see [5].

If an elastic body is reinforced by a hard coating, or shell, the very same spectral problem (2) occurs; the eigenvalues are related to vibration frequencies, and the purpose of the coating is the reduction of the frequencies. After completion of our work we became aware of Friedman's interesting 1980 paper [3] in this context. His concern is suppression of the principal eigenvalue. In section 1.1 we compare our results with his. An overall technical difference is that Friedman's proofs need estimates on the second derivatives of eigenfunctions; we use only estimates of the H^1 norm, which are easily obtained via variational characterization of the eigenvalues. We note that we measure thickness of the coating in the normal direction from the body whereas [3] measures thickness using the *conormal* direction determined by (what is here called) the thermal tensor of the coating.

1.1. Summary of our results. We consider a region Ω_1 composed of some isotropically conducting material protected by an external insulator coating Ω_2 with thermal tensor $(\overline{a}_{ij})$. See Figure 1.

Let $A(\mathbf{x}) = (a_{ij}(\mathbf{x}))$ (the thermal tensor of the *composite* material, i.e., the coated body) be a 3×3 or 2×2 thermal tensor, $\mathbf{x} \in \overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$:

$$A(\mathbf{x}) = (a_{ij}(\mathbf{x})) = \begin{cases} kI_{n\times n}, & \mathbf{x} \in \Omega_1, \\ (\overline{a}_{ij})_{n\times n}, & \mathbf{x} \in \Omega_2, \end{cases} \tag{5}$$

where $k > 0$ is the constant thermal conductivity of the body occupying Ω_1, $(\overline{a}_{ij})$ is a constant positive-definite matrix ($n = 2$ or $n = 3$), and I is the identity matrix. Denote by $\overline{\sigma}_1, \overline{\sigma}_2, \ldots, \overline{\sigma}_n$ the eigenvalues of the thermal tensor of the coating. Let the largest and smallest be denoted by $\overline{\sigma}_{\max}$ and $\overline{\sigma}_{\min}$. After an appropriate rotation of coordinates (in two dimensions (2D) or three dimensions (3D)) the tensor $(\overline{a}_{ij})$ and hence also $(a_{ij}(\mathbf{x}))$ are diagonalized. From now on we work in this principal coordinate system.

Here follows a sketch, not complete, of our results.

Coating fixed, not thin. (In [3] this is called "thick reinforcement.") Suppose the thermal tensor of the coating depends on a small parameter ϵ in such a way that for some constants $\sigma_1, \ldots, \sigma_n$, $\overline{\sigma}_i \sim \epsilon\sigma_i$ as $\epsilon \downarrow 0$. Then there exist $\alpha_m(\Omega)$, $m \geq 1$, such that the Dirichlet eigenvalues $\lambda_m \sim \epsilon\alpha_m(\Omega)$ as $\epsilon \downarrow 0$, $m \geq 1$. The principal eigenfunction is close to a plateau over the body. The $\{\alpha_m(\Omega)\}$ are characterized as eigenvalues of a Rayleigh quotient in a certain Hilbert space. See Theorems 1 and 2. *Thus the evolution varies uniformly according to a slow time ϵt.* The paper [3] has the result for $m = 1$ but with a different characterization of α_1.

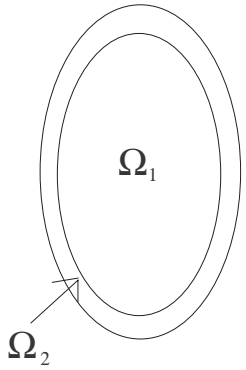


FIG. 1. *The* completely *coated body is the domain* $\Omega = \Omega_2 \cup \overline{\Omega}_1$*, the coating is the open set* Ω_2*, the uncoated body* Ω_1 *is a proper subdomain of the coated body, and the distance between* $\partial\Omega_1$ *and* $\partial\Omega$ *is positive. The uncoated body may have holes, also protected. Thus the coating* Ω_2 *may be disconnected.*

Section 2.2 has some comments on incompletely coated bodies.

All $\overline{\sigma}_i = \mathrm{o}(\delta)$. Suppose the thickness of the coating is δ, that as $\delta \downarrow 0$ all eigenvalues of the thermal tensor of the coating are $\mathrm{o}(\delta)$, and that $\overline{\sigma}_{\max}/\overline{\sigma}_{\min}$ is bounded. Then $\lambda_1 = \mathrm{o}(1)$ as $\delta \downarrow 0$: For small enough δ,

$$\frac{|\partial\Omega_1|}{|\Omega_1|}\frac{\overline{\sigma}_{\min}}{\delta}(1+\mathrm{o}(1)) \le \lambda_1 \le \frac{|\partial\Omega_1|}{|\Omega_1|}\frac{\overline{\sigma}_{\max}}{\delta}(1+\mathrm{o}(1)).$$

Moreover, the normalized positive principal eigenfunction ϕ_1 converges to the constant $1/\sqrt{|\Omega|}$ strongly in $L^2(\Omega_1)$ as $\delta \downarrow 0$. See Theorems 3 and 4. The paper [3] has this result in terms of the "conormal" thickness δ' rather than δ, with a more complicated proof.[2] Our Theorem 3 actually implies the upper bound *without* assuming $\overline{\sigma}_{\max} = \mathrm{o}(\delta)$. Theorems 6 and 7 generalize the above results to the optimally aligned case.

All $\overline{\sigma}_i = \mathrm{o}(\delta^2)$. Assume $\overline{\sigma}_{\max}/\overline{\sigma}_{\min}$ is bounded as $\delta \downarrow 0$. If the $\overline{\sigma}_i$ are *very small*, i.e., all $\overline{\sigma}_i = \mathrm{o}(\delta^2)$, then the higher eigenvalues are all $\mathrm{O}(\overline{\sigma}_{\max}/\delta^2) = \mathrm{o}(1)$ while $\lambda_1 = \mathrm{O}(\overline{\sigma}_{\max}/\delta) = \mathrm{o}(\delta)$ as $\delta \downarrow 0$. The normalized higher eigenfunctions converge to zero strongly in $L^2(\Omega_1)$. See Theorem 9. In this case, changes occur very slowly until eventually (4) holds and thereafter the changes are even slower. Theorem 11 generalizes this to the optimally aligned case.

All $\overline{\sigma}_i = \mathrm{o}(\delta)$ *while all* $\overline{\sigma}_i/\delta^2 \to \infty$. If, on the other hand, all the $\overline{\sigma}_i$ are small but not *very small*, i.e., $\overline{\sigma}_i = \mathrm{o}(\delta)$ while all $\overline{\sigma}_i/\delta^2 \to \infty$ as $\delta \downarrow 0$, then $\lambda_1 = \mathrm{o}(1)$ while $\lambda_2 \to k\mu_2$ as $\delta \downarrow 0$, where μ_2 is the first positive Neumann eigenvalue of $-\Delta$ on Ω_1. *So we have a new interpretation of the first positive Neumann eigenvalue in terms of a sequence of second Dirichlet eigenvalues.* See Theorem 8. In this case (4) is accurate after an $\mathrm{O}(1)$ time, after which changes are slow. Theorem 10 generalizes this to the optimally aligned case.

Thermal tensor of coating scaled by $\beta\delta'$ for some constant β, and conormal thickness $\delta' \downarrow 0$. It is shown in [3] that λ_1 converges to a certain (conormal) Robin eigenvalue. We did not treat this case.

[2]The paper [3] allows more general elliptic operators, including first-order terms, in the coating and the body. If the body operator has no such terms and if the coating operator is multiplied by a scalar $\mu = \mathrm{o}(\delta')$, then it is shown in [3] that $\lambda_1\delta'/\mu \to |\partial\Omega_1|/|\Omega_1|$.

2. Fixed coating. In this section the coating Ω_2 is fixed. In section 2.1 we assume that the body is completely coated; see Figure 1. In section 2.2 we briefly consider how our results change for incompletely coated bodies.

Suppose that the anisotropic thermal tensor $(\overline{a}_{ij})$ depends on a small parameter ϵ, $0 < \epsilon < 1$. From now on write ϕ as ϕ^ϵ.

2.1. Completely coated bodies.

THEOREM 1. *Assume the body Ω_1 is completely coated, as explained in the caption to Figure 1, and assume the eigenvalues $\overline{\sigma}_1, \overline{\sigma}_2, \dots, \overline{\sigma}_n$ of the thermal tensor of the coating satisfy*

$$\overline{\sigma}_i = \epsilon\sigma_i + o(\epsilon) \quad \text{as } \epsilon \downarrow 0, \quad i = 1, \dots, n, \tag{6}$$

for some positive $\sigma_1, \dots, \sigma_n$.

(i) *As $\epsilon \downarrow 0$*

$$\lambda_1(\Omega) = \epsilon\alpha_1(\Omega) + o(\epsilon),$$

and in the $L^2(\Omega)$-sense the normalized positive principal eigenfunction satisfies

$$\phi_1^\epsilon = \phi_1^0 + o(1),$$

where

$$\begin{cases} 0 < \alpha_1(\Omega) = \inf\limits_{u \in H,\, u \not\equiv 0} \dfrac{\int_\Omega \sigma_1 u_{x_1}^2 + \cdots + \sigma_n u_{x_n}^2}{\int_\Omega u^2}, \\ H := \left\{ u \in H_0^1(\Omega) | u \text{ is constant on } \Omega_1 \right\} \end{cases} \tag{7}$$

and ϕ_1^0 is the unique minimizer of (7) with $\int_\Omega (\phi_1^0)^2 = 1$ and $\phi_1^0 > 0$ on Ω. (Thus the shape of ϕ_1^ϵ is close to a plateau over Ω_1.)

(ii) *$\alpha_1(\Omega)$ is less than the principal (smallest) eigenvalue $\widetilde{\lambda}(\Omega_2)$ of the Dirichlet problem*

$$\begin{cases} \sigma_1 u_{x_1 x_1} + \cdots + \sigma_n u_{x_n x_n} + \lambda u = 0, & \mathbf{x} \in \Omega_2, \\ u = 0, & \mathbf{x} \in \partial\Omega_2. \end{cases} \tag{8}$$

Remark 1. By rescaling k in (5) to be 1 we see that the small "o" terms are uniformly small over k if k is bounded away from zero.

Proof of Theorem 1. *Step* 1. The set H introduced in the statement of the theorem is a Hilbert subspace of $H_0^1(\Omega)$. By the Banach theorem and the Kondrachov compact imbedding, $H \hookrightarrow L^2(\Omega)$, $\alpha_1(\Omega)$ is achieved by a minimizer $\phi_1^0 \in H$ with $\phi_1^0 \geq 0$ and $\int_\Omega \left(\phi_1^0\right)^2 = 1$. In fact, any minimizer must be of one sign. The Euler–Lagrange equation is

$$\int_\Omega \sigma_1 \left(\phi_1^0\right)_{x_1} v_{x_1} + \cdots + \sigma_n \left(\phi_1^0\right)_{x_n} v_{x_n} - \alpha_1(\Omega)\phi_1^0 v = 0 \quad \text{for all } v \in H. \tag{9}$$

In particular (by taking $v \in H_0^1(\Omega_2) \subset H_0^1(\Omega)$), we have

$$\begin{cases} \sigma_1 \left(\phi_1^0\right)_{x_1 x_1} + \cdots + \sigma_n \left(\phi_1^0\right)_{x_n x_n} + \alpha_1(\Omega)\phi_1^0 = 0, & \mathbf{x} \in \Omega_2, \\ \phi_1^0 = 0, & \mathbf{x} \in \partial\Omega, \\ \phi_1^0 = \text{const} = C^0 \geq 0, & \mathbf{x} \in \partial\Omega_1. \end{cases} \tag{10}$$

We claim that $C^0 > 0$. Multiplying both sides of (10) by $v \in H$ we have

$$0 = \int_{\Omega_2} \sigma_1 \phi^0_{1x_1} v_{x_1} + \cdots + \sigma_n \phi^0_{1x_n} v_{x_n} - \alpha_1(\Omega)\phi^0 v - \int_{\partial\Omega_1} \left(\sigma_1 \phi^0_{1x_1}\nu_1 + \cdots + \sigma_n \phi^0_{1x_n}\nu_n\right) v, \tag{11}$$

where ν is the outer normal of Ω_2 on $\partial\Omega_1$. Now subtract (9) from (11):

$$\int_{\Omega_1} \alpha_1(\Omega)\phi^0_1 v - \int_{\partial\Omega_1} \left(\sigma_1 \phi^0_{1x_1}\nu_1 + \cdots + \sigma_n \phi^0_{1x_n}\nu_n\right) v = 0.$$

Since $v|_{\partial\Omega_1} = \text{const}$ for $v \in H$, we have

$$\int_{\Omega_1} \alpha_1(\Omega)\phi^0_1 = \int_{\partial\Omega_1} \left(\sigma_1 \phi^0_{1x_1}\nu_1 + \cdots + \sigma_n \phi^0_{1x_n}\nu_n\right). \tag{12}$$

If $\phi^0_1|_{\partial\Omega_1} = 0$, then by the Hopf boundary point lemma (recall $\phi^0_1 \geq 0$ in Ω) we have that the integrand on the right-hand side of (12) is negative. On the other hand, since $\phi^0_1|_{\Omega_1}$ is constant, $\phi^0_1|_{\Omega_1} = 0$ if $\phi^0_1|_{\partial\Omega_1} = 0$. We have reached a contradiction and thus $\phi^0_1|_{\partial\Omega_1} > 0$.

Step 2. We prove $\limsup_{\epsilon\downarrow 0} \lambda_1(\Omega)/\epsilon \leq \alpha_1(\Omega)$. By the variational characterization[3]

$$\begin{aligned}
\lambda_1(\Omega) &= \inf_{u\in H^1_0(\Omega), u\not\equiv 0} \frac{\int_\Omega a_{ij}(\mathbf{x}) u_{x_i} u_{x_j}}{\int_\Omega u^2} \\
&\leq \frac{\int_\Omega a_{ij}(\mathbf{x}) \left(\phi^0_1\right)_{x_i} \left(\phi^0_1\right)_{x_j}}{\int_\Omega \left(\phi^0_1\right)^2} \\
&= \epsilon \int_\Omega \left(\sigma_1 + \mathrm{o}(1)\right) \left(\phi^0_1\right)^2_{x_1} + \cdots + \left(\sigma_n + \mathrm{o}(1)\right) \left(\phi^0_1\right)^2_{x_n}.
\end{aligned}$$

(Note $\nabla\phi^0_1 = 0$ in Ω_1.) This completes Step 2.

Step 3. By Step 2 we have

$$\mathrm{O}(\epsilon) = k \int_\Omega |\nabla \phi^\epsilon_1|^2 + \epsilon \int_{\Omega_2} \left(\sigma_1 + \mathrm{o}(1)\right) \left(\phi^\epsilon_1\right)^2_{x_1} + \cdots + \left(\sigma_n + \mathrm{o}(1)\right) \left(\phi^\epsilon_1\right)^2_{x_n}. \tag{13}$$

Thus ϕ^ϵ_1 is bounded in $H^1_0(\Omega)$ as $\epsilon \downarrow 0$. So for each sequence $\epsilon \downarrow 0$ there exists a subsequence (still denoted by ϵ) such that $\phi^\epsilon_1 \to \widetilde{\phi}^0_1 \in H^1_0(\Omega)$ and strongly in $L^2(\Omega)$. In particular $\widetilde{\phi}^0_1 \geq 0$ and $\int_\Omega \left(\widetilde{\phi}^0_1\right)^2 = 1$. By (13) we easily see $\widetilde{\phi}^0_1$ is constant on Ω_1.

[3]Here and henceforth we use the summation convention.

Thus $\widetilde{\phi}_1^0 \in H$ and

$$\begin{aligned}\liminf_{\epsilon\downarrow 0}\frac{\lambda_1(\Omega)}{\epsilon} &\geq \liminf_{\epsilon\downarrow 0}\int_{\Omega_2}(\sigma_1+\mathrm{o}(1))\,(\phi_1^\epsilon)_{x_1}^2+\cdots+(\sigma_n+\mathrm{o}(1))\,(\phi_1^\epsilon)_{x_n}^2\\ &\geq \int_{\Omega_2}\sigma_1\left(\widetilde{\phi}_1^0\right)_{x_1}^2+\cdots+\sigma_n\left(\widetilde{\phi}_1^0\right)_{x_n}^2\\ &= \frac{\int_{\Omega_2}\sigma_1\left(\widetilde{\phi}_1^0\right)_{x_1}^2+\cdots+\sigma_n\left(\widetilde{\phi}_1^0\right)_{x_n}^2}{\int_\Omega\left(\widetilde{\phi}_1^0\right)^2}\\ &\geq \alpha_1(\Omega).\end{aligned}$$

This and Step 2 imply $\lim_{\epsilon\downarrow 0}\lambda_1(\Omega)/\epsilon=\alpha_1(\Omega)$. The above arguments also show that $\widetilde{\phi}_1^0$ is a minimizer of $\alpha_1(\Omega)$.

Step 4. We show that $\alpha_1(\Omega)<\widetilde{\lambda}(\Omega_2)$. Let ψ be an eigenfunction of (8) corresponding to $\widetilde{\lambda}(\Omega_2)$, with $\psi>0$ on Ω_2. That is, (8) holds with $u=\psi$ and $\lambda=\widetilde{\lambda}(\Omega_2)$. Multiplying it by ϕ_1^0 and integrating by parts, we have

$$\begin{aligned}-\int_{\Omega_2}\sigma_1\psi_{x_1}\phi_{1x_1}^0+\cdots+\sigma_n\psi_{x_n}\phi_{1x_n}^0+\int_{\partial\Omega_1}\phi_1^0\left(\sigma_1\psi_{x_1}\nu_1+\cdots+\sigma_n\psi_{x_n}\nu_n\right)\\ +\widetilde{\lambda}(\Omega_2)\int_{\Omega_2}\psi\phi_1^0=0.\end{aligned}\tag{14}$$

In (9) take $v=\psi$ and then add (9) to (14). We then have

$$\left(\widetilde{\lambda}(\Omega_2)-\alpha_1(\Omega)\right)\int_{\Omega_2}\psi\phi_1^0+\int_{\partial\Omega_1}\phi_1^0\left(\sigma_1\psi_{x_1}\nu_1+\cdots+\sigma_n\psi_{x_n}\nu_n\right)=0.$$

The Hopf boundary point lemma applied to ψ and the fact that $\phi_1^0|_{\partial\Omega_1}>0$ imply $\widetilde{\lambda}(\Omega_2)-\alpha_1(\Omega)>0$.

Step 5. We prove the uniqueness of the minimizer for $\alpha_1(\Omega)$. We just need to show $\phi_1^0=\widetilde{\phi}_1^0$. As in Step 1, we have $\widetilde{\phi}_1^0\big|_{\partial\Omega_1}>0$. Since both $\widetilde{\phi}_1^0\big|_{\partial\Omega_1}$ and $\phi_1^0\big|_{\partial\Omega_1}$ are constants, there is a constant μ such that $(\widetilde{\phi}_1^0+\mu\phi_1^0)\big|_{\partial\Omega_1}=0$. Define $\theta=\widetilde{\phi}_1^0+\mu\phi_1^0$. Then θ satisfies

$$\begin{cases}\sigma_1\theta_{x_1x_1}+\cdots+\sigma_n\theta_{x_nx_n}+\alpha_1(\Omega)\theta=0, & \mathbf{x}\in\Omega_2,\\ \theta=0, & \mathbf{x}\in\partial\Omega_2,\\ \theta=0, & \mathbf{x}\in\Omega_1,\\ \theta\in H.\end{cases}$$

If $\theta\not\equiv 0$, then $\alpha_1(\Omega)$ is an eigenvalue of (8). This contradicts Step 4. The uniqueness is proved.

(The subsequence of ϵ's that was found in Step 2 is actually not needed. By the uniqueness of the minimizer, $\phi^\epsilon\to\phi^0$ strongly in $L^2(\Omega)$, $\lambda(\Omega)/\epsilon\to\alpha_1(\Omega)$, without passing to a subsequence.) □

We next prove that all *higher* eigenvalues $\lambda_m(\Omega)$ are of order $\mathrm{O}(\epsilon)$; in fact we shall give first-order expansions in ϵ for $\lambda_m(\Omega)$ as well as ϕ_m^ϵ. Before doing so we discuss eigenvalues of the Rayleigh quotient (7) in the space H:

$$I(u)=\frac{\int_\Omega\sigma_1u_{x_1}^2+\cdots+\sigma_nu_{x_n}^2}{\int_\Omega u^2}.$$

The eigenvalue problem associated to $I(u)$ in H is

$$\int_\Omega \sigma_1 u_{x_1} v_{x_1} + \cdots + \sigma_n u_{x_n} v_{x_n} = \alpha \int_\Omega uv \quad \text{for all } v \in H.$$

If this equation has a nontrivial solution $u \in H$ for some α, then α is called an eigenvalue of $I(u)$ on H, and u is a corresponding eigenfunction. (If H were replaced by $H_0^1(\Omega)$, then we would have the classical Dirichlet eigenvalue problem.) By standard variational arguments (using among other things the compact imbedding $H \hookrightarrow L^2(\Omega)$; see [4] for the Dirichlet case) we have that

(i) all the eigenvalues of $I(u)$ on H form an unbounded sequence

$$\alpha_1(\Omega) < \alpha_2(\Omega) \le \alpha_3(\Omega) \le \cdots \to \infty;$$

(ii) corresponding to each $\alpha_m(\Omega)$, there exists a normalized eigenfunction satisfying the orthogonality condition

$$\int_\Omega \phi_m^0 \phi_l^0 = 0 = \int_\Omega \sigma_1 \left(\phi_m^0\right)_{x_1} \left(\phi_l^0\right)_{x_1} + \cdots + \sigma_n \left(\phi_m^0\right)_{x_n} \left(\phi_l^0\right)_{x_n}, \quad m \ne l; \tag{15}$$

(iii) $\alpha_1(\Omega)$ is described as in Theorem 1 and the other eigenvalues can be characterized variationally as

$$\alpha_m(\Omega) = \inf_{\substack{u \in H \\ \int_\Omega u\phi_l^0 = 0,\ 1 \le l \le m-1}} I(u);$$

(iv) each eigenvalue is repeated according to its geometric multiplicity.

THEOREM 2. *Assume the conditions of Theorem* 1. *Then as* $\epsilon \downarrow 0$, $\lambda_m(\Omega) = \epsilon\alpha_m(\Omega) + o(\epsilon)$, *and in the* L^2*-sense, after passing to a subsequence,* $\phi_m^\epsilon \to \widetilde{\phi}_m^0$, $m = 1, 2, \ldots$, *where* $\widetilde{\phi}_m^0$ *is a normalized eigenfunction corresponding to* $\alpha_m(\Omega)$.

Proof. Step 1. We prove $\limsup_{\epsilon \downarrow 0} \lambda_m(\Omega)/\epsilon \le \alpha_m(\Omega)$. Let

$$V_m = \operatorname{span}\{\phi_1^0, \ldots, \phi_m^0\} \subset H.$$

By the Poincaré principle (see, for example, [1, p. 97]),

$$\begin{aligned}
\lambda_m(\Omega) &\le \max_{u \in V_m} \frac{\int_\Omega a_{ij}(\mathbf{x}) u_{x_i} u_{x_j}}{\int_\Omega u^2} \\
&= \epsilon \max_{u \in V_m} \frac{\int_\Omega (\sigma_1 + \mathrm{o}(1)) u_{x_1}^2 + \cdots + (\sigma_n + \mathrm{o}(1)) u_{x_n}^2}{\int_\Omega u^2} \quad (\text{note that } \nabla u = 0 \text{ in } \Omega_1) \\
&= \epsilon(\alpha_m(\Omega) + \mathrm{o}(1)).
\end{aligned}$$

In the last equation, we used (15) and the elementary fact

$$\frac{a_1 + \cdots + a_n}{b_1 + \cdots + b_n} \le \max\left\{\frac{a_1}{b_1}, \ldots, \frac{a_n}{b_n}\right\} \quad \text{for positive } a\text{'s and } b\text{'s}. \tag{16}$$

Step 2. We prove $\liminf_{\epsilon \downarrow 0} \lambda_m(\Omega)/\epsilon \ge \alpha_m(\Omega)$. For $m = 1$ this has been proved in Theorem 1. For $m = 2$, arguing as in Step 3 of Theorem 1, we have that, after passing to a subsequence, $\phi_2^\epsilon \to$ some $\widetilde{\phi}_2^0$, where $\widetilde{\phi}_2^0$ satisfies

$$\widetilde{\phi}_2^0 \in H, \quad \int_\Omega \left(\widetilde{\phi}_2^0\right)^2 = 1, \quad \int_\Omega \phi_1^0 \widetilde{\phi}_2^0 = 0;$$

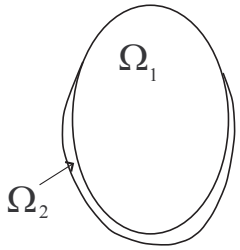


FIG. 2. *Incompletely coated body.*

thus

$$\begin{aligned}\liminf_{\epsilon\downarrow 0}\frac{\lambda_2(\Omega)}{\epsilon} &\geq \liminf_{\epsilon\downarrow 0}\int_{\Omega_2}(\sigma_1+\mathrm{o}(1))\,(\phi_2^\epsilon)^2_{x_1}+\cdots+(\sigma_n+\mathrm{o}(1))\,(\phi_2^\epsilon)^2_{x_n}\\ &\geq \int_{\Omega_2}\sigma_1\left(\widetilde{\phi}_2^0\right)^2_{x_1}+\cdots+\sigma_n\left(\widetilde{\phi}_2^0\right)^2_{x_n}\\ &= I(\widetilde{\phi}_2^0)\\ &\geq \alpha_2(\Omega).\end{aligned}$$

This and Step 1 imply that $\lim_{\epsilon\downarrow 0}\lambda_2(\Omega)/\epsilon=\alpha_2(\Omega)$ and that $\widetilde{\phi}_2^0$ is an eigenfunction corresponding to $\alpha_2(\Omega)$. Repeating the above argument, we complete the proof of the theorem. □

2.2. Incompletely coated bodies. In case Ω_2 does not completely coat Ω_1 (see Figure 2), we can show, for fixed $k>0$,

$$\lambda_1(\Omega)=\epsilon\widetilde{\lambda}(\Omega_2)+\mathrm{o}(\epsilon),\quad \epsilon\downarrow 0 \tag{17}$$

($\widetilde{\lambda}(\Omega_2)$ is defined as in part (2) of Theorem 1), and in the $L^2(\Omega)$-sense

$$\phi_1^\epsilon=\phi^0+\mathrm{o}(1), \tag{18}$$

where ϕ^0 on Ω_2 is the normalized eigenfunction corresponding to $\widetilde{\lambda}(\Omega_2)$ and ϕ^0 is understood as being zero on Ω_1. This can be proved by slightly modifying the proof of Theorem 1. One might be tempted to conclude, on the basis of (17), that even an incomplete coating will protect the body: The principal Dirichlet eigenvalue remains small if ϵ is small enough.[4] It is, however, physically obvious that the coating ought to be complete and that incomplete coating *is* unsatisfactory. The mathematical resolution has to do with the issue raised in item (iii) of section 1: By (18) the eigenfunction is *small* on the body, and hence the body is *not* protected from overheating.

In addition to (18), the size of ϕ_1^ϵ on Ω_1 can also be estimated more explicitly in terms of ϵ as follows. On Ω_1, we have

$$\Delta\phi_1^\epsilon=-\frac{\lambda_1(\Omega)}{k}\phi_1^\epsilon\,,$$

and so $\|\Delta\phi_1^\epsilon\|_{L^2(\Omega_1)}=\mathrm{O}(\epsilon/k)$. On the other hand,

$$\mathrm{O}(\epsilon)=\lambda_1(\Omega)\geq k\int_{\Omega_1}|\nabla\phi_1^\epsilon|^2\geq k\mu(\Omega_1)\int_{\Omega_1}(\phi_1^\epsilon)^2,$$

[4] Although if the coating is a tiny bump on the body, the coefficient of ϵ is very large.

where $\mu(\Omega_1)$ is the first eigenvalue of $-\Delta$ with Dirichlet boundary conditions on $\partial\Omega_1 \backslash \partial\Omega_2$, and Neumann on $\partial\Omega_1 \cap \partial\Omega_2$ (*assumed nonempty*). Thus $\|\phi_1^\epsilon\|_{L^2} = \mathrm{O}(\sqrt{\epsilon/k})$. By interior and boundary elliptic estimates, we infer

$$\|\phi_1^\epsilon\|_{H^2(K)} = \mathrm{O}(\sqrt{\epsilon/k}),$$

where K is any subdomain of Ω_1 with $\mathrm{dist}(K, \partial\Omega_2) > 0$. By the Sobolev imbedding $H^2(K) \subset L^\infty(K)$ (in 2D or 3D), $\|\phi_1^\epsilon\|_{L^\infty(K)} = \mathrm{O}(\sqrt{\epsilon/k})$.

3. How thin can the coating Ω_2 be? In this section, we assume that the coating Ω_2 has uniform thickness δ. In the intended applications, δ is much smaller than the length scale of the protected body Ω_1. We wish to find conditions that keep $\lambda_1(\Omega)$ small, while keeping Ω_2 as thin as possible. (If the coating is too thin, it can't insulate well. We will quantify this.)

Let $\mathbf{n}(p)$ be the unit outward normal vector field of $\partial\Omega_1$ at point p. For any small positive δ, define a mapping F as follows:

$$F : (p, \tau) \in \partial\Omega_1 \times [-\delta, \delta] \to (x_1, \ldots, x_n) \in R^n, \quad n > 1,$$

$$(x_1, \ldots, x_n) = F(p, \tau) = p + \tau \mathbf{n}(p).$$

If δ is small enough,

$$\delta \times (\text{maximum of the principal curvatures of } \partial\Omega_1) < 1$$

and perhaps smaller, then F is a diffeomorphism. We shall assume this throughout this section. We take

$$\Omega_2 = F(\partial\Omega_1 \times (0, \delta)), \tag{19}$$

and thus the thickness of Ω_2 is δ.

3.1. Upper bounds for $\lambda_1(\Omega)$.

THEOREM 3. *For any thermal diffusion coefficient $k > 0$ (see (5))*

$$\lambda_1(\Omega) \le \begin{cases} \dfrac{\overline{\sigma}_{\max}}{\delta|\Omega_1|}\left(|\partial\Omega_1| - \delta\left(\dfrac{\pi\chi(\partial\Omega_1)}{2} + \dfrac{|\partial\Omega_1|^2}{3|\Omega_1|}\right) + O(\delta^2)\right) & \text{in } 2D \\ \dfrac{\overline{\sigma}_{\max}}{\delta|\Omega_1|}\left(|\partial\Omega_1| - \delta\left(\overline{H} + \dfrac{|\partial\Omega_1|^2}{3|\Omega_1|}\right) + O(\delta^2)\right) & \text{in } 3D, \end{cases}$$

where $\overline{\sigma}_{max}$ is the largest of the eigenvalues of the thermal tensor of the coating Ω_2, $\overline{H}$ (in 3D) is the integral of the mean curvature of $\partial\Omega_1$ over $\partial\Omega_1$, and $\chi(\partial\Omega_1) = 2(1 - $ the number of holes in $\Omega_1)$ is the Euler characteristic of $\partial\Omega_1$.

Remark 2. Thus to keep $\lambda_1(\Omega)$ small, we just need to have $\overline{\sigma}_{\max}/\delta$ small. We emphasize that these upper bounds are independent of k (they apply even when the coated body is a near perfect conductor, $k \to \infty$), and in 2D, of the geometry of Ω_1. (In 2D, the upper bound depends only on the *topology* of Ω_1 and the relative magnitudes of $|\partial\Omega_1|$ and $|\Omega_1|$.) In 3D, the upper bound is more geometry-dependent but still only in a global way.

Proof. By the variational characterization,

$$\lambda_1(\Omega) = \inf_{u \in H_0^1(\Omega), u \not\equiv 0} \frac{\int_\Omega a_{ij}(\mathbf{x}) u_{x_i} u_{x_j}}{\int_\Omega u^2},$$

where (a_{ij}) is the thermal tensor of the composite; see (5).

We take

$$\phi(x_1,x_2,x_3)=\begin{cases}1-\dfrac{\tau}{\delta}, & 0\le\tau\le\delta,\ p\in\partial\Omega_1,\\ 1 & \text{otherwise,}\end{cases}$$

where we have used the coordinates introduced in (19). Then

$$\lambda_1(\Omega)\le\overline{\sigma}_{\max}\frac{\int_{\Omega_2}|\nabla\phi|^2}{\int_\Omega\phi^2}. \tag{20}$$

To compute this Rayleigh-type quotient in the three-dimensional case we need to use a convenient coordinate system. For every $q\in\partial\Omega_1$, parametrize $\partial\Omega_1$ near q by $p=p(u,v)$ (u, v real) such that on $\partial\Omega_1$ the curves $u\to p(u,v)$ and $v\to p(u,v)$ are curves of principal curvature of $\partial\Omega_1$ with speed equal to 1:

$$\begin{cases}|\partial p/\partial u|=1,\\ |\partial p/\partial v|=1,\\ \partial p/\partial u\perp\partial p/\partial v,\\ (\partial/\partial u)\mathbf{n}(p(u,v))=-k_1(p(u,v))(\partial/\partial u)p(u,v),\\ (\partial/\partial v)\mathbf{n}(p(u,v))=-k_2(p(u,v))(\partial/\partial u)p(u,v).\end{cases}$$

Here k_1 and k_2 are principal curvatures of $\partial\Omega_1$ (with the convention that they are positive if Ω_1 is convex).

Then the surface element on $\partial\Omega_1$ is given by

$$dS_p=\left|\frac{\partial p}{\partial u}\times\frac{\partial p}{\partial v}\right|du\,dv=du\,dv.$$

The volume element on Ω_2 is

$$\begin{aligned}dx_1\,dx_2\,dx_3&=\left|\left(\frac{\partial F}{\partial u}\times\frac{\partial F}{\partial v}\right)\cdot\frac{\partial F}{\partial\tau}\right|du\,dv\,d\tau\\&=\left|\left((p_u+\tau\mathbf{n}_u)\times(p_v+\tau\mathbf{n}_v)\right)\cdot\mathbf{n}\right|du\,dv\,d\tau\\&=\left|\left[(1-\tau k_1(p))p_u\times(1-\tau k_2(p))p_v\right]\cdot\mathbf{n}\right|du\,dv\,d\tau\\&=\left|(1-\tau k_1(p))(1-\tau k_2(p))\right|du\,dv\,d\tau\\&=(1-\tau k_1(p))(1-\tau k_2(p))\,dS_p\,d\tau\\&=\left(\tau^2G(p)-2\tau H(p)+1\right)\,dS_p\,d\tau,\end{aligned}$$

where $G(p)=k_1k_2$ is the Gauss curvature and $H(p)$ is the mean curvature of $\partial\Omega_1$.

By the chain rule, $\partial\phi/\partial u=\nabla_{(x_1,x_2,x_3)}\phi\cdot\frac{\partial F}{\partial u}=(1-\tau k_1(p))\,\nabla_{(x_1,x_2,x_3)}\phi\cdot p_u$ and also $\partial\phi/\partial v=(1-\tau k_2(p))\,\nabla_{(x_1,x_2,x_3)}\phi\cdot p_v$ and $\partial\phi/\partial\tau=\nabla_{(x_1,x_2,x_3)}\phi\cdot\mathbf{n}$. Thus

$$\left|\nabla_{(x_1,x_2,x_3)}\phi\right|^2=\phi_\tau^2+\frac{\phi_u^2}{(1-\tau k_1)^2}+\frac{\phi_v^2}{(1-\tau k_2)^2}.$$

Now we go back to (20) to compute

$$\begin{aligned}\int_{\Omega_2}\left|\nabla_{(x_1,x_2,x_3)}\phi\right|^2&=\int_0^\delta\int_{\partial\Omega_1}\phi_\tau^2(\tau^2G(p)-2\tau H(p)+1)\,dS_p\,d\tau\\&=\frac{1}{\delta^2}\int_0^\delta\int_{\partial\Omega_1}\left(\tau^2G(p)-2\tau H(p)+1\right)\,dS_p\,d\tau.\end{aligned}$$

Introducing $\overline{H} = \int_{\partial\Omega_1} H(p)\, dS_p$ and using the Gauss–Bonnet theorem $\int_{\partial\Omega_1} G(p) dS_p = 2\pi\chi(\partial\Omega_1)$, we have

$$\begin{aligned}\int_{\Omega_2} \left|\nabla_{(x_1,x_2,x_3)}\phi\right|^2 &= \frac{1}{\delta^2}\int_0^\delta \left(2\pi\chi(\partial\Omega_1)\tau^2 - 2\tau\overline{H} + |\partial\Omega_1|\right)\, d\tau \\ &= \frac{1}{\delta}\left[|\partial\Omega_1| - \delta\overline{H} + \frac{2\pi\chi(\partial\Omega_1)\delta^2}{3}\right]\end{aligned}$$

while

$$\begin{aligned}\int_\Omega \phi^2 &= |\Omega_1| + \int_0^\delta \left(1-\frac{\tau}{\delta}\right)^2 \left(|\partial\Omega_1| - 2\overline{H}\tau + 2\pi\chi(\partial\Omega_1)\tau^2\right) d\tau \\ &= |\Omega_1| + \frac{|\partial\Omega_1|\delta}{3} + O(\delta^2).\end{aligned}$$

When the quotient is assembled, and expanded for small δ, the result is the theorem in the three-dimensional case.

In the two-dimensional case we simply parametrize the curve $\partial\Omega_1$ by the arclength variable s (so $u = s$). In the computation of the Rayleigh quotient, set $G(p) = 0$ and $2H(p) = K(p)$, the curvature of $\partial\Omega_1$. In 2D the Gauss–Bonnet theorem is reduced to $\int_{\partial\Omega_1} K(p)\, ds = \pi\chi(\partial\Omega_1)$. We now have

$$\begin{aligned}\int_\Omega |\nabla\phi|^2 &= \frac{1}{\delta^2}\int_0^\delta \left(|\partial\Omega_1| - \pi\chi(\partial\Omega_1)\tau\right) d\tau \\ &= \frac{1}{\delta}\left(|\partial\Omega_1| - \frac{\pi\chi(\partial\Omega_1)\delta}{2}\right)\end{aligned}$$

and

$$\begin{aligned}\int_\Omega \phi^2 &= |\Omega_1| + \int_0^\delta \left(1-\frac{\tau}{\delta}\right)^2 \left(|\partial\Omega_1| - \pi\chi(\partial\Omega_1)\tau\right)\, d\tau \\ &= |\Omega_1| + \frac{|\partial\Omega_1|\delta}{3} + O(\delta^2).\end{aligned}$$

The two-dimensional inequality of the theorem now follows from assembling the quotient and expanding for small δ. □

3.2. Theorem 3 is sharp. We now show that Theorem 3 is sharp. This is accomplished by studying the situation in which $\delta \downarrow 0$ and k and the thermal tensor are allowed to vary with δ. As mentioned previously, the upper bound for $\lambda_1(\Omega)$ in Theorem 3 is *universal*: It is valid even when $k \to \infty$. The point of allowing the thermal tensor to vary with δ is to identify exactly the scaling relationships among δ, k, and the thermal tensor in determining the size of $\lambda_1(\Omega)$.

THEOREM 4. *Let $\overline{\sigma}_{\max}$ and $\overline{\sigma}_{\min}$ be the largest and smallest of the eigenvalues of the thermal tensor of the coating Ω_2. Assume that as $\delta \downarrow 0$*

$$\frac{\overline{\sigma}_{\max}}{\delta k} \to 0 \tag{21}$$

($k > 0$ is the thermal diffusion coefficient of the uncoated body) and

$$\frac{\overline{\sigma}_{\max}}{\overline{\sigma}_{\min}} \quad \textit{is bounded.} \tag{22}$$

Then

$$\liminf_{\delta\downarrow 0} \frac{\lambda_1(\Omega)\delta}{\overline{\sigma}_{\min}} \geq \frac{|\partial\Omega_1|}{|\Omega_1|}. \tag{23}$$

Moreover, the normalized eigenfunction ϕ_1 *corresponding to* $\lambda_1(\Omega)$ *converges to* $1/\sqrt{|\Omega_1|}$ *strongly in* $L^2(\Omega_1)$ *and* $C^2_{\rm loc}(\Omega_1)$ *as* $\delta \downarrow 0$.

Proof. Step 1. Let

$$\lambda_\delta = \inf_{u\in H^1_0(\Omega), u\not\equiv 0} \frac{k\int_{\Omega_1}|\nabla u|^2 + \overline{\sigma}_{\min}\int_{\Omega_2}|\nabla u|^2}{\int_\Omega u^2}. \tag{24}$$

Then $\lambda_1(\Omega) \geq \lambda_\delta$. Denote by ϕ^δ the normalized ($\int_\Omega(\phi^\delta)^2 = 1$) minimizer of (24). It is positive on Ω and is the weak solution of

$$\begin{cases} \sum(a(\mathbf{x})u_i)_i + \lambda_\delta u = 0 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where $u_i \equiv \partial u/\partial x_i$ and

$$a(\mathbf{x}) = \begin{cases} k & \text{in } \Omega_1, \\ \overline{\sigma}_{\min} & \text{in } \Omega_2. \end{cases}$$

By the De Giorgi–Nash estimates (see [4, Theorem 8.29]), ϕ^δ is Hölder-continuous throughout $\overline{\Omega}$. Since

$$\Delta\phi^\delta = -\frac{\lambda_\delta}{k}\phi^\delta \quad \text{in } \Omega_1$$

with $\lambda_\delta/k \leq \lambda_1(\Omega)/k \leq O(\frac{\overline{\sigma}_{\max}}{\delta k}) = o(1)$ (by Theorem 3 and (21)), we have, by the interior elliptic estimates and a bootstrap argument, that ϕ^δ is compact in $C^2_{\rm loc}(\Omega_1)$ as $\delta \downarrow 0$.

Step 2. Since $\int_{\Omega_1}|\nabla\phi^\delta|^2 \leq \lambda_\delta/k \to 0$ as $\delta \downarrow 0$, we have that, after passing to a subsequence of $\delta \downarrow 0$,

$$\phi^\delta \to \text{ some constant } m, \text{ weakly in } H^1(\Omega_1) \text{ and strongly in } L^2(\Omega_1). \tag{25}$$

Because ϕ^δ is compact in $C^2_{\rm loc}(\Omega_1)$ as $\delta \downarrow 0$, we also have

$$\phi^\delta \to m \text{ in } C^2_{\rm loc}(\Omega_1) \text{ as } \delta \downarrow 0 \tag{26}$$

after possibly passing to yet another subsequence.

By Lemma 5 and (22),

$$\int_{\Omega_2}(\phi^\delta)^2 \leq O(\delta^2)\int_{\Omega_2}|\nabla\phi^\delta|^2 \leq O\left(\frac{\delta^2}{\overline{\sigma}_{\min}}\right)\lambda_\delta \leq O\left(\frac{\delta^2\overline{\sigma}_{\max}}{\overline{\sigma}_{\min}\delta}\right) \to 0 \quad \text{as } \delta \downarrow 0.$$

Thus

$$1 = \int_\Omega(\phi^\delta)^2 = \int_{\Omega_1} m^2 + o(1) \quad \text{as } \delta \downarrow 0$$

and hence

$$m = \frac{1}{\sqrt{|\Omega_1|}}. \tag{27}$$

The proof that the eigenfunction ϕ_1 converges to $1/\sqrt{|\Omega_1|}$ is similar.

Step 3. We now show

$$\lim_{\delta \downarrow 0} \int_{\partial\Omega_1} \phi^\delta = m|\partial\Omega_1|. \tag{28}$$

Near $\partial\Omega_1$ think of ϕ^δ as a function of τ and $p \in \partial\Omega_1$ through the dependence $\phi^\delta = \phi^\delta(F(p,\tau))$. Fix a small $\delta_0 > 0$. Then

$$\left|\phi^\delta(-\delta_0, p) - \phi^\delta(0,p)\right| = \left|\int_0^{-\delta_0} \frac{\partial \phi^\delta}{\partial \tau}\, d\tau\right| \le \left(\int_{-\delta_0}^0 \left(\frac{\partial \phi^\delta}{\partial \tau}\right)^2 d\tau\right)^{1/2} \delta_0^{1/2}$$

and thus

$$\begin{aligned}
\int_{\partial\Omega_1} \left(\phi^\delta(-\delta_0,p) - \phi^\delta(0,p)\right)^2 dS_p &\le \delta_0 \int_{\partial\Omega_1} \int_{-\delta_0}^0 \left(\frac{\partial \phi^\delta}{\partial \tau}\right)^2 d\tau\, dS_p \\
&\le \frac{\delta_0}{1 - O(\delta_0)} \int_{\partial\Omega_1} \int_{-\delta_0}^0 \left(\frac{\partial \phi^\delta}{\partial \tau}\right)^2 \left(1 - 2\tau H(p) + \tau^2 G(p)\right) d\tau\, dS_p \\
&\le \frac{\delta_0}{1 - O(\delta_0)} \int_{F(\partial\Omega_1 \times (-\delta_0, 0))} \left|\nabla_{(x_1,x_2,x_3)} \phi^\delta\right|^2 dx\, dy\, dz \\
&\to 0 \quad \text{as } \delta \downarrow 0.
\end{aligned}$$

Thus

$$\begin{aligned}
\lim_{\delta \downarrow 0} \int_{\partial\Omega_1} \phi^\delta(0,p)\, dS_p &= \lim_{\delta \downarrow 0} \int_{\partial\Omega_1} \phi^\delta(-\delta_0, p)\, dS_p \\
&= m|\partial\Omega_1| \quad \text{by (26)}.
\end{aligned}$$

Step 4.

$$\begin{aligned}
\int_{\partial\Omega_1} \phi^\delta(0,p)\, dS_p &= \int_{\partial\Omega_1} \left(\phi^\delta(0,p) - \phi^\delta(\delta,p)\right) dS_p \\
&= \int_{\partial\Omega_1} \int_\delta^0 \frac{\partial \phi^\delta}{\partial \tau}\, d\tau\, dS_p \\
&\le \left(\int_{\partial\Omega_1} \int_0^\delta \left(\frac{\partial \phi^\delta}{\partial \tau}\right)^2 d\tau\, dS_p\right)^{1/2} \left(\delta|\partial\Omega_1|\right)^{1/2},
\end{aligned}$$

which implies

$$\begin{aligned}
\left(\int_{\partial\Omega_1} \phi^\delta(0,p)\, dS_p\right)^2 &\le \delta|\partial\Omega_1| \int_{\partial\Omega_1} \int_0^\delta \left(\frac{\partial \phi^\delta}{\partial \tau}\right)^2 d\tau\, dS_p \\
&\le \frac{\delta|\partial\Omega_1|}{1 - O(\delta)} \int_{\partial\Omega_1} \int_0^\delta \left(\frac{\partial \phi^\delta}{\partial \tau}\right)^2 \left(1 - 2\tau H(p) + \tau^2 G(p)\right) d\tau\, dS_p \\
&\le \frac{\delta|\partial\Omega_1|}{1 - O(\delta)} \int_{\Omega_2} \left|\nabla_{(x_1,x_2,x_3)} \phi^\delta\right|^2 dx_1\, dx_2\, dx_3 \\
&\le \frac{\delta|\partial\Omega_1|}{1 - O(\delta)} \frac{\lambda_\delta}{\overline{\sigma}_{\min}}.
\end{aligned}$$

Consequently,

$$\liminf_{\delta\downarrow 0} \frac{\delta\lambda_\delta}{\overline{\sigma}_{\min}} \geq \frac{(m|\partial\Omega_1|)^2}{|\partial\Omega_1|} = \frac{|\partial\Omega_1|}{|\Omega_1|}. \qquad \square$$

LEMMA 5. *For all small $\delta > 0$ and $\phi \in H^1(\Omega_2)$ with $\phi = 0$ on the outer boundary of Ω_2,*

$$\int_{\Omega_2} |\nabla\phi|^2 \geq \frac{\pi^2\left(1 - 2\delta H_{\max} + \delta^2 G_{\min}\right)}{4\delta^2\left(1 - 2\delta H^-_{\min} + \delta^2 G_{\max}\right)} \int_{\Omega_2} \phi^2,$$

where $H_{\max} \equiv \max_{p\in\partial\Omega_1} H(p)$, $H^-_{\min} \equiv \min\left(0, \min_{p\in\partial\Omega_1} H(p)\right)$ and $G_{\max}$ and $G_{\min}$ are the maximum and minimum of G on $\partial\Omega_1$. In 2D, $G(p) \equiv 0$ and $2H(p)$ is understood as $K(p)$.

Proof. The smallest eigenvalue of $-d^2/dx^2$ on $(0,\delta)$ with Neumann condition on the left and Dirichlet on the right is $\frac{\pi^2}{4\delta^2}$. Thus

$$\int_0^\delta \left(\frac{\partial\phi(\tau,p)}{\partial\tau}\right)^2 d\tau \geq \frac{\pi^2}{4\delta^2}\int_0^\delta \phi^2(\tau,p)\,d\tau,$$

which implies

$$\frac{1}{1 - 2\delta H_{\max} + \delta^2 G_{\min}} \int_{\partial\Omega_1}\int_0^\delta \left(\frac{\partial\phi}{\partial\tau}\right)^2 \left(1 - 2\tau H(p) + \tau^2 G(p)\right)\, d\tau\, dS_p$$
$$\geq \frac{\pi^2}{4\delta^2} \frac{\int_{\partial\Omega_1}\int_0^\delta \phi^2(\tau,p)\left(1 - 2\tau H(p) + \tau^2 G(p)\right)\, d\tau\, dS_p}{1 - 2\delta H^-_{\min} + \delta^2 G_{\max}}.$$

Now the conclusion follows from the fact that the integral on the left is less than or equal to $\int_{\Omega_2} |\nabla\phi|^2$. $\square$

3.3. Optimally aligned coating. If the thermal tensor of the coating varies spatially, i.e., $\overline{a}_{ij} = \overline{a}_{ij}(\mathbf{x})$, then it is possible for the principal Dirichlet eigenvalue to be small, as in Theorem 2, even if *not all* the eigenvalues of this thermal tensor are small. We call the coating *optimally aligned* if the smallest of the eigenvalues has, at each boundary point of the coated body, eigenvector normal to the boundary (and this persists into the coating; see below). Only *that* eigenvalue needs to be controlled (assumed small everywhere); the size of the others is irrelevant. An example is the two-dimensional disk of radius $1-\delta$ coated by an annulus of thickness δ, the annulus having thermal tensor

$$A = \begin{pmatrix} \epsilon x^2 + y^2 & xy(-1+\epsilon) \\ xy(-1+\epsilon) & x^2 + \epsilon y^2 \end{pmatrix}. \tag{29}$$

The eigenvalues are $x^2 + y^2$ and $\epsilon(x^2 + y^2)$ with eigenvectors angular and radial, respectively.

THEOREM 6. *Suppose for all $p \in \partial\Omega_1$ and all points $\mathbf{x} = p + \tau\mathbf{n}(p)$ $(\tau > 0)$ in Ω_2 the smallest eigenvalue $\overline{\sigma}^{\rm op}_{\min}(\mathbf{x})$ of $\overline{a}_{ij}(\mathbf{x})$ has eigenvector $\mathbf{n}(p)$. Let*

$$\overline{\sigma}^{\rm op}_{\min} = \max_{\mathbf{x}\in\overline{\Omega}_2} \overline{\sigma}^{\rm op}_{\min}(\mathbf{x}).$$

Then

$$\lambda_1(\Omega) \le \begin{cases} \dfrac{\overline{\sigma}^{\rm op}_{\rm min}}{\delta|\Omega_1|}\left(|\partial\Omega_1| - \delta\left(\dfrac{\pi\chi(\partial\Omega_1)}{2} + \dfrac{|\partial\Omega_1|^2}{3|\Omega_1|}\right) + O(\delta^2)\right) & \textit{in } 2D, \\ \dfrac{\overline{\sigma}^{\rm op}_{\rm min}}{\delta|\Omega_1|}\left(|\partial\Omega_1| - \delta\left(\overline{H} + \dfrac{|\partial\Omega_1|^2}{3|\Omega_1|}\right) + O(\delta^2)\right) & \textit{in } 3D. \end{cases}$$

Proof. The proof is a slight modification of that of Theorem 3. Using the same test function ϕ and observing that $\nabla\phi(x_1, x_2, x_3) = \frac{\partial\phi}{\partial\tau}\mathbf{n}(p) = -\frac{\mathbf{n}(p)}{\delta}$ is an eigenvector of $\overline{a}_{ij}(\mathbf{x})$ corresponding to $\overline{\sigma}^{\rm op}_{\rm min}(\mathbf{x})$, we have

$$\begin{aligned} \lambda_1(\Omega) &\le \frac{\int_{\Omega_2} \overline{a}_{ij}(\mathbf{x})\phi_{x_i}\phi_{x_j}}{\int_\Omega \phi^2} \\ &= \frac{\int_{\Omega_2} \overline{\sigma}^{\rm op}_{\rm min}(\mathbf{x})|\nabla\phi|^2}{\int_\Omega \phi^2} \\ &\le \frac{\overline{\sigma}^{\rm op}_{\rm min}}{\delta^2}\frac{\int_{\Omega_2} 1}{\int_\Omega \phi^2}. \end{aligned}$$

The remaining calculations will be the same as in the proof of Theorem 3. ∎

THEOREM 7. *Let*

$$\underline{\sigma}^{\rm op}_{\rm min} = \min_{\mathbf{x}\in\overline{\Omega}_2} \overline{\sigma}^{op}_{min}(\mathbf{x}).$$

If

$$\lim_{\delta\downarrow 0}\frac{\overline{\sigma}^{\rm op}_{\rm min}}{\delta k} = 0 \quad \textit{and} \quad \frac{\overline{\sigma}^{\rm op}_{\rm min}}{\underline{\sigma}^{\rm op}_{\rm min}} \quad \textit{is bounded,}$$

then

$$\liminf_{\delta\downarrow 0}\frac{\lambda_1(\Omega)\delta}{\underline{\sigma}^{\rm op}_{\rm min}} \ge \frac{|\partial\Omega_1|}{|\Omega_1|},$$

and the normalized eigenfunction ϕ_1 corresponding to $\lambda_1(\Omega)$ converges to $1/\sqrt{|\ \Omega_1\ |}$ (strongly in $L^2(\Omega_1)$ and $C^2_{\rm loc}(\Omega_1)$) as $\delta \downarrow 0$.

Proof. The proof is the same as the proof of Theorem 4 but with $\overline{\sigma}_{\rm min} \to \underline{\sigma}^{\rm op}_{\rm min}$, $\overline{\sigma}_{\rm max} \to \overline{\sigma}^{\rm op}_{\rm min}$, and "Theorem 3" → "Theorem 6." ∎

4. Estimation of higher eigenpairs. A fuller understanding of the long-time behavior (i.e., of the accuracy of the truncated eigenfunction expansion approximation) can be achieved by estimating the higher eigenpairs $\lambda_m(\Omega)$, $\phi^\delta_m(\mathbf{x})$, $m \ge 2$. A by-product is a new characterization of the first positive Neumann eigenvalue as the limit of a sequence of second Dirichlet eigenvalues.

4.1. The case in which all eigenvalues of the thermal tensor are small.

THEOREM 8. *Assume the conditions of Theorem 4 with $k > 0$ fixed. Suppose also that*

$$\frac{\delta^2}{\overline{\sigma}_{\rm min}} \downarrow 0 \quad \textit{as } \delta \downarrow 0. \tag{30}$$

Then

$$\lim_{\delta\downarrow 0}\lambda_2(\Omega)=k\mu_2(\Omega_1),$$

where $\mu_2(\Omega_1)$ is the first positive eigenvalue of the Neumann eigenproblem

$$\begin{cases}\Delta u+\mu u=0 & \text{in } \Omega_1,\\ \dfrac{\partial u}{\partial \mathbf{n}}=0 & \text{on } \partial\Omega_1,\ \mathbf{n} \text{ the unit outward normal to } \partial\Omega_1.\end{cases}\tag{31}$$

Moreover, after passing to a subsequence the normalized eigenfunction ϕ_2^δ corresponding to $\lambda_2(\Omega)$ converges to a normalized eigenfunction ψ_2^0 associated to $\mu_2(\Omega_1)$ strongly in $L^2(\Omega_1)$ and $C^2_{\rm loc}(\Omega_1)$.

Proof. Step 1. We prove $\limsup_{\delta\downarrow 0}\lambda_2(\Omega)\le k\mu_2(\Omega_1)$. Recall the Poincaré principle

$$\lambda_2(\Omega)=\min_{\dim V_2=2}\max_{u\in V_2}\frac{\int_\Omega a_{ij}(\mathbf{x})u_iu_j\,d\mathbf{x}}{\int_\Omega u^2\,d\mathbf{x}},\tag{32}$$

where the minimum is taken over all two-dimensional subspaces V_2 of $H_0^1(\Omega)$. Let ψ^0 be a fixed normalized eigenfunction associated to $\mu_2(\Omega_1)$. Since $\psi^0\in C^2(\overline{\Omega}_1)$ we can extend it so that $\psi^0\in C^2(R^n)$. Let ϕ_1^δ be the normalized eigenfunction corresponding to $\lambda_1(\Omega)$. In (32) take $V_2=\operatorname{span}\{\phi_1^\delta,\phi\psi^0\}$, where ϕ is as defined in the proof of Theorem 3. For any $u\in V_2$ write $u=c_1\phi_1^\delta+c_2\phi\psi^0$, where c_1 and c_2 are constants. Now by (32)

$$\begin{aligned}\lambda_2&\le\max_{u\in V_2}\frac{\int_\Omega a_{ij}(\mathbf{x})u_iu_j\,d\mathbf{x}}{\int_\Omega u^2\,d\mathbf{x}}\\&\le\max_{(c_1,c_2)\in R^2}\frac{c_1^2\,(\lambda_1(\Omega)+A)+c_2^2(B+A)}{c_1^2(1-D)+c_2^2(C-D)},\end{aligned}\tag{33}$$

where

$$\begin{aligned}A&=\left|\int_\Omega a_{ij}(\mathbf{x})(\phi_1^\delta)_{x_i}(\phi\psi^0)_{x_j}\right|\\&\le\left(\int_\Omega a_{ij}(\mathbf{x})(\phi_1^\delta)_{x_i}(\phi_1^\delta)_{x_j}\right)^{1/2}\left(\int_\Omega a_{ij}(\mathbf{x})(\phi\psi^0)_{x_i}(\phi\psi^0)_{x_j}\right)^{1/2}\\&=\sqrt{\lambda_1(\Omega)}\sqrt{B};\\B&=\int_\Omega a_{ij}(\mathbf{x})(\phi\psi^0)_{x_i}(\phi\psi^0)_{x_j}\\&=k\int_{\Omega_1}|\nabla\psi^0|^2+\int_{\Omega_2}\overline{a}_{ij}(\phi_{x_i}\psi^0+\phi\psi^0_{x_i})(\phi_{x_j}\psi^0+\phi\psi^0_{x_j})\\&\le k\int_{\Omega_1}|\nabla\psi^0|^2+2\int_{\Omega_2}\overline{a}_{ij}\phi_{x_i}\phi_{x_j}(\psi^0)^2+2\int_{\Omega_2}\overline{a}_{ij}\psi^0_{x_i}\psi^0_{x_j}\phi^2\\&\le k\mu_2(\Omega_1)+\mathrm{O}\left(\frac{\overline{\sigma}_{\max}}{\delta}\right)+\mathrm{O}(\delta);\\C&=\int_\Omega(\phi\psi^0)^2\end{aligned}$$

$$
\begin{aligned}
&= \int_{\Omega_1} (\psi^0)^2 + \int_{\Omega_2} (\phi\psi^0)^2 \geq 1;\\
D &= \left| \int_\Omega \phi_1^\delta \phi\psi^0 \right| \\
&\leq \left| \int_{\Omega_1} \phi_1^\delta \psi^0 \right| + \left| \int_{\Omega_2} \phi_1^\delta \phi\psi^0 \right| \\
&\leq \left| \int_{\Omega_1} \left(\phi_1^\delta - \frac{1}{\sqrt{|\Omega_1|}} \right) \psi^0 \right| + \left(\int_{\Omega_2} (\psi^0)^2 \right)^{1/2} \left(\int_\Omega (\phi_1^\delta)^2 \right)^{1/2} \\
&\leq \left(\int_{\Omega_2} \left(\phi_1^\delta - \frac{1}{\sqrt{|\Omega_1|}} \right)^2 \right)^{1/2} \left(\int_{\Omega_1} (\psi^0)^2 \right)^{1/2} + \left(\int_{\Omega_2} (\psi^0)^2 \right)^{1/2} \\
&= \mathrm{o}(1) + \mathrm{O}(\sqrt{\delta}) \quad \text{(by Theorem 4).}
\end{aligned}
$$

These estimates, the fact that $\lambda_1(\Omega) = \mathrm{O}(\overline{\sigma}_{\max}/\delta)$, and (33) imply

$$\lambda_2(\Omega) \leq \frac{c_1^2 \mathrm{o}(1) + c_2^2 \left(k\mu_2(\Omega_1) + \mathrm{o}(1) \right)}{c_1^2(1 - \mathrm{o}(1)) + c_2^2(1 - \mathrm{o}(1))}.$$

From the elementary inequality $\frac{a+b}{c+d} \leq \max\left(\frac{a}{c}, \frac{b}{d}\right)$ for positive a, b, c, d we infer

$$\lambda_2(\Omega) \leq \max\left(\frac{\mathrm{o}(1)}{1 - \mathrm{o}(1)}, \frac{k\mu_2(\Omega_1) + \mathrm{o}(1)}{1 - \mathrm{o}(1)} \right).$$

Step 2. We prove $\liminf_{\delta \downarrow 0} \lambda_2(\Omega) \geq k\mu_2(\Omega_1)$. Observe that

$$\lambda_2(\Omega) \geq k \int_{\Omega_1} |\nabla \phi_2^\delta|^2 + \overline{\sigma}_{\min} \int_{\Omega_2} |\nabla \phi_2^\delta|^2,$$

where $\int_\Omega (\phi_2^\delta)^2 = 1$ and $\int_\Omega \phi_1^\delta \phi_2^\delta = 0$. In particular ϕ_2^δ is bounded in $H^1(\Omega_1)$, and hence after passing to a subsequence, $\phi_2^\delta \to$ some ψ_2^0 weakly in $H^1(\Omega_1)$ and strongly in $L^2(\Omega_1)$. Applying interior elliptic estimates to $\Delta\phi_2^\delta + \lambda_2(\Omega)\phi_2^\delta = 0$, we also have $\phi_2^\delta \to \psi_2^0$ in $C^2_{\mathrm{loc}}(\Omega_1)$. From this it follows that

$$\liminf_{\delta \downarrow 0} \lambda_2(\Omega) \geq k \int_{\Omega_1} |\nabla \psi_2^0|^2 \tag{34}$$

and that

$$
\begin{aligned}
0 &= \lim_{\delta \downarrow 0} \int_\Omega \phi_1^\delta \phi_2^\delta \\
&= \lim_{\delta \downarrow 0} \left(\int_{\Omega_1} + \int_{\Omega_2} \right) \phi_1^\delta \phi_2^\delta \\
&= \int_{\Omega_1} \frac{1}{\sqrt{|\Omega_1|}} \psi_2^0 + \lim_{\delta \downarrow 0} \left(\mathrm{O}\left(\frac{\delta \overline{\sigma}_{\max}}{\overline{\sigma}_{\min}} \right) \right)^{1/2} \quad \text{(by Theorem 4 and its proof)} \\
&= \int_{\Omega_1} \frac{1}{\sqrt{|\Omega_1|}} \psi_2^0;
\end{aligned}
$$

i.e.,

$$\int_{\Omega_1} \psi_2^0 = 0. \tag{35}$$

Moreover, by Lemma 5, we have

$$\int_{\Omega_2} \left(\phi_2^\delta\right)^2 \le \mathrm{O}(\delta^2) \int_{\Omega_2} |\nabla \phi_2^\delta|^2 = \mathrm{O}\left(\frac{\delta^2}{\overline{\sigma}_{\min}}\right) \lambda_2(\Omega) = \mathrm{o}(1)$$

and thus

$$1 = \lim_{\delta \downarrow 0} \left(\int_{\Omega_1} \left(\phi_2^\delta\right)^2 + \int_{\Omega_2} \left(\phi_2^\delta\right)^2 \right) = \int_{\Omega_1} \left(\psi_2^0\right)^2.$$

Now this and (35) imply

$$\mu_2(\Omega) = \inf_{\int_{\Omega_1} u = 0} \frac{\int_{\Omega_1} |\nabla u|^2}{\int_{\Omega_1} u^2} \le \frac{\int_{\Omega_1} |\nabla \psi_2^0|^2}{\int_{\Omega_1} \left(\psi_2^0\right)^2} = \int_{\Omega_1} |\nabla \psi_2^0|^2.$$

This and (34) yield the desired conclusion.

Step 3. Combining Steps 1 and 2 we have

$$\lim_{\delta \downarrow 0} \lambda_2(\Omega) = k \mu_2(\Omega_1).$$

The arguments in Step 2 show that ψ_2^0 is a normalized eigenfunction corresponding to $\mu_2(\Omega_1)$. □

THEOREM 9.

(i) *For any $k > 0$ (see (5)) and $m = 2, 3, \ldots,$*

$$\lambda_m(\Omega) \le \begin{cases} \dfrac{\overline{\sigma}_{\max} m^2 \pi^2}{\delta^2} \left(1 + \dfrac{\delta \pi}{|\partial \Omega_1|} |\chi(\partial \Omega_1)| + O(\delta^2)\right) & \text{in } 2D, \\ \dfrac{\overline{\sigma}_{\max} m^2 \pi^2}{\delta^2} \left(1 + \dfrac{2\delta}{|\partial \Omega_1|} |\overline{H}| + O(\delta^2)\right) & \text{in } 3D. \end{cases}$$

In particular, if

$$\frac{\overline{\sigma}_{\max}}{\delta^2} \to 0 \quad \left(\iff \frac{\delta^2}{\overline{\sigma}_{\max}} \to \infty \right) \quad \text{as } \delta \downarrow 0, \tag{36}$$

then $\lambda_m(\Omega) \to 0$ as $\delta \downarrow 0$ uniformly in $k > 0$.

(ii) *Fix $k > 0$ and assume (36) and (22). Then the normalized eigenfunction ϕ_m^δ corresponding to $\lambda_m(\Omega)$ converges to zero strongly in $L^2(\Omega_1)$ and $C^2_{\mathrm{loc}}(\Omega_1)$ as $\delta \downarrow 0$, $m \ge 2$.*

(iii) *Under the same conditions as in part* (ii) *above, there exists a positive constant C such that*

$$\lambda_2(\Omega) \ge C \frac{\overline{\sigma}_{\max}}{\delta^2}.$$

Thus for $m \ge 2$ each $\lambda_m(\Omega)$ is of order $\overline{\sigma}_{\max}/\delta^2$, while $\lambda_1(\Omega)$ is of order $\overline{\sigma}_{\max}/\delta$.

Proof of (i). Fix $w \in H_0^1(\Omega_2) \subset H_0^1(\Omega)$, which depends only on $\tau \in [0,\delta]$. We compute the Rayleigh quotient

$$\begin{aligned}\frac{\int_\Omega a_{ij}(\mathbf{x}) w_i w_j \, d\mathbf{x}}{\int_\Omega w^2 \, d\mathbf{x}} &\le \frac{\overline{\sigma}_{\max} \int_{\Omega_2} |\nabla w|^2 \, d\mathbf{x}}{\int_{\Omega_2} w^2 \, d\mathbf{x}} \\ &= \overline{\sigma}_{\max} \frac{\int_0^\delta \int_{\partial\Omega_1} w_\tau^2 \left(\tau^2 G(p) - 2\tau H(p) + 1\right) \, dS_p \, d\tau}{\int_0^\delta \int_{\partial\Omega_1} w^2 \left(\tau^2 G(p) - 2\tau H(p) + 1\right) \, dS_p \, d\tau} \\ &\le \overline{\sigma}_{\max} \frac{\left(|\partial\Omega_1| - 2\delta \min(0, \overline{H}) + \mathrm{O}(\delta^2)\right) \int_0^\delta w_\tau^2 \, d\tau}{\left(|\partial\Omega_1| - 2\delta \max(0, \overline{H}) + \mathrm{O}(\delta^2)\right) \int_0^\delta w^2 \, d\tau} \\ &= \overline{\sigma}_{\max} \left(1 + \frac{2\delta|\overline{H}|}{|\partial\Omega_1|} + \mathrm{O}(\delta^2)\right) \frac{\int_0^\delta w_\tau^2 \, d\tau}{\int_0^\delta w^2 \, d\tau}.\end{aligned}$$

Take an m-dimensional subspace of $H_0^1(\Omega_2) \subset H_0^1(\Omega)$:

$$V_m = \operatorname{span}\left\{\sin\frac{\pi\tau}{\delta}, \sin\frac{2\pi\tau}{\delta}, \ldots, \sin\frac{m\pi\tau}{\delta}\right\}.$$

Then by the Poincaré principle

$$\begin{aligned}\lambda_m(\Omega) &\le \max_{w \in V_m} \frac{\int_\Omega a_{ij}(\mathbf{x}) w_i w_j \, d\mathbf{x}}{\int_\Omega w^2 \, d\mathbf{x}} \\ &\le \overline{\sigma}_{\max} \left(1 + \frac{2\delta|\overline{H}|}{|\partial\Omega_1|} + \mathrm{O}(\delta^2)\right) \max_{w \in V_m} \frac{\int_0^\delta w_\tau^2 \, d\tau}{\int_0^\delta w^2 \, d\tau} \\ &= \overline{\sigma}_{\max} \left(1 + \frac{2\delta|\overline{H}|}{|\partial\Omega_1|} + \mathrm{O}(\delta^2)\right) \frac{m^2\pi^2}{\delta^2}. \qquad \square\end{aligned}$$

Proof of (ii). By Theorem 4 and its proof, ϕ_1^δ (the normalized eigenfunction corresponding to $\lambda_1(\Omega)$) and ϕ_m^δ satisfy

$$\phi_1^\delta \to \frac{1}{\sqrt{|\Omega_1|}} \quad \text{and} \quad \phi_m^\delta \to \text{constant } C_m \quad \text{strongly in } L^2(\Omega_1) \text{ and } C^2_{\mathrm{loc}}(\Omega_1)$$

as $\delta \downarrow 0$. On the other hand,

$$0 = \int_\Omega \phi_1^\delta \phi_m^\delta = \int_{\Omega_1} \phi_1^\delta \phi_m^\delta + \int_{\Omega_2} \phi_1^\delta \phi_m^\delta,$$

while

$$\int_{\Omega_1} \phi_1^\delta \phi_m^\delta \to \frac{C_m |\Omega_1|}{\sqrt{|\Omega_1|}}$$

and

$$\left| \int_{\Omega_2} \phi_1^\delta \phi_m^\delta \right| \le \left(\int_{\Omega_2} \left(\phi_1^\delta\right)^2 \right)^{1/2} \left(\int_\Omega \left(\phi_m^\delta\right)^2 \right)^{1/2} = \mathrm{O}\left(\frac{\overline{\sigma}_{\max}\delta}{\overline{\sigma}_{\min}} \right) \to 0.$$

Thus $C_m = 0$. $\square$

Proof of (iii). Suppose there exists a sequence $\delta \downarrow 0$ such that

$$\frac{\delta^2 \lambda_2(\Omega)}{\overline{\sigma}_{\max}} \to 0. \tag{37}$$

Observe that

$$\begin{aligned}
1 &= \int_\Omega \left(\phi_2^\delta\right)^2 \\
&= \int_{\Omega_1} \left(\phi_2^\delta\right)^2 + \int_{\Omega_2} \left(\phi_2^\delta\right)^2 \\
&= \mathrm{o}(1) + \mathrm{O}(\delta^2) \int_{\Omega_2} |\nabla \phi_2^\delta|^2 \quad \text{by (ii) and Lemma 5} \\
&\le \mathrm{o}(1) + \frac{\mathrm{O}(\delta^2)}{\overline{\sigma}_{\min}} \lambda_2(\Omega) \\
&= \mathrm{o}(1) + \mathrm{O}(1) \frac{\delta^2 \lambda_2(\Omega)}{\overline{\sigma}_{\max}} \\
&= \mathrm{o}(1) \quad \text{by (37)},
\end{aligned}$$

which is a contradiction. □

4.2. The optimally aligned case.

THEOREM 10. *Let all the conditions of Theorem* 7 *hold with thermal conductivity coefficient* $k > 0$ *(see* (5)*) fixed. Assume also that*

$$\frac{\delta^2}{\underline{\sigma}^{\mathrm{op}}_{\min}} \to 0 \quad \textit{as } \delta \downarrow 0.$$

Then $\lim_{\delta \downarrow 0} \lambda_2(\Omega) = k\mu_2(\Omega_1)$ *and after passing to a subsequence* $\lim_{\delta \downarrow 0} \psi_2^\delta = \phi_2^0$ *in* $L^2(\Omega_1)$ *and* $C^2_{\mathrm{loc}}(\Omega_1)$.

Proof. We slightly modify the proof of Theorem 8. The first modification occurs in the estimate for B, where we use the fact that $\nabla\phi$ is an eigenvector of $(\overline{a}_{ij}(\mathbf{x}))$ associated with eigenvalue $\overline{\sigma}^{\mathrm{op}}_{\min}(\mathbf{x})$, and hence "$\mathrm{O}(\overline{\sigma}_{\max}/\delta)$" can be replaced by "$\mathrm{O}(\overline{\sigma}^{\mathrm{op}}_{\min}/\delta)$." The term "$\mathrm{O}(\delta)$" is kept unchanged because $\overline{a}_{ij}(\mathbf{x})$ is bounded on Ω_2. The other modifications are obvious ones: "Theorem 4" → "Theorem 7," $\overline{\sigma}_{\min} \to \underline{\sigma}^{\mathrm{op}}_{\min}$, and $\overline{\sigma}_{\max} \to \overline{\sigma}^{\mathrm{op}}_{\min}$. □

THEOREM 11.

(i) *For any thermal diffusion constant* $k > 0$ *(see* (5)*) and* $m = 2, 3, \ldots,$

$$\lambda_m(\Omega) \le \begin{cases} \dfrac{\overline{\sigma}^{\mathrm{op}}_{\min} m^2 \pi^2}{\delta^2} \left(1 + \dfrac{\delta\pi}{|\partial\Omega_1|} |\chi(\Omega_1)| + O(\delta^2)\right) & \textit{in } 2D, \\ \dfrac{\overline{\sigma}^{\mathrm{op}}_{\min} m^2 \pi^2}{\delta^2} \left(1 + \dfrac{2\delta|\overline{H}|}{|\partial\Omega_1|} + O(\delta^2)\right) & \textit{in } 3D. \end{cases}$$

In particular, if

$$\overline{\sigma}^{\mathrm{op}}_{\min}/\delta^2 \to 0 \quad \textit{as} \quad \delta \downarrow 0, \tag{38}$$

then $\lambda_m(\Omega) \to 0$ *as* $\delta \downarrow 0$, *uniformly for* $k > 0$.

(ii) *Fix* $k > 0$ *and assume* (38) *and that* $\overline{\sigma}^{\mathrm{op}}_{\min}/\underline{\sigma}^{\mathrm{op}}_{\min}$ *is bounded. Then* $\phi_m^\delta \to 0$ *strongly in* $L^2(\Omega_1)$ *and* $C^2_{\mathrm{loc}}(\Omega_1)$ *as* $\delta \downarrow 0$.

(iii) *Under the conditions in part* (ii)*, there exists a positive constant* c *such that* $c\overline{\sigma}^{\mathrm{op}}_{\min}/\delta^2 \leq \lambda_2(\Omega)$.

Proof. Modify the proof of Theorem 9 with the same obvious changes listed in the proof of Theorem 10. The key observation is that if $w = w(\tau)$, then in Ω_2, ∇w is an eigenvector of $\overline{a}_{ij}(\mathbf{x})$ associated to $\overline{\sigma}_{\min}(\mathbf{x})$. □

Acknowledgments. We would like to thank Greg Forest, who started us thinking about these questions and invited us to the February 2004 SAMSI conference, Workshop on Multi-scale Challenges in Soft Matter Materials, which we found very stimulating. Later we had several very useful conversations with Greg Forest, Rob Lipton, Wei-Ming Ni, and Hans Weinberger. We thank the referee for suggesting the consideration of spatially varying thermal tensors, and for several other helpful remarks. Part of the work of X. Wang was performed while he was supported by Wei-Ming Ni during his visit to University of Minnesota.

REFERENCES

[1] C. Bandle, *Isoperimetric Inequalities and Applications*, Pitman, Boston, London, 1980.
[2] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, New York, 1978.
[3] A. Friedman, *Reinforcement of the principal eigenvalue of an elliptic operator*, Arch. Rational Mech. Anal., 73 (1980), pp. 1–17.
[4] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 3rd ed., Springer-Verlag, Berlin, 1998.
[5] G. W. Milton, *The Theory of Composites*, Cambridge University Press, Cambridge, UK, 2002.
[6] X. Zheng, M. G. Forest, R. Lipton, R. Zhou, and Q. Wang, *Exact scaling laws for electrical conductivity properties of nematic polymer nano-composite monodomains*, Adv. Funct. Mat., 15 (2005), pp. 627–638.

SIAM J. APPL. MATH.
Vol. 68, No. 4, pp. 1202

CORRIGENDUM: SUPPRESSION OF THE DIRICHLET EIGENVALUES OF A COATED BODY*

STEVE ROSENCRANS† AND XUEFENG WANG†

Abstract. In our paper [*SIAM J. Appl. Math.*, 66 (2006), pp. 1895–1916] there are several mistakes in signs in the statements of Theorems 3 and 6.

Key words. nanocomposite, Dirichlet eigenvalue, anisotropic heat conduction, thermal tensor, thermal management, insulation, reinforcement

AMS subject classifications. 35J05, 35J20, 80A20, 80M30, 80M40

DOI. 10.1137/070705404

In the upper bounds on $\lambda_1(\Omega)$ stated in Theorems 3 and 6, there should be "$-$" signs in front of π and $\overline{H}$.

On page 1905 all "$-$" signs in front of the principal curvatures k_1 and k_2 should be changed to "$+$" signs. This leads to some obvious minor changes in the rest of the proof of Theorem 3 and the statement and proof of Lemma 5. The speed of the curves of principal curvature are 1 only at q, and so the six equations following "speed equal to 1" hold only at q.

These corrections necessitate *only* the above-mentioned changes in the statements of Theorems 3 and 6. The statements of all the other theorems remain unchanged.

*Received by the editors October 15, 2007; accepted for publication October 22, 2007; published electronically March 28, 2008.

http://www.siam.org/journals/siap/68-4/70540.html

†Mathematics Department, Tulane University, New Orleans, LA 70118 (srosenc@tulane.edu, xdw@math.tulane.edu).

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 1917–1948

RETURN MAP CHARACTERIZATIONS FOR A MODEL OF BURSTING WITH TWO SLOW VARIABLES*

ROGER E. GRIFFITHS† AND MARK PERNAROWSKI‡

Abstract. Various physiological systems display bursting electrical activity (BEA). There exist numerous three-variable models to describe this behavior. However, higher-dimensional models with two slow processes have recently been used to explain qualitative features of the BEA of some experimentally observed systems [T. Chay and D. Cook, *Math. Biosci.*, 90 (1988), pp. 139–153; P. Smolen and J. Keizer, *J. Memb. Biol.*, 127 (1992), pp. 9–19; R. Bertram et al., *Biophys. J.*, 79 (2000), pp. 2880–2892; R. Bertram et al.,, *Biophys. J.*, 68 (1995), pp. 2323–2332; J. Keizer and P. Smolen, *Proc. Nat. Acad. Sci. USA*, 88 (1991), pp. 3897–3901]. In this paper we present a model with two slow and two fast variables. For some parameter values the system has stable equilibria, while for other values there exist bursting solutions. Singular perturbation methods are used to define a one-dimensional return map, wherein fixed points correspond to singular bursting solutions. We analytically demonstrate that bursting solutions may exist even with a combination of activating and inactivating slow processes. We also demonstrate that for different parameters, bursting solutions may coexist with stable equilibria. Hence small variations in the initial conditions may drastically affect the dynamics.

Key words. bursting, return map, singular perturbation solutions

AMS subject classifications. 34A, 34C15, 34C29, 34D15, 34E15

DOI. 10.1137/050635201

1. Introduction.

Bursting electrical activity (BEA) is a phenomenon in which the membrane potential of a cell goes through a succession of alternating active (spiking) and silent states (cf. Figure 1.1). Such patterns of electrical activity were first observed experimentally in the electrical activity of the Aplysia R-15 neuron [52, 1]. Biophysical mechanisms of bursting in the pancreatic β-cell were proposed by Atwater et al. [2], which were later used by Chay and Keizer [15] to create the first "minimal" mathematical model based on the Hodgkin–Huxley model. Since then, there have been a large number of β-cell models [14, 27, 31, 46, 32, 13, 49] and other cellular models exhibiting bursting behavior [18, 26, 56, 55, 8, 33].

Most mathematical models of BEA are variants of the Hodgkin–Huxley model [29] of the squid giant axon. Generically, these models make up a set of (dimensional) differential equations,

$$C_m \frac{dv}{dt} = -\sum_X I_X(v,z), \tag{1.1}$$

$$\frac{dz_i}{dt} = \frac{(z_{i\infty}(v) - z_i)}{\tau_i(v)}, \quad i = 1, 2, \ldots, n, \tag{1.2}$$

where t is time, v is the transmembrane potential, and z_i are typically channel activation (resp., inactivation) variables. In some instances, z_i may be concentrations of regulatory chemicals. Regardless, all such models have a current balance equation

*Received by the editors July 5, 2005; accepted for publication (in revised form) May 10, 2006; published electronically September 12, 2006.
http://www.siam.org/journals/siap/66-6/63520.html
†Department of Mathematics, Mercyhurst College, Erie, PA 16546 (rgriffiths@mercyhurst.edu).
‡Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717 (pernarow@math.montana.edu).

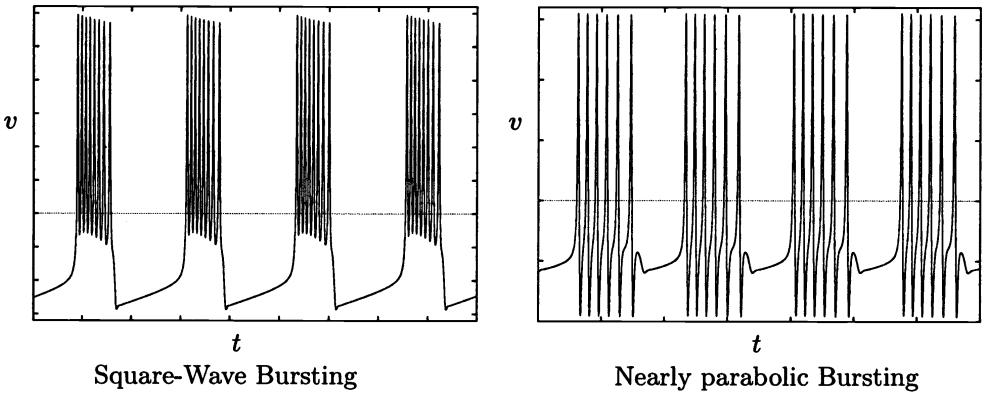


FIG. 1.1. *Some examples of bursting, with their classification type indicated. Here voltage v is plotted against time t.*

such as (1.1), where C_m is the cell's total capacitance and I_X are currents (of type X, i.e., voltage-gated calcium) thought to be relevant to the particular cell being examined. In models of BEA, the time constants τ_i often have greatly different magnitudes. Thus, from a modeling (resp., mathematical) perspective, bursting depends on processes with distinctly different time scales, typically termed *fast* and *slow*. The fast processes remain in quasi-steady state except for the rapid transitions between states, while the slow processes modulate the fast dynamics between the silent and active states.

The nonlinearities essential in biophysical models of bursting such as (1.1)–(1.2) make any form of analysis difficult. Consequently, phenomenological models have often been used to explore issues related to bursting (see Hindmarsh and Rose [28], Pernarowski [36], Baer, Rinzel, and Carrillo [3]). Regardless of what type of model is studied, the vast majority of models involve multiple time scales and can be written as a system of the form

$$\frac{dx}{dt} = f(x,y)\,, \quad x \in \mathbb{R}^2, \tag{1.3}$$

$$\frac{dy}{dt} = \varepsilon g(x,y)\,, \quad y \in \mathbb{R}^K, \tag{1.4}$$

where $\varepsilon \ll 1$ is a small parameter. Cast in this form, x are fast variables, while y are slow variables.

Many models of BEA consisting of one slow variable ($K = 1$) have been studied. The goals of such studies have varied. Some studies use numerical methods to simulate postulated models to explain cell mechanisms. Such is the case in numerous studies of the insulin-secreting pancreatic β-cell [15, 12, 46, 45, 43, 17, 44]. In other studies, the goal has been to use singular perturbation ideas and methods to classify the different type of oscillations which can arise from such systems. Such classification studies originated with work by Rinzel [41] and were subsequently continued by others [36, 4, 16, 30]. Lastly, other studies have focused on proving the existence of periodic bursting orbits. For instance, Terman [53] formalized the singular perturbation construction. Interest in such fast-slow systems with one slow variable has even spawned studies of

topological-based proof techniques [24].

Other models of BEA have more than one slow variable [14, 49, 5, 32, 6, 40]. In some recent studies, bursting cycles have been characterized using two-dimensional maps to aid in the construction of singular solutions [51, 9]. In other recent works, one-dimensional maps have been used to explore bifurcations in systems with one slow variable [34]. In all of these works, periodic bursting cycles equate to fixed points of the map, making the use of the map construct simplistically elegant. Despite such recent uses of maps to describe bursting cycles, explicit singular perturbation and numerical constructions of the maps for models exhibiting bursting remain scant.

In this paper we study a phenomenological model of bursting with two slow variables. As in previous works, singular perturbation methods are used here to define a return map to describe the bursting cycle. Unlike previous analyses of models having two slow and two fast variables, our return map is one-dimensional. Also, because the model is simple, many of the perturbation calculations can be performed analytically. The goal of this work is three-fold: (1) to outline new analytical and numerical techniques for such singular constructions; (2) to analytically demonstrate that bursting solutions can exist even when the slow processes are activating and inactivating; and (3) to demonstrate that bursting solutions can coexist with stable equilibria. The latter result, for example, can be used to explain why some isolated pancreatic β-cells burst while others do not [50]. The implications and importance of these results are discussed in the conclusion.

In section 2, the model is introduced and its leading-order fast, slow, and averaged fast subsystems are defined. Since the model is an extension of a previously studied model [36, 37, 16], stability analyses related to the fast subsystem are referenced. A detailed multiple scales averaging calculation used to derive the averaged fast subsystem in section 2 is included in an appendix. This derivation is a multivariable generalization of the single slow variable derivation presented in [39].

As each of the slow and averaged fast subsystems is two-dimensional, each defines a two-dimensional map between the transition curves[1] between the silent and active phases. These maps and their dimensionality reduction are carefully defined in section 3. Singular bursting solutions are determined by the fixed points of the composition ϕ of these maps.

When the time constants $\tau_i, i = 1, 2$, of the slow variables are equal, a transformation is used in section 4 to demonstrate that the model dynamics can be described by a single slow variable. As a consequence, singular bursting solutions exist even if one slow variable is activating and the other is inactivating. Furthermore, we show that the map ϕ defined in section 3 can be computed explicitly. However, when the time constants τ_i are not equal, these analyses do not apply. For this case, a numerical method for computing ϕ is developed and implemented in section 5. The method uses AUTO [19] to solve two one-parameter families of boundary value problems whose solutions can then be used to construct ϕ.

Lastly, in section 6, we present numerical simulations which demonstrate that for certain parameter values the model exhibits bistability between stable equilibria and bursting solutions. There, this dynamic is shown to relate to the domains of the maps used to construct the singular bursting solutions.

[1] Saddle-node and homoclinic bifurcation curves of the fast subsystem.

2. Model and subsystem definitions. In this paper we will study the following two slow variable models:

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u},\mathbf{z}) = \begin{pmatrix} f(u) - w - x - \gamma y \\ g(u) - w \end{pmatrix}, \tag{2.1}$$

$$\frac{d\mathbf{z}}{dt} = \varepsilon \mathbf{G}(\mathbf{u},\mathbf{z}) = \varepsilon \begin{pmatrix} \frac{h_1(u)-x}{\tau_1} \\ \frac{h_2(u)-y}{\tau_2} \end{pmatrix}, \tag{2.2}$$

where $0 < \varepsilon \ll 1$, $\mathbf{u} = (u,w)^T$ are fast variables and $\mathbf{z} = (x,y)^T$ are slow variables.

In the model, τ_1, τ_2, and γ are constants; the equations

$$f(u) = -\frac{a}{3}u^3 + a\mu u^2 + (1 - a(\mu^2 - \eta^2))u\,, \tag{2.3}$$

$$g(u) = \left(1 - \frac{a}{3}\right)u^3 + a\mu u^2 - (2 + a(\mu^2 - \eta^2))u - 3 \tag{2.4}$$

are the same functions used in [36, 37], and (a,μ,η) are parameters. The complicated form of the polynomials f and g is due in part to their derivation from a Liénard form in Pernarowski [37]. An advantage of the Liénard form is the availability of the Melnikov theory to analytically approximate homoclinic bifurcation points [39, 36]. Also, as shall be seen in the next section, the location of the fast subsystem equilibria does not depend on (a,μ,η). We mention these facts here for reference purposes only, and will not need them in subsequent analysis.

Finally,

$$h_i(u) = \beta_i(u - \alpha_i), \quad i = 1,2, \tag{2.5}$$

where α_i and β_i are also constants.

In this model u should be interpreted as the membrane potential, whereas w is a fast conductance and x, y are slow conductances for gating channels of the same ion. Hereafter we shall refer to (2.1)–(2.2) as (FULL). Lastly, we note that throughout this paper the notation $\dot{(\,)}$ will be used to denote differentiation in t, as in $\dot{u} = \frac{du}{dt}$.

In the next few sections the fast subsystem (FS), slow subsystem (SS) and averaged fast subsystem (AFS) associated with (FULL) will be defined. These preliminary definitions will be needed to accurately define the return map in section 3.

2.1. (FS) dynamics. On the fast t time scale the dynamics of (FULL) is governed by the (FS) obtained by letting $\varepsilon = 0$,

$$\frac{du}{dt} = f(u) - w - z\,, \quad z \equiv x + \gamma y\,, \tag{2.6}$$

$$\frac{dw}{dt} = g(u) - w\,, \tag{2.7}$$

with the slow variables x and y treated as parameters and combined as shown above. In Figure 2.1 we show a numerically generated (FS) bifurcation diagram in $z = x + \gamma y$. The projection of the equilibria of (FS) onto the (u,z)-plane yields a Z-shaped curve

$$z = G(u) \equiv f(u) - g(u) = -u^3 + 3u + 3. \tag{2.8}$$

Note here that despite the dependence of f and g on the fast parameter set $\lambda_f = (a,\mu,\eta)$, G depends on no parameters. For the remainder of this paper we fix these

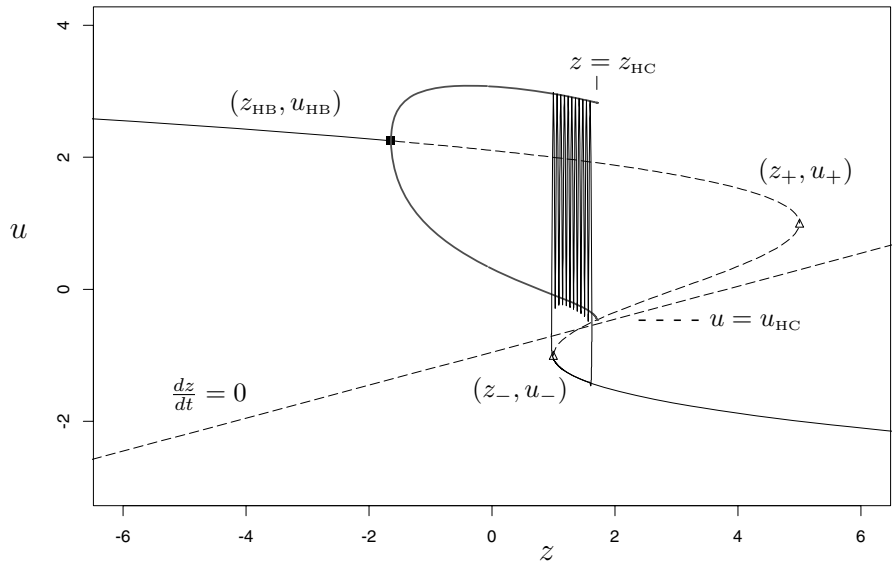


FIG. 2.1. *(FS) bifurcation diagram for* (2.6)–(2.7) *when* $\lambda_f = (a, \eta, \mu) = \left(\frac{1}{4}, \frac{3}{4}, \frac{3}{2}\right)$ *with a bursting solution superimposed. Other parameter values for this illustration are* $\beta = 4, \alpha = -0.954, \varepsilon = 0.0025$.

fast-parameter values at $(a, \mu, \eta) = \left(\frac{1}{4}, \frac{3}{4}, \frac{3}{2}\right)$; for details on fast parameter selection, see [36].

In Figure 2.1, solid lines on the $z = G(u)$ equilibria curve indicate stable equilibria, whereas the dashed portion indicates unstable equilibria. Equilibria on the lower branch are stable nodes, whereas equilibria on the middle branch are saddle points. The stability of the steady states on the upper branch changes at a supercritical Hopf bifurcation at $z = z_{\rm HB}$. Though Figure 2.1 was computed numerically using XPPAUT [21], the aforementioned stabilities and bifurcations were proven analytically in [36]. Stable periodic orbits (the dark, thick lines) emanate from the Hopf point and terminate at a homoclinic bifurcation on the middle branch at $z = z_{\rm HC}$. The upper and lower portions indicate the extreme values of u on the limit cycles of the (FS). Saddle-node bifurcations are indicated at $z = z_-$ and $z = z_+$. Note that the (FS) has a region of bistability, where stable lower branch equilibria and periodic orbits coexist for $z \in (z_-, z_{\rm HC})$. In later sections we will make reference to the values $u_+, u_-, u_{\rm HC}$ as the u value at which the upper saddle-node, lower saddle-node, and homoclinic bifurcations occur, respectively. Also, as indicated in Figure 2.1, when we refer to $u_{\rm HC}$ we mean the u value on the middle branch when $z = z_{\rm HC}$.

The (FS) bifurcation diagram in z described above is identical to that of the one slow variable model discussed in [37]. In that model, z evolves according to the differential equation

$$\frac{dz}{dt} = \varepsilon(h(u) - z), \tag{2.9}$$

where $h(u) = \beta(u - \alpha)$; α and β are parameters; and $\beta > 0$. Collectively, (2.6), (2.7), and (2.9) define the one slow variable model in [37]. Before we examine bursting solutions of (FULL), we briefly describe a bursting cycle of the aforementioned one slow variable model for comparison. Toward this end, the z (slow) nullcline associated with (2.9) is superimposed on the (FS) bifurcation diagram in Figure 2.1 (dashed line

passing through the middle branch of the Z curve). Here we have used the slow parameter values $(\alpha, \beta) = (-0.954, 4)$.

In this example with $\beta > 0$, $\dot{z}$ is negative only below the nullcline. Thus, z slowly increases above the nullcline and decreases below it. Keeping this in mind, a bursting solution of the one slow variable model (2.6), (2.7), (2.9) is superimposed on the (FS) bifurcation diagram in Figure 2.1. In what is often referred to as the "silent phase," trajectories lie close to the lower branch of equilibria. Since $\dot{z} < 0$ on the lower branch, trajectories move to the left until bistability is lost at the saddle-node bifurcation point (z_-, u_-). As z decreases below z_-, trajectories are then attracted to the (FS) limit cycles, initiating the "active phase" marked by high-frequency oscillations. In the active phase, $\dot{z} > 0$ so that solutions slowly drift to the right until bistability is lost at the homoclinic bifurcation point at $z = z_{\text{HC}}$. Trajectories are then attracted to the lower branch, initiating the silent phase again.

This explanation of the resulting "square wave" bursting cycle was given by Rinzel in [41], where he classified several other types of bursting cycles depending on their fast (and slow) subsystem structure. In a later classification scheme [4], the cycle depicted in Figure 2.1 is known as type I bursting. In a subsequent and more extensive classification scheme [30], the same cycle is described as a "fold/homoclinic burster"; its name is due to the fact that the silent phase ends via a *fold* bifurcation and the active phase terminates via a saddle *homoclinic* orbit bifurcation.

Though the (FS) of (FULL) is identical to that of the one slow variable model just discussed, in (FULL) the slow variables x and y do not evolve according to (2.9) but by (2.2). Instead, silent phase trajectories of (FULL) are attracted to the stable manifold

$$\mathbb{S}_L = \{(u, w, x, y) : x + \gamma y = G(u), w = g(u), u < u_-\}, \tag{2.10}$$

formed by the lower branch equilibria of the (FS). In contrast to the single slow variable model, this is a two-dimensional manifold in $\mathbb{R}^4$, making the preceding explanation of the bursting cycle using (2.9) inapplicable. Later, we shall make use of the fact that bursting cycles of (FULL) are conveniently described by projecting them onto the (x, y)-plane. Toward this end, we define the projection $P(\mathbb{S}_L)$ of $\mathbb{S}_L$ onto the (x, y)-plane as

$$S_L = \{(x, y) : x + \gamma y = G(u), u < u_-\}. \tag{2.11}$$

The reason we distinguish S_L from $\mathbb{S}_L$ is that later it will become easier to visualize trajectories on S_L rather than on $\mathbb{S}_L$. We note, for instance, that equilibria of (FULL) truly exist only on $\mathbb{S}_L$.

2.2. (SS) dynamics. In this section we define the (SS) of (FULL). Using the (slow time) transformation $\tau = \varepsilon t$ in (FULL) and then setting $\varepsilon = 0$ results in the system

$$z = x + \gamma y = G(u), \tag{2.12}$$

$$w = g(u), \tag{2.13}$$

$$\frac{dx}{d\tau} = \frac{h_1(u) - x}{\tau_1}, \tag{2.14}$$

$$\frac{dy}{d\tau} = \frac{h_2(u) - y}{\tau_2}. \tag{2.15}$$

Collectively, (2.12)–(2.15) define the (SS) of (FULL) as long as $u < u_-$. Solutions of (2.12)–(2.15) yield trajectories on $\mathbb{S}_L$ which are leading-order approximations to the silent phase of (FULL).

Equations (2.12)–(2.13) are algebraic conditions which ensure the (SS) flow remains on $\mathbb{S}_L$. Rewritten, this condition is equivalent to the cubic

$$u^3 - 3u - 3 + z = 0. \tag{2.16}$$

Roots of (2.16) can be explicitly computed. Here we use the trigonometric form of these roots discussed in [7]. Since the root associated with $\mathbb{S}_L$ has $u < u_-$, it is readily verified that on the lower branch of the (FS),

$$u = u_{\text{LB}}(z) = \begin{cases} 2\cos\left(\frac{1}{3}\arccos\left(\frac{3-z}{2}\right) + \frac{2\pi}{3}\right), & 1 \le z \le 5, \\ -2\cosh\left(\frac{1}{3}\ln\left(\frac{z-3}{2} + \sqrt{\left(\frac{z-3}{2}\right)^2 - 1}\right)\right), & z > 5. \end{cases} \tag{2.17}$$

This allows one to rewrite (2.14)–(2.15) as

$$\frac{dx}{d\tau} = F_1(x,y) \equiv \frac{\beta_1(u_{\text{LB}}(z) - \alpha_1) - x}{\tau_1}, \tag{2.18}$$

$$\frac{dy}{d\tau} = F_2(x,y) \equiv \frac{\beta_2(u_{\text{LB}}(z) - \alpha_2) - y}{\tau_2}. \tag{2.19}$$

Solutions of (2.18)–(2.19) are leading-order silent phase approximations of (FULL) projected onto the (x,y)-plane when initial conditions are close to $\mathbb{S}_L$. In section 3, system (2.18)–(2.19) will be used to define the portion of the map needed to describe the dynamics of (FULL) in the silent phase.

At this point we also note that the (FS) and the (SS) depend on different parameters. For future reference, we define the *fast* parameter set λ_f as those parameters which occur explicitly in the (FS) but not in the (SS). In this case, $\lambda_f = (a, \eta, \mu)$. In a similar fashion, we define the *slow* parameter set λ_s as those parameters which occur explicitly in the (SS) but not in the (FS). For (FULL), $\lambda_s = (\beta_1, \beta_2, \alpha_1, \alpha_2, \tau_1, \tau_2)$.

2.3. (AFS). In this section we define the (AFS) associated with (FULL). Like the (SS), the (AFS) is a leading-order approximation for the x and y components of (FULL) valid for slow times $\tau = O(1)$. In contrast to the (SS), the (AFS) is an approximation valid only for initial conditions near the (FS) limit cycles.[2] The details of the expansions, assumptions, and multiple scales procedure used to derive the (AFS) are included in the appendix. Here we merely summarize the relevant points.

Defining the set

$$S_A = \{(x,y) : z_{HB} < x + \gamma y < z_{\text{HC}}\}, \tag{2.20}$$

we note from Figure 2.1 that the (FS) has a stable $T(z)$-periodic limit cycle $(u,w) = \Omega(t,z)$ for every $(x,y) \in S_A$. Here the limit cycle $\Omega(t,z)$ and its associated period T

[2]Since the averaging is performed over these (FS) limit cycles, we named the system the "averaged fast subsystem." This naming scheme was chosen to avoid confusion were the averaging performed over periodic orbits lying near the slow manifold of the (SS). Indeed, in [25], the (SS) is shown to possess Hopf bifurcations for some parameter values. An exploration of averaged slow subsystems near such Hopf points is not done in this paper.

are functions of $z = x + \gamma y$ alone. If we define the average

$$\hat{u}(z) \equiv \frac{1}{T(z)} \int_0^{T(z)} \Omega_1(\eta, z) d\eta, \tag{2.21}$$

where $\Omega(t,z) = (\Omega_1(t,z), \Omega_2(t,z))$, and take advantage of the linearity of $h_i(u)$, $i = 1, 2$, in u, the (AFS) of (FULL) is

$$\frac{dx}{d\tau} = \widehat{G}_1(x,y) = \frac{h_1(\hat{u}(x+\gamma y)) - x}{\tau_1}\ , \tag{2.22}$$

$$\frac{dy}{d\tau} = \widehat{G}_2(x,y) = \frac{h_2(\hat{u}(x+\gamma y)) - y}{\tau_2}\ . \tag{2.23}$$

For initial conditions sufficiently close to Ω, solutions of (2.22)–(2.23) are leading-order asymptotic approximations of (x,y) in (FULL) for times $\tau = O(1)$ as long as $(x(\tau), y(\tau)) \in S_A$. In the singular limit, bistability of the (FS) is lost when $z = z_{\rm HC}$, at which point a transition to the silent phase occurs.

Subsequently, we will need to perform computations using the (AFS). For these calculations we used AUTO [19] to compute $\hat{u}(z)$ for a range of z values. These values are shown in Figure 2.2, where they are compared to an approximating function $\hat{u}_a(z)$ of the form

$$\widehat{u}_a(z) = u_{\rm HC} + a_0 \left(z_{\rm HC} - z\right)^{\frac{1}{p}} + a_1 \left(z_{\rm HC} - z\right). \tag{2.24}$$

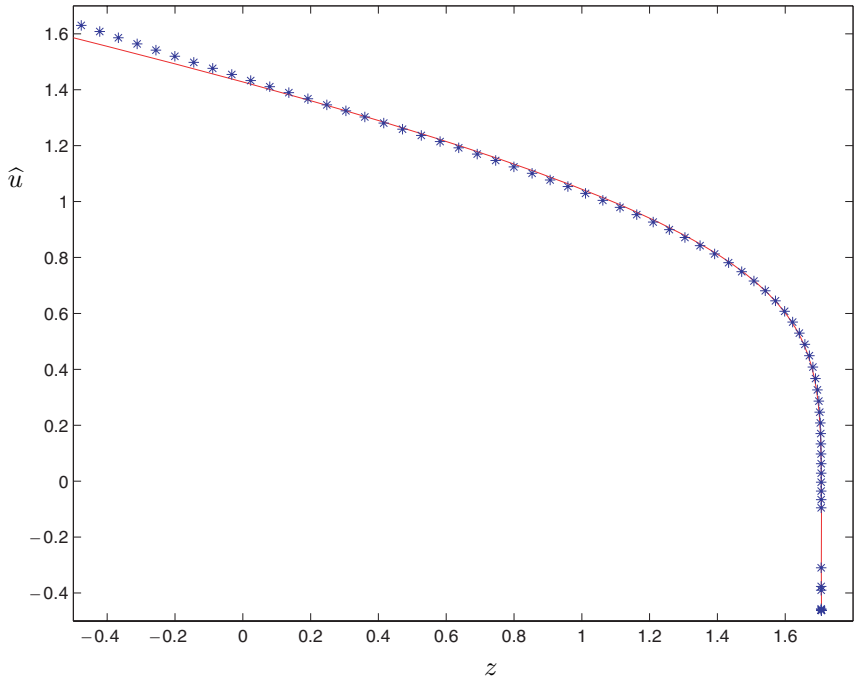


FIG. 2.2. *$\widehat{u}(z)$ plotted as AUTO-generated data along with the approximating function $\widehat{u}_a(z)$. $\widehat{u}(z)$ is u averaged over the limit cycles $\Omega(t,z)$. The AUTO-generated data are plotted as *'s and the superimposed curve is the approximating function $\widehat{u}_a(z)$ defined in* (2.24).

As can be seen in Figure 2.2, for the choice $(p, a_0, a_1) = (8, 1.378, 0.260)$, $\hat{u}(z)$ and $\hat{u}_a(z)$ are almost indistinguishable over the active phase range $z_- < z < z_{\rm HC}$. Here $z_- = 1$ and $(z_{\rm HC}, u_{\rm HC}) \simeq (1.70633, -0.46466)$ were estimated using AUTO. We note that (2.24) may be viewed as a two term asymptotic expansion for $\hat{u}(z)$. However, we currently have no theoretical explanation as to why $\hat{u}_a(z)$ approximates $\hat{u}(z)$ so well.

We conclude this section with a numerical example showing how well the (AFS) approximates (FULL) in the active phase. In Figure 2.3, an active phase trajectory of (FULL) projected into the (x, y)-plane is shown together with a solution of the (AFS) with the same initial conditions. Superimposed is the line $x + \gamma y = z_{\rm HC}$, labeled as $\Gamma_{\rm HC}$ in the figures. On this line the (FS) has a homoclinic bifurcation. For the

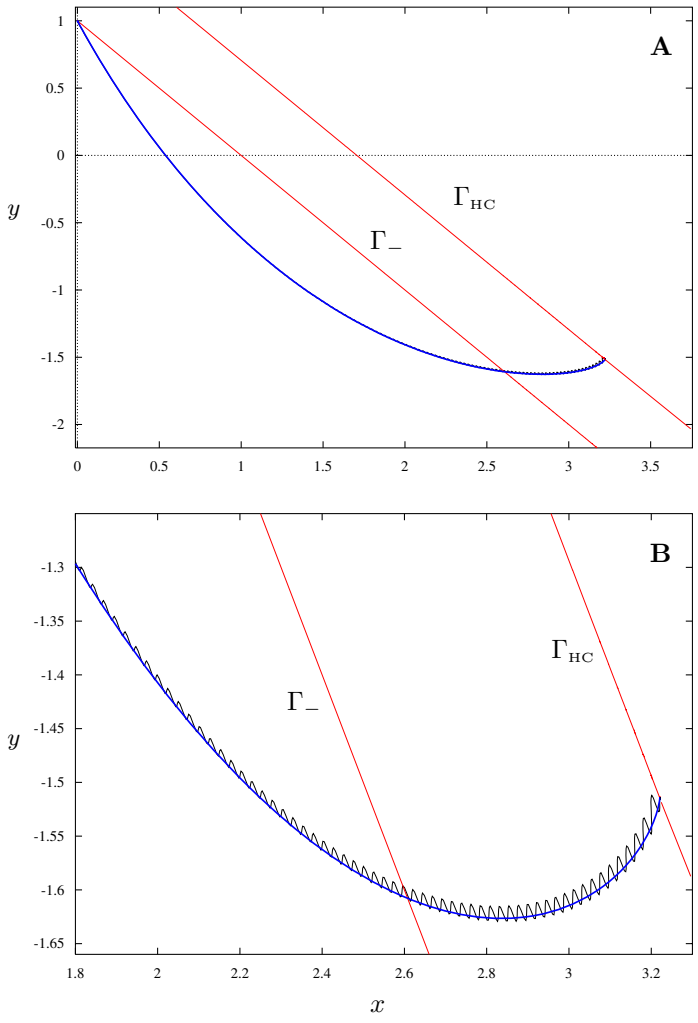


FIG. 2.3. *Both figures show an active phase trajectory of (FULL) projected into the (x, y)-plane along with the (AFS) approximation using* (2.24), $\lambda_s = (\beta_1, \beta_2, \alpha_1, \alpha_2, \tau_1, \tau_2) = (4, -1, -1, -0.7, 1, 0.5)$, *and* $\varepsilon = 0.0025$. **A** *shows the entire active phase, whereas in* **B** *we have enlarged the region near the jump to the silent phase at the homoclinic bifurcation of the (FS). The lines Γ_{HC} and Γ_- are defined in* (3.1), (3.2), *respectively.*

(x, y) values above Γ_{HC} the (FS) does not have periodic orbits. Thus, for those values the (AFS) is undefined. Moreover, to leading order, as trajectories of (FULL) cross above Γ_{HC}, a rapid transition back to the silent phase occurs. Analogous transitions of (FULL) from the silent phase into the active phase would occur as trajectories traverse below the line $x + \gamma y = z_-$. This line, labeled as Γ_-, is also superimposed in the figures purely for reference purposes and represents the (x, y) pairs for which the (FS) has a saddle-node bifurcation. A more complete discussion of these transition curves and their relevance to an overall bursting cycle is relegated to the next section, where the return maps for the leading dynamics are defined. Here, the point is simply to illustrate just how well the (AFS) using (2.24) approximates (FULL) in the active phase. In this particular example, parameter values were chosen so that the projected active phase solution of (FULL) extended over a wide range of $z = x + \gamma y$ for which the (FS) has limit cycles. The figures illustrate that the projected trajectory $(x(t), y(t))$ of (FULL) is very well approximated by the (AFS) trajectories over just such a large range of z values. Indeed, even in the enlargement shown in Figure 2.3B, the trajectories remain very close in the region between Γ_- and Γ_{HC}, where the (FS) is bistable.

3. Definition of the return map. We define a bursting solution of (FULL) as a trajectory which both is periodic and traverses the active and silent phases once during each period. By a singular bursting solution of (FULL) we mean a bursting solution whose leading-order approximation consists of a silent phase portion approximated by the (SS), an active phase portion approximated by the (AFS), and two rapid transitions between each phase at $z = z_-$ and $z = z_{\text{HC}}$, as illustrated in Figure 2.1.

The return map we construct to describe singular bursting solutions of (FULL) is actually the composition of two maps, as illustrated in Figure 3.1. One map accounts for the silent phase and the other for the active phase. The silent phase portion of the singular solution is constructed from a trajectory $\tilde{\gamma}_x(\tau)$ of the (SS) which starts at $\mathbf{X} = (x, y)$ on the curve

$$\Gamma_{\text{HC}} = \{(x, y) : x + \gamma y = z_{\text{HC}}\} \subset \overline{S}_A \tag{3.1}$$

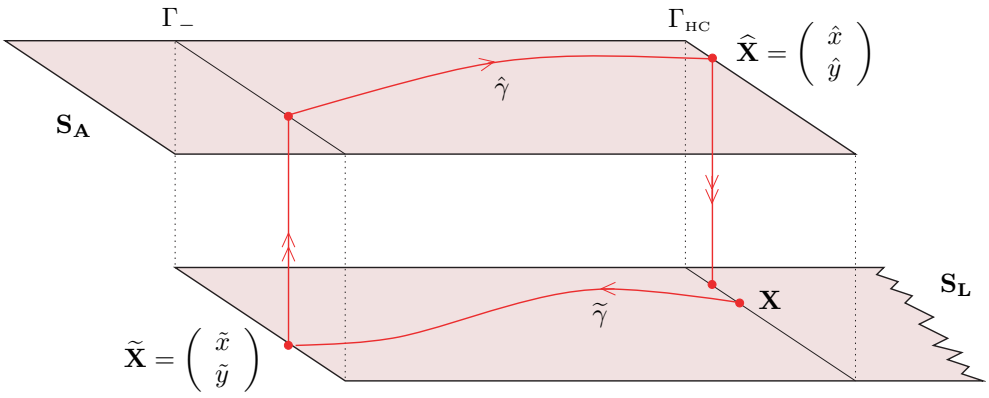


FIG. 3.1. *Illustration of the return map definition for singular bursting solutions. Illustration S_L is shown on bottom with a trajectory of the (SS) $\tilde{\gamma}$ beginning at $\mathbf{X} = (x, y) \in \Gamma_{HC}$. After a fast transition to S_A on the top, the (AFS) trajectory $\hat{\gamma}$ is shown to terminate at $\widehat{\mathbf{X}} \in \Gamma_{HC}$.*

and terminates at $\widetilde{\mathbf{X}}$ on

$$\Gamma_- = \{(x,y) : x + \gamma y = z_-\} \subset \overline{S}_A \tag{3.2}$$

after time T_s. When such transitions occur, this defines a map $\widetilde{\Phi} : \Gamma_{\text{HC}} \to \Gamma_-$ such that $\widetilde{\mathbf{X}} = \widetilde{\Phi}(\mathbf{X})$.

Similarly, the active phase portion is constructed from a trajectory $\hat{\gamma}_x(\tau)$ of the (AFS) which starts at $\widetilde{\mathbf{X}}$ on the lower saddle-node curve Γ_- and terminates at $\widehat{\mathbf{X}}$ on the curve Γ_{HC}, which forms a boundary of S_A. This transition defines a map $\widehat{\Phi} : \Gamma_- \to \Gamma_{\text{HC}}$ such that $\widehat{\mathbf{X}} = \widehat{\Phi}(\widetilde{\mathbf{X}})$.

The composition of these two maps may be written as $\Phi : \Gamma_{\text{HC}} \to \Gamma_{\text{HC}}$, $\mathbf{X} = (x,y)$,

$$\Phi(\mathbf{X}) = \left(\widehat{\Phi} \circ \widetilde{\Phi}\right)(\mathbf{X}). \tag{3.3}$$

Given these definitions, there is then a 1-1 correspondence between singular bursting solutions and fixed points of Φ.

At this point we also remark that $\widetilde{\Phi}$ and $\widehat{\Phi}$ are, technically, maps from $\mathbb{R}^2$ into $\mathbb{R}^2$, but they have restricted domains, i.e., $\Gamma_{\text{HC}} \subset \mathbb{R}^2$ for $\widetilde{\Phi}$. Clearly such domain restrictions are not needed. For instance, one could have defined $\widehat{\Phi}$ as the flow function associated with the (SS) defined on $D_\Phi = \{(x,y) : x + \gamma y > z_-\}$. Although such a definition has much theoretical appeal, our subsequent dimensionality reduction below is slightly more transparent using the domain restrictions.

A feature we use to our advantage is that y is a function of x on both Γ_- and Γ_{HC}—specifically, on Γ_-: $y = \frac{z_- - x}{\gamma}$ and on Γ_{HC}: $y = \frac{z_{\text{HC}} - x}{\gamma}$. As such, $\widehat{\Phi}$ reduces to a one-dimensional map $\widehat{\phi} : \mathbb{R} \to \mathbb{R}$, $\widehat{\phi}(x_0) = \hat{x}(T_a)$, where T_a is the active phase duration. Here $\widehat{x}_0 = \widehat{\phi}(x_0)$ is the x-coordinate of an (AFS) trajectory, where bistability is lost at $z = z_{\text{HC}}$ and a transition to the silent phase occurs. Adopting this convention, the domain of the map $\widehat{\phi}$ can be written

$$D(\widehat{\phi}) = \{x \in \mathbb{R} \,|\, \exists T_a < \infty \ni \widehat{\gamma}_x(T_a) \cap \Gamma_{\text{HC}} \neq \emptyset\},$$

where $\emptyset$ is the empty set.

An analogous one-dimensional map for the (SS) may be defined. Again exploiting the functional relationship of x and y on Γ_{HC} and Γ_-, we define the map $\widetilde{\phi} : \mathbb{R} \to \mathbb{R}$ with domain

$$D(\widetilde{\phi}) = \{x \in \mathbb{R} \,|\, \exists T_s < \infty \ni \widetilde{\gamma}_x(T_s) \cap \Gamma_- \neq \emptyset\},$$

where T_s is the silent phase duration.

Then, as with the map Φ, there is a 1-1 correspondence of singular bursting solutions and fixed points $\bar{x}$ of the map

$$\phi(x) = (\widehat{\phi} \circ \widetilde{\phi})(x), \quad x \in D(\phi) = \{x \in \mathbb{R} \,|\, x \in D(\widetilde{\phi})\,,\ \widetilde{\phi}(x) \in D(\widehat{\phi})\}.$$

Domain issues associated with the map are complicated, but some will be addressed in section 6. If $R(\widehat{\phi})$ is the range of $\widehat{\phi}$, then, given our previous definitions, the set $W \equiv R(\widehat{\phi}) \cap D(\widetilde{\phi})$ could be empty, equal to $D(\widetilde{\phi})$, or a (nonempty) strict subset of $D(\widetilde{\phi})$. In the next section we show that for a subset of slow parameter space, singular bursting solutions exist and W is nonempty. The case when W is a strict subset of $D(\widetilde{\phi})$ relates to bistability between bursting solutions and stable equilibria on $\mathbb{S}_L$. This will be discussed in section 6.

4. Degenerate case. In this section we examine a degenerate case of (FULL), where the time constants τ_1 and τ_2 are equal. In this degenerate case, (FULL) is shown to have a three-dimensional manifold on which (for certain parameter values) bursting solutions exist. Furthermore, we explicitly compute the maps defined in the previous section and the fixed point of ϕ associated with the singular bursting solution.

We accomplish these goals using the following linear transformation of the slow variables:

$$\mathbf{p} = A\mathbf{z} + \mathbf{b} = \begin{bmatrix} a_{11} & a_{12} \\ 1 & \gamma \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} b_1 \\ 0 \end{pmatrix}, \quad \mathbf{z} = (x, y)^T, \tag{4.1}$$

where $\mathbf{p} = (p, z)^T$ are the new slow variables, $z = x + \gamma y$ are as before, and a_{11}, a_{12}, b_1 are to be determined. Since z is one of the new variables, the new (FS) of the transformed (FULL) will depend solely on z. Assuming $a_{11}\gamma - a_{12} \neq 0$ so that A is invertible, (4.1) and (2.2) imply

$$\frac{d\mathbf{p}}{dt} = \varepsilon A\mathbf{G}\left(u, A^{-1}(\mathbf{p} - \mathbf{b})\right). \tag{4.2}$$

Since $\mathbf{G}(u, \mathbf{z})$ depends linearly on u, x, and y, (4.2) can be equivalently written as

$$\frac{d\mathbf{p}}{dt} = \begin{bmatrix} \eta_{11} & \eta_{12} & \eta_{13} \\ \eta_{21} & \eta_{22} & \eta_{23} \end{bmatrix} \begin{pmatrix} u \\ z \\ p \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \tag{4.3}$$

for appropriate definitions of η_{ij}, ζ_i. To ensure that the new differential equation for z does not depend explicitly on p, we require that $\eta_{23} = 0$. Since calculations reveal

$$\eta_{23} = \frac{\gamma(\tau_1 - \tau_2)}{\tau_1\tau_2 \; detA},$$

the new slow variable p will be decoupled from the resulting (u, w, z) system only if $\tau_1 = \tau_2$ (which is precisely how we defined the degenerate case).

By assuming $\tau_1 = \tau_2$, the choice

$$a_{11} = 1\,, \quad a_{12} = -\frac{\beta_1\tau_2}{\beta_2\tau_1}\,, \quad b_1 = \beta_1(\alpha_1 - \alpha_2) + 1 \tag{4.4}$$

then results in the transformed system

$$\frac{du}{dt} = f(u) - w - z, \tag{4.5}$$

$$\frac{dw}{dt} = g(u) - w, \tag{4.6}$$

$$\frac{dz}{dt} = \varepsilon\frac{H(u) - z}{\tau_1}, \tag{4.7}$$

$$\frac{dp}{dt} = \varepsilon\frac{1 - p}{\tau_1}, \tag{4.8}$$

where

$$H(u) = \beta^*(u - \alpha^*), \tag{4.9}$$

$$\beta^* = \beta_1 + \gamma\beta_2, \tag{4.10}$$

$$\alpha^* = \frac{\beta_1\alpha_1 + \gamma\beta_2\alpha_2}{\beta_1 + \gamma\beta_2}. \tag{4.11}$$

As claimed, we see that the variable p is decoupled from the rest of the system (4.5)–(4.7), showing the reduction of (FULL) to a one slow variable model (4.5)–(4.7). Furthermore, we see in (4.8) that as $t \to \infty$, $p \to 1$. In other words, $p = 1$ is a globally stable three-dimensional manifold on which the dynamics are determined by (4.5)–(4.7). Given (4.1) and (4.4), this implies that the projected trajectories of (FULL) are attracted to the line

$$y = \frac{\beta_2}{\beta_1}x + \beta_2(\alpha_1 - \alpha_2) \tag{4.12}$$

in the (x, y)-plane.

By comparing (4.7) to (2.9) and making the identification $(\beta, \alpha) = (\beta^*, \alpha^*)$, it is evident that on $p = 1$ there exist (β^*, α^*) with $\beta^* > 0$ such that (FULL) has bursting solutions. This assures us, in this degenerate case, that for some subset of slow parameter space there must exist a fixed point $\bar{x}$ for the return map $\phi(x)$.

We illustrate one such bursting solution and fixed point in Figure 4.1. There, XPPAUT [21] was used to numerically integrate (FULL) for the degenerate case $\lambda_s = (\beta_1, \beta_2, \alpha_1, \alpha_2, \tau_1, \tau_2) = (3, 0.5, -1, -3, 1, 1)$ with $\gamma = 0.7$. In Figure 4.1A we see square-wave bursting in the u versus t time trace for this run. In Figure 4.1B we see how the projected trajectory is attracted to the line $y = \frac{\beta_2}{\beta_1}x + \beta_2(\alpha_1 - \alpha_2)$, indicated as $p = 1$. Also superimposed in Figure 4.1B are the two curves Γ_- and Γ_{HC}. Lastly, we note that for this run, since $\beta_1 > 0$ and $\beta_2 > 0$, both x and y are activation variables. Given (4.10), it is clear that the signs of β_i need not be the same for β^* to be positive. Thus, in the degenerate case, combinations of activating and inactivating variables can result in bursting solutions.

4.1. Explicit construction of ϕ in the degenerate case. In this section we find an explicit formula for the map ϕ in the degenerate case. We do this by separately deriving $\widetilde{\phi}$ and $\widehat{\phi}$, after which a direct composition yields ϕ.

First, we determine the silent phase duration T_s. Differentiating (2.12) and using (4.7),

$$\frac{dz}{d\tau} = G'(u)\frac{du}{d\tau} = \frac{H(u) - G(u)}{\tau_1}.$$

Integrating this result,

$$\int_{u_{\text{HC}}}^{u_-} \frac{G'(u)}{H(u) - G(u)}\, du = \int_0^{T_s} \frac{1}{\tau_1}\, d\tau,$$

one may solve for

$$T_s = \tau_1 \int_{u_{\text{HC}}}^{u_-} \frac{G'(u)}{H(u) - G(u)}\, du. \tag{4.13}$$

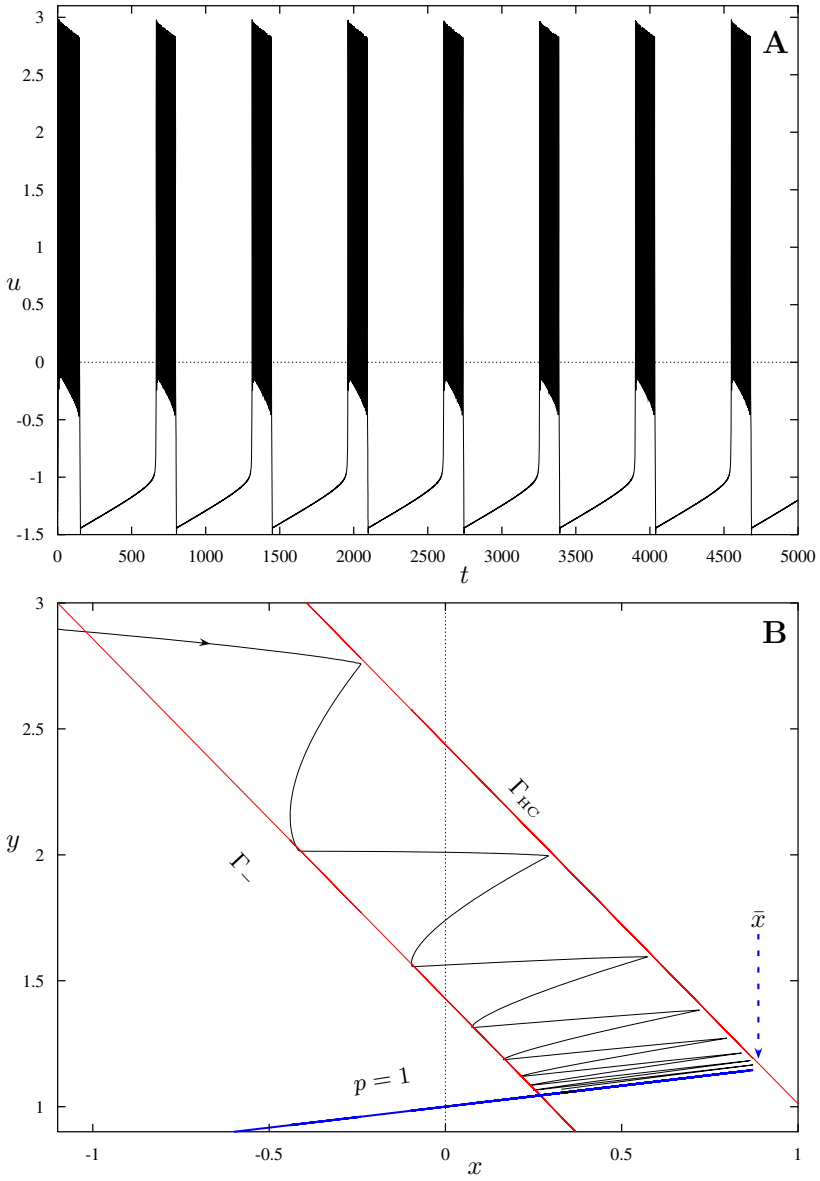


FIG. 4.1. *Illustration of the attraction of projected trajectories to $p = 1$ in the degenerate case.* **A** *shows u versus t for this run. In* **B** *we see the compression of trajectories projected into the (x, y)-plane to the line $y = \frac{\beta_2}{\beta_1} x + \beta_2(\alpha_1 - \alpha_2)$. Parameter values for the run are listed in the text.*

This is the time taken by a projected (SS) trajectory to traverse from $\mathbf{X}_n \in \Gamma_{\rm HC}$ to $\mathbf{X}_{n+1} \in \Gamma_-$, as illustrated in Figure 4.2. Also shown is the line $p = 1$, to which such trajectories are attracted, and the p coordinates p_n and p_{n+1} on $\Gamma_{\rm HC}$ and Γ_-, respectively.

Having found T_s, we integrate (4.8) in the slow time τ,

$$\int_{p_n}^{p_{n+1}} \frac{1}{1-p}\, dp = \int_0^{T_s} \frac{1}{\tau_1}\, d\tau,$$

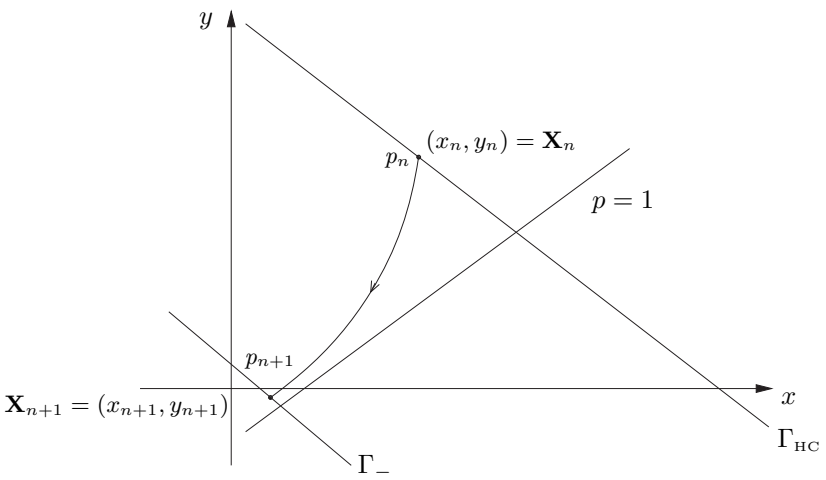


FIG. 4.2. *Projected trajectories on S_L associated with $\widetilde{\phi}$, the degenerate $\tau_1 = \tau_2$ case.*

and find that

$$\ln\left|\frac{1-p_n}{1-p_{n+1}}\right| = \frac{T_s}{\tau_1}. \tag{4.14}$$

However, (4.8) guarantees that $1-p_n$ and $1-p_{n+1}$ have the same sign, so we may drop the absolute values sign in (4.14) and solve for p_{n+1}:

$$p_{n+1} = 1 - (1-p_n)\, e^{\frac{-T_s}{\tau_1}}. \tag{4.15}$$

To convert this expression back into our original (x, y)-coordinate system, we note that by using $\tau_1 = \tau_2$ and (4.4) in (4.1) one finds

$$p = \bar{P}(x, y) = x - \frac{\beta_1}{\beta_2} y + \beta_1(\alpha_1 - \alpha_2) + 1. \tag{4.16}$$

Thus, (4.15) becomes

$$\bar{P}\left(x_{n+1}, \frac{z_- - x_{n+1}}{\gamma}\right) = 1 - \left(1 - \bar{P}\left(x_n, \frac{z_{\text{HC}} - x_n}{\gamma}\right)\right) e^{\frac{-T_s}{\tau_1}}, \tag{4.17}$$

which when solved for x_{n+1} allows us to finally obtain

$$x_{n+1} = \widetilde{\phi}(x_n) = e^{\frac{-T_s}{\tau_1}} x_n + b_s, \tag{4.18}$$

where

$$b_s = \bar{B}(T_s, z_-, z_{\text{HC}}) \equiv \frac{\left(1 - e^{\frac{-T_s}{\tau_1}}\right)\beta_1\beta_2\gamma(\alpha_2 - \alpha_1) + \beta_1\left(z_- - z_{\text{HC}} e^{\frac{-T_s}{\tau_1}}\right)}{\beta_1 + \gamma\beta_2}. \tag{4.19}$$

In a fashion analogous to the derivation of $\widetilde{\phi}$, we compute $\widehat{\phi}$ by integrating (4.8) over the active phase duration T_a. A leading-order value for T_a can be computed by integrating the averaged fast subsystem corresponding to the transformed system

(4.5)–(4.8). Applying the method of averaging to system (4.5)–(4.8), one finds its associated (AFS) is

$$\frac{dz}{d\tau} = \frac{\widehat{H}(z) - z}{\tau_1}, \tag{4.20}$$

$$\frac{dp}{d\tau} = \frac{1-p}{\tau_1}, \tag{4.21}$$

where

$$\widehat{H}(z) = H(\hat{u}(z)) = \beta^*(\hat{u}(z) - \alpha^*), \tag{4.22}$$

and $\hat{u}(z)$ is as defined in (2.21).

In order to compute the active phase duration T_a, we integrate (4.20) as follows:

$$\int_{z_-}^{z_{\rm HC}} \frac{dz}{\widehat{H}(z) - z} = \int_0^{T_a} \frac{1}{\tau_1}\, d\tau,$$

and then solve for T_a to find

$$T_a = \tau_1 \int_{z_-}^{z_{\rm HC}} \frac{dz}{\widehat{H}(z) - z}. \tag{4.23}$$

We let p_n be the p coordinate of an (AFS) trajectory with initial conditions $(x_n, y_n) \in \Gamma_-$. Similarly, we define p_{n+1} to be the p-coordinate of the (AFS) trajectory as it leaves the active phase at $(x_{n+1}, y_{n+1}) \in \Gamma_{\rm HC}$. Then, integrating (4.21),

$$\int_{p_n}^{p_{n+1}} \frac{1}{1-p}\, dp = \int_0^{T_a} \frac{1}{\tau_1}\, d\tau,$$

and proceeding exactly as we did in (4.14)–(4.17) while using the value of T_a from (4.23), we find

$$x_{n+1} = \widehat{\phi}(x_n) = e^{\frac{-T_a}{\tau_1}} x_n + b_a\,, \quad b_a = \bar{B}(T_a, z_{\rm HC}, z_-), \tag{4.24}$$

where the function $\bar{B}$ was defined in (4.19). Also, since $\phi(x) = \widehat{\phi}(\widetilde{\phi}(x))$, (4.18) and (4.24) imply

$$\phi(x) = e^{\frac{-(T_s+T_a)}{\tau_1}} x + b_{as}\,, \tag{4.25}$$

where $b_{as} = e^{\frac{-T_a}{\tau_1}} b_s + b_a$.

From this expression, the fixed point $\bar{x}$ of ϕ is easily computed as

$$\bar{x} = \frac{b_{as}}{1 - e^{\frac{-(T_s+T_a)}{\tau_1}}}. \tag{4.26}$$

It should be noted that the dependence of $\bar{x}$ on the active and silent durations in (4.26) is misleading. As can be seen in Figure 4.1, the fixed point can also be computed as that x value, where the line $p = 1$ intersects $\Gamma_{\rm HC}$. Using (4.16), this value can be found by solving $\bar{P}(x, y) = 1$ and $x + \gamma y = z_{\rm HC}$ for x to find

$$\bar{x} = \frac{\gamma\beta_1\beta_2(\alpha_2 - \alpha_1) + \beta_1 z_{\rm HC}}{\beta_1 + \gamma\beta_2} = \frac{\gamma\beta_1\beta_2(\alpha_2 - \alpha_1) + \beta_1 z_{\rm HC}}{\beta^*}. \tag{4.27}$$

However, given the previous definitions, it is readily verified that (4.26) indeed simplifies to (4.27).

To conclude this section we make some observations about the dependence of the map ϕ on the (activation/inactivation) parameters β_1 and β_2. As was pointed out earlier, β_1 and β_2 need not be of the same sign for ϕ to have the fixed point calculated in (4.27). The issue here is, what degrees of activation and inactivation can lead to such degenerate bursting solutions? Since (4.8) has no dependence on these parameters, the answer to this question is equivalent to knowing the parameter sets (α^*, β^*) for which (4.5)–(4.7) has bursting solutions. A complete description of the set $\mathcal{B}$ of (α^*, β^*) pairs for which such bursting solutions exist is complex but well studied. For instance, for an appropriate $\beta^* > 0$, as α^* is varied the system makes a transition from bursting to continuous spiking via a complex sequence of bifurcations [54]. However, one can glean a few simple results from the explicit expressions for $\bar{x}$ and $\phi(x)$.

For example, when $\beta^* < 0$, system (4.5)–(4.7) typically does not have a bursting solution. Now suppose β^* is initially positive and $\beta_1 < 0$ (inactivation) is decreased while $\beta_2 > 0$ (activation) and $\alpha_k, k = 1, 2$ are held fixed. Given (4.10) and (4.27), one then sees that $\beta^* \to 0+$ and $|\bar{x}| \to \infty$. Alternately, as inactivation increases (with other parameters fixed), the fixed point of the map ϕ will increase in magnitude.

Lastly, we emphasize that other limiting cases may be possible since (α^*, β^*) depend on all $\beta_1, \beta_2, \alpha_1$, and α_2. However, from (4.25),

$$|\phi'(\bar{x})| = e^{\frac{-(T_s+T_a)}{\tau_1}} < 1$$

implies that no such limiting cases in the degenerate case can involve a destabilization of the fixed point $\bar{x}$. Moreover, the singular bursting solutions are always stable (in this degenerate case).

5. Numerically approximating ϕ in the nondegenerate case. In the nondegenerate case when $\tau_1 \neq \tau_2$, explicit formulas for $\widehat{\phi}$ and $\widetilde{\phi}$ remain elusive. In this section we outline a continuation technique for approximating these maps numerically. The technique requires recasting the map values as boundary conditions in two point boundary value problems which are homotopic to simpler problems whose solutions are known.

To be specific, the map $\widetilde{\phi}$ for the slow flow on $\mathbb{S}_L$ is computed as a solution of the following boundary value problem:

$$\frac{dx}{d\tau} = Tf_1(x, y, \lambda) = T\left(-(1-\lambda)x + \lambda F_1(x, y)\right), \tag{5.1}$$

$$\frac{dy}{d\tau} = Tf_2(x, y, \lambda) = T\left(-(1-\lambda)y + \lambda F_2(x, y)\right), \tag{5.2}$$

$$x(0) = x_0, \tag{5.3}$$

$$y(0) = (z_{\text{HC}} - x_0)/\gamma, \tag{5.4}$$

$$x(1) = x_f, \tag{5.5}$$

$$y(1) = (z_- - x_f)/\gamma, \tag{5.6}$$

where the vector field $\mathbf{F} = (F_1, F_2)$ is that of the (SS) in (2.18)–(2.19), T is a constant, λ is a homotopy parameter, and $\mathbf{f}(x, y, \lambda) \equiv (f_1, f_2)$. The boundary conditions (5.3)–(5.4) imply $(x(0), y(0)) \in \Gamma_{\text{HC}}$. Similarly, (5.5)–(5.6) imply $(x(1), y(1)) \in \Gamma_-$. Thus, solutions of this boundary value problem describe trajectories which start on Γ_{HC} and terminate on Γ_-. When $\lambda = 1$, $\mathbf{f}(x, y, 1) = \mathbf{F}(x, y)$, so that $x_f = \widetilde{\phi}(x_0)$ providing

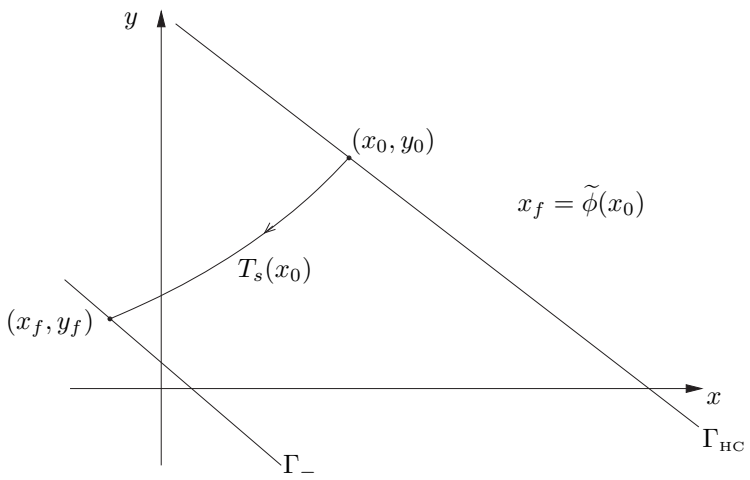


FIG. 5.1. *An illustration of $\widetilde{\phi}$ and the (SS) projected into the (x, y)-plane.*

$T = T_s$, the silent phase duration. A diagram illustrating this $\lambda = 1$ case is shown in Figure 5.1.

When $\lambda = 0$, the solution of (5.1)–(5.6) is known explicitly:

$$x(\tau) = x_0 e^{-\tau T}, \quad y(\tau) = \frac{z_{\rm HC} - x_0}{\gamma} e^{-\tau T}, \quad T = ln\left(\frac{z_{\rm HC}}{z_-}\right), \quad x_f = x_0 e^{-T}.$$

AUTO [19, 20] was then used to numerically continue this known solution to the $\lambda = 1$ case while letting the three parameters (λ, x_f, T) vary. Then, keeping $\lambda = 1$ fixed, (x_0, x_f, T) are allowed to vary in a subsequent run over a prescribed range of initial x values x_0. Given Figure 5.1, the resulting x_f values are $\widetilde{\phi}(x_0)$, and T is the silent phase duration T_s.

Results of these calculations are illustrated in Figures 5.2(a) and 5.2(b). There, $\widetilde{\phi}(x)$ is illustrated for the two-parameter sets listed in Table 5.1. For the $(+,+)$ case (Figures 5.2(a), (c), (e)) where $\beta_i > 0$, both x and y model activating variables in (FULL). In an analogous fashion, the $(+,-)$ case (Figures 5.2(b), (d), (f)) models the competing effect of activating and inactivating variables. Superimposed on the figures is the line $y = x$ as a point of reference. As a note, however, the fixed points of $\widetilde{\phi}(x)$ do not correspond to bursting solutions of (FULL).

The technique for generating $\widehat{\phi}$ from the (AFS) is similar. One first defines the boundary value problem

$$\frac{dx}{d\tau} = T g_1(x, y, \lambda) = T\left(x(1-\lambda) + \lambda \widehat{G}_1(x, y)\right), \tag{5.7}$$

$$\frac{dy}{d\tau} = T g_2(x, y, \lambda) = T\left(\lambda \widehat{G}_2(x, y)\right), \tag{5.8}$$

$$x(0) = x_0, \tag{5.9}$$

$$y(0) = (z_- - x_0)/\gamma, \tag{5.10}$$

$$x(1) = x_f, \tag{5.11}$$

$$y(1) = (z_{\rm HC} - x_f)/\gamma, \tag{5.12}$$

TABLE 5.1
Standard parameter sets.

	β_1	α_1	β_2	α_2	γ	τ_1	τ_2
$(+,+)$	3	-1	0.5	-3	0.7	0.9	1
$(+,-)$	4	-1	-1	-0.7	1	1	0.3

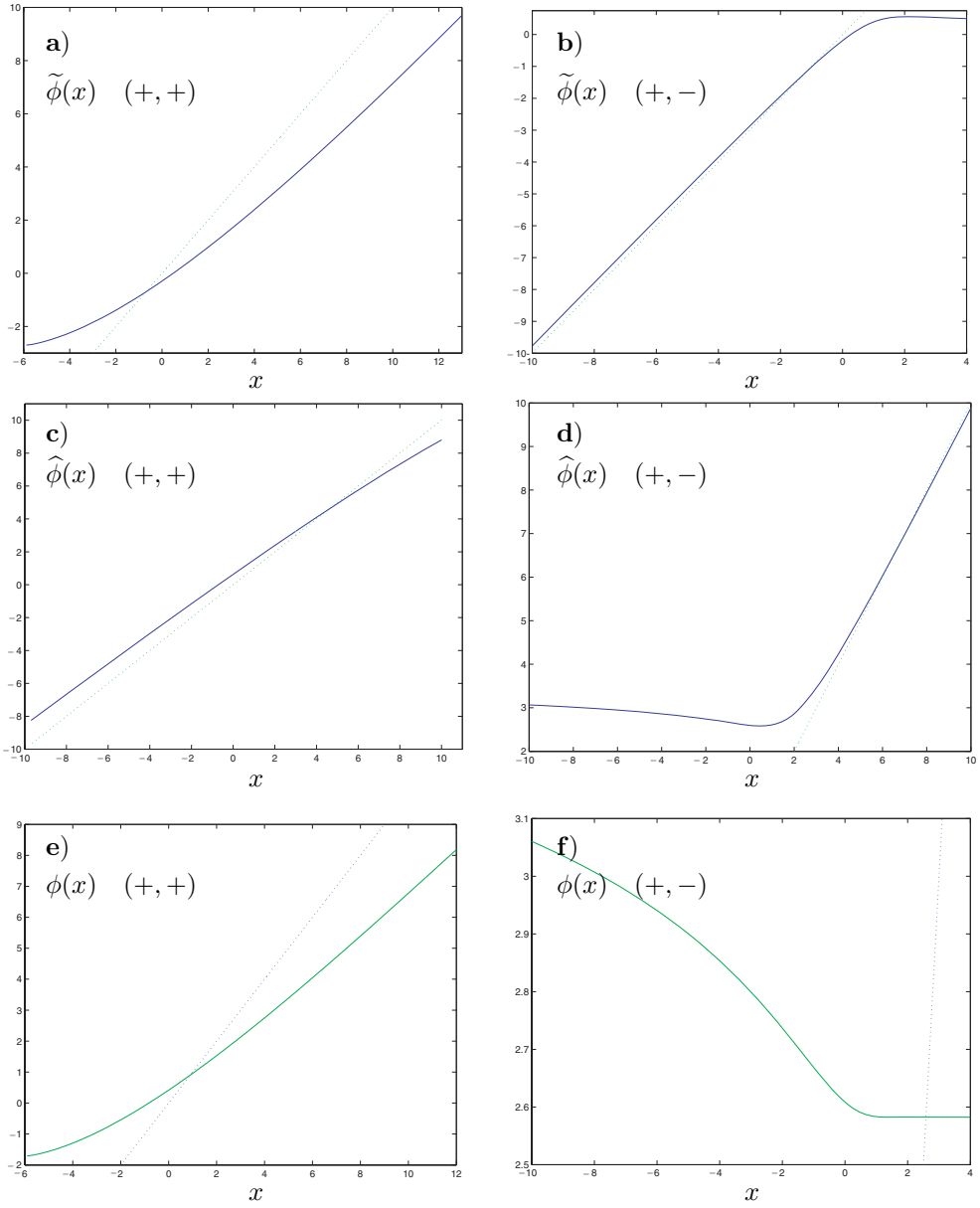


FIG. 5.2. *Maps $\widetilde{\phi}, \widehat{\phi}$, and $\phi(x)$ generated numerically using AUTO. Shown are $\widetilde{\phi}(x)$, $\widehat{\phi}(x)$, and $\phi(x)$ for a range of initial x values; the $(+,+)$ case is shown in* **a**, **c**, **e**; *the $(+,-)$ case in* **b**, **d**, **f**. *Parameter values for each computation are tabulated in Table* 5.1.

where the vector field $\widehat{\mathbf{G}} = (\widehat{G}_1, \widehat{G}_2)$ is that defined in (2.22)–(2.23), and $\mathbf{g}(x, y, \lambda) \equiv (g_1, g_2)$. As before, $\mathbf{g}(x, y, 1) = \widehat{\mathbf{G}}(x, y)$ implies $x_f = \widehat{\phi}(x_0)$ when T is the active phase duration T_a.

However, the problem (5.7)–(5.12) differs from (5.1)–(5.6) in an essential way. In the former, the vector field $\mathbf{g}(x, y, 0) = (x, 0)$ was chosen to match the flow direction of the (AFS) *from* Γ_- to Γ_{HC}. In contrast, the flow direction of $\mathbf{f}(x, y, 0) = (-x, -y)$ was chosen so that trajectories starting on Γ_{HC} would terminate on Γ_-. The choice of initial ($\lambda = 0$) vector fields is not unique. Here we have chosen simple ones which retain the flow directionality of each subsystem and whose analytic solution is known. For the choice $\mathbf{g}(x, y, 0) = (x, 0)$, the exact initial solution is

$$x(\tau) = x_0 e^{\tau T}, \quad y(\tau) = \frac{z_- - x_0}{\gamma}, \quad T = ln\left(\frac{z_{\text{HC}} - z_- + x_0}{x_0}\right), \quad x_f = x_0 e^T.$$

In an analogous fashion, AUTO was used to continue this solution in λ and then in x_0 to generate the map $x_f = \widehat{\phi}(x_0)$. In all these runs the approximation (2.24) of $\hat{u}(z)$ was used.

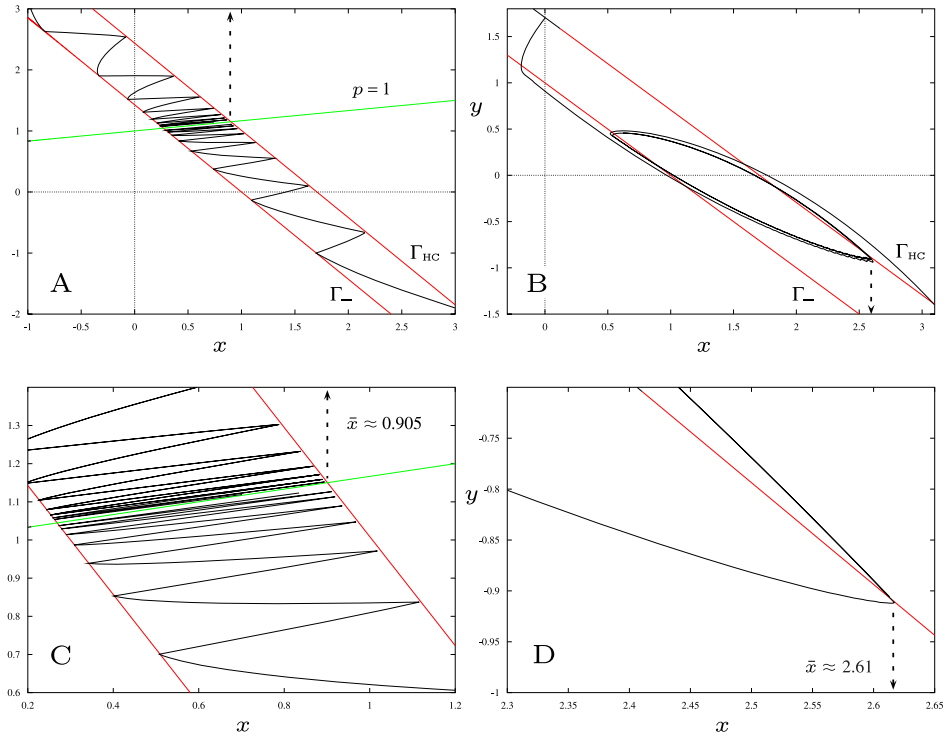


FIG. 5.3. *Verification of the fixed point $\bar{x} = \phi(\bar{x})$ obtained in the map composition illustrated in Figure 5.2. The two figures on the left are the $(+,+)$ case and the two on the right are the $(+,-)$ case. In* **A** *we see the projection of two trajectories of (FULL) into the (x, y)-plane appearing to compress toward the superimposed line $p = 1$. In* **C** *we have enlarged the region of interest about the map fixed point $\bar{x} \approx 0.905$. Also shown in* **A** *and* **C** *are the curves Γ_- and Γ_{HC}. In* **B** *we also see the projection of two trajectories of (FULL) into the (x, y)-plane, but in the $(+,-)$ case. The two projected trajectories wind onto a cycle in the (x, y)-plane. In* **D** *we are able to discern the map fixed point of $\bar{x} \approx 2.61$.*

Results of the calculations for $\widehat{\phi}$ and the respective compositions $\underset{\sim}{\phi} = \widehat{\phi} \circ \widetilde{\phi}$ are shown in Figure 5.2 for the same parameter values used to compute ϕ. The latter compositions were computed by numerically composing the data which generated $\widehat{\phi}$ and $\widetilde{\phi}$. Superimposed on the graphs of $\phi(x)$ is the line $y = x$ to indicate the location of the fixed point $\bar{x}$ corresponding to bursting solutions of (FULL). The interpolated values of the fixed points found in this manner were $\bar{x} \approx 0.905$ and $\bar{x} \approx 2.58$ for the $(+,+)$ and $(+,-)$ cases, respectively.

To separately verify these fixed point values, (FULL) was numerically integrated in Figure 5.3 over many silent and active phase cycles for the parameter values in Table 5.1. As can be seen, the projections of these solutions ultimately approach a periodic orbit, and the fixed point values computed from these figures closely agree with those computed from Figure 5.2. In all, these numerical results suggest that singular bursting solutions persist for nondegenerate parameter values. In the (+,+) case shown in Figure 5.3A, the parameter set is "nearly" degenerate and the projected bursting solution lies near the line $p = 1$. In contrast, the bursting cycle in the $(+,-)$ (activation/inactivation) case does not lie near the $p = 1$ line. In fact, the projected active and silent phase trajectories of that $(+,-)$ case are nearly tangent to the transition curves Γ_{HC} and Γ_{-}. From this observation, one might conjecture that bursting solutions do not persist for all parameter values. For example, the presence of a strongly attractive equilibria of (FULL) on $\mathbb{S}_L$ might significantly alter the (SS) flow to the point that trajectories may not be attracted to a bursting solution. These issues are explored in the next section.

6. A bistable case. In this section we demonstrate numerically that system (FULL) can exhibit bistability between bursting solutions and equilibria on $\mathbb{S}_L$. Just such an example is illustrated in Figure 6.1. Slow parameters λ_s were chosen so that (FULL) had a stable fixed point $\mathbf{X}_e$ on the lower branch $\mathbb{S}_L$ (see [25] for a detailed treatment of how to determine equilibria location and stability dependence on λ_s). In Figure 6.1B the u component of a bursting solution resulting from a particular set of initial conditions is shown. The projection of this solution onto the xy-plane is shown in Figure 6.1A. However, by choosing initial conditions near $\mathbb{S}_L$ in the basin of attraction of $\mathbf{X}_e$, the solution of (FULL) is shown to approach $\mathbf{X}_e$ in Figures 6.1C, D. The projection of the bursting solution in Figure 6.1A is also shown in Figure 6.1C for comparison.

For the bistability demonstrated in Figure 6.1 to occur, a few things must happen simultaneously. Minimally, (FULL) must have a stable equilibrium $\mathbf{X}_e$ on $\mathcal{S}_L$. However, even if parameters are chosen so that $\mathbf{X}_e \in \mathcal{S}_L$, it is not immediately clear if $\mathbf{X}_e$ will be stable or if the map ϕ can simultaneously have a stable fixed point. In this section, we shall give a brief synopsis of some issues regarding these points. First, we address issues concerning equilibria stability and location. Later, we determine some necessary conditions that map domains and ranges must satisfy for this type of bistability to occur.

The equilibria $\mathbf{X}_e$ of (FULL) have coordinates $\mathbf{X}_e = (\bar{u}, \bar{w}, \bar{x}, \bar{y})$ where, given (2.8), $\bar{u}$ are roots of

$$\Delta(u) = -G(u) + h_1(u) + \gamma h_2(u) = u^3 + au + b, \tag{6.1}$$

where

$$a = \beta_1 + \gamma\beta_2 - 3, \tag{6.2}$$

$$b = -(\alpha_1\beta_1 + \gamma\alpha_2\beta_2 + 3), \tag{6.3}$$

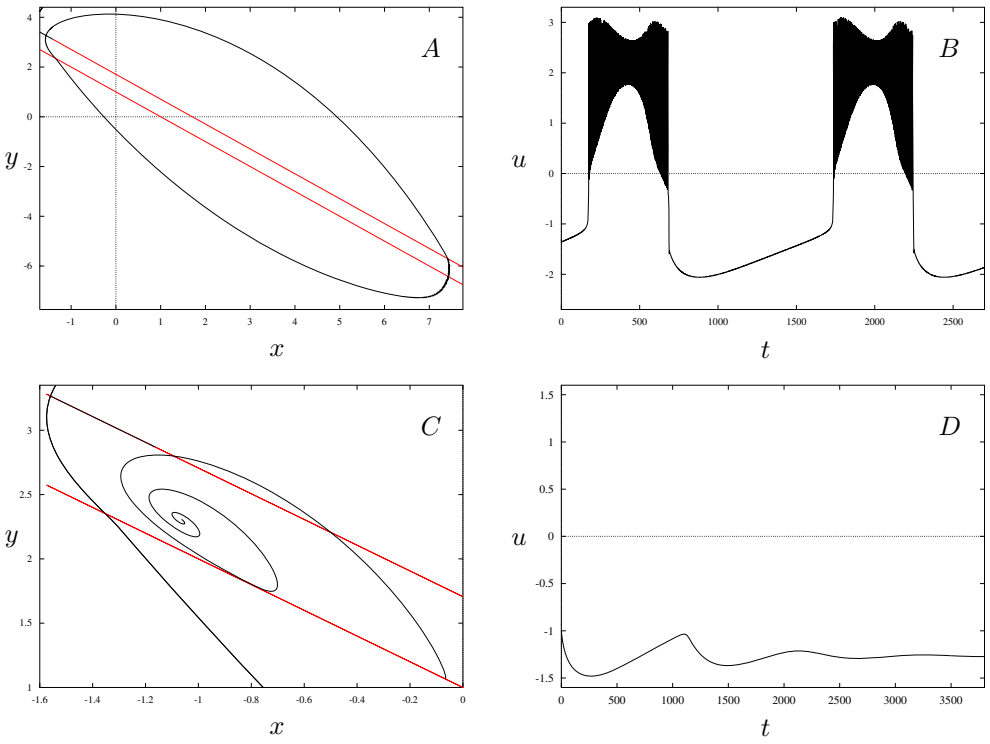


FIG. 6.1. *Numerical illustration of bistability. In both simulations the slow parameters were* $\lambda_s = (\beta_1, \beta_2, \alpha_1, \alpha_2, \tau_1, \tau_2) = (4, -3, -1, -0.5, 1, 0.3)$. *A and C are projections of the solutions of (FULL) for different initial conditions. B and D show the corresponding* $u(t)$ *component in each simulation.*

and $\bar{w} = g(\bar{u})$, $\bar{x} = h_1(\bar{u})$, and $\bar{y} = h_2(\bar{u})$. From this we note that $\mathbf{X}_e \in \mathcal{S}_L$ only if (a, b) is an element of the parameter space:

$$D_L = \{(a, b) : b = -ua - u^3, u < u_- = -1\}.$$

Thus, D_L can be characterized as the union of all those lines in the (a, b)-plane having slope $-u$ and that intercept $-u^3$ with $u < u_-$. We note, however, that even if the slow parameters λ_s are chosen so that $(a, b) \in D_L$, (FULL) may have other equilibria (see [25] for a detailed treatment of how to determine equilibria numbers and locations). Moreover, if equilibria occur on the upper branch ($u > u_+$) near the limit cycles of the (FS), it is possible that the (AFS) itself can have a fixed point. With the dynamics of the (AFS) changed, bursting solutions may no longer be possible. Thus, when searching a parameter space for the type of bistability described in this section, it is reasonable to restrict oneself to seeking parameters for which (FULL) has a unique equilibrium on $\mathcal{S}_L$. Given the cubic form of $\Delta(u)$, (FULL) will have a *unique* equilibrium $\mathbf{X}_e \in \mathcal{S}_L$ if and only if $(a, b) \in D_L$ and the discriminant

$$D_\Delta = \frac{b^2}{4} + \frac{a^3}{27} > 0. \tag{6.4}$$

Next, we address the stability of the (unique) equilibrium $\mathbf{X}_e \in \mathcal{S}_L$. Toward this end, we define $P(\lambda)$ as the characteristic polynomial of the Jacobian $D\vec{\mathbf{F}}(\mathbf{X}_e)$ of

(FULL) at $\mathbf{X}_e$ with roots λ of P being the eigenvalues. Similarly, we define $P_0(\lambda)$ as the characteristic polynomial of the Jacobian of the (FS). Using these definitions and the expansion

$$\lambda = \lambda_0 + \varepsilon\lambda_1 + \varepsilon^2\lambda_2 + \cdots$$

in $P(\lambda)$, one finds

$$P(\lambda) = \lambda_0^2 P_0(\lambda_0) + \varepsilon P_1(\lambda_0, \lambda_1) + \varepsilon^2 P_2(\lambda_0, \lambda_1, \lambda_2) + O(\varepsilon^3), \tag{6.5}$$

where for the moment we do not explicitly state the functions P_0, P_1, P_2. Since (FULL) has two fast variables and two slow variables, two of the eigenvalues $\lambda = O(1)$, while the remaining two eigenvalues $\lambda = O(\varepsilon)$. For $\mathbf{X}_e$ to be stable we require all four eigenvalues to have $\Re e(\lambda) < 0$. Clearly, if $\mathbf{X}_e \in \mathcal{S}_L$, the two $O(1)$ eigenvalues have $\Re e(\lambda) < 0$ since such λ's equal λ_0 to leading order and the roots $\lambda_0 \neq 0$ of P_0 are the same eigenvalues as in the (FS). Thus, it suffices to examine the $O(\varepsilon)$ eigenvalues, whose expansions have $\lambda_0 = 0$.

For the $\lambda_0 = 0$ case, it is easily verified that

$$P_1(0, \lambda_1) = 0 \qquad \forall \lambda_1.$$

So the "small" $O(\varepsilon)$ eigenvalues associated with the (SS) require examining the $O(\varepsilon^2)$ term in (6.5). Explicit calculations reveal

$$P_2(0, \lambda_1, \lambda_2) = Q(\lambda_1) \equiv q_2\lambda_1^2 + q_1\lambda_1 + q_0, \tag{6.6}$$

where the coefficients of Q in (6.6) are given by

$$q_2 = -G'(\bar{u}), \tag{6.7}$$

$$q_1 = \frac{\beta_1\tau_2 + \gamma\beta_2\tau_1 - G'(\bar{u})(\tau_1 + \tau_2)}{\tau_1\tau_2}, \tag{6.8}$$

$$q_0 = \frac{\beta_1 + \gamma\beta_2 - G'(\bar{u})}{\tau_1\tau_2}. \tag{6.9}$$

Notice that there is no dependence on λ_2 in $P_2(0, \lambda_1, \lambda_2)$, and that Q is quadratic in λ_1. Also, for the remainder of this section we drop the overbar notation on u.

The leading term λ_1 of these small eigenvalues is determined as the roots of Q in (6.6):

$$\lambda_1^{\pm} = \frac{-q_1 \pm \sqrt{q_1^2 - 4q_2q_0}}{2q_2}. \tag{6.10}$$

Since $u < u_- = -1$ on $\mathcal{S}_L$, the quantity $G'(u) = -3u^2 + 3$ is negative on $\mathcal{S}_L$ and

$$\mathbf{X}_e \in \mathcal{S}_L \Rightarrow q_2 > 0. \tag{6.11}$$

This result and the signs of the coefficients (q_1, q_0), based on the signs of the activation parameters[3] (β_1, β_2), are organized in Table 6.1 and are discussed in the following text. Given the signs of the coefficients q_i, we are then able to determine the stability of $\mathcal{S}_L$ equilibria. Throughout, recall $0 < \gamma,\ 0 < \tau_i,\ i = 1, 2$, and (6.11), i.e., $q_2 > 0$.

[3]Recall that x is a slow activation variable if $\beta_1 > 0$ and is a slow inactivation variable if $\beta_1 < 0$. Analogous remarks hold for y and β_2.

TABLE 6.1

Stability dependence of the equilibria $\mathbf{X}_e \in \mathcal{S}_L$ on the activation/inactivation parameters (β_1, β_2), where (q_2, q_1, q_0) are the coefficients of the quadratic $Q(\lambda)$ defined in (6.6). In the "Case" column, $(+,-)$ indicates the respective signs of β_1 and β_2. In the other columns, $+/-$ indicates that both signs of the column quantity are possible. In both the $(+,-)$ and $(-,+)$ cases, the equilibria $\mathbf{X}_e$ can undergo a Hopf bifurcation, and thus may be stable or unstable.

Case	sign(q_2)	sign(q_1)	sign(q_0)	Equilibria $\mathbf{X}_e$
$(+,+)$	$+$	$+$	$+$	stable
$(-,-)$	$+$	$+$	$+$	stable
		$+/-$	$-$	saddle
$(+,-)$	$+$	$+/-$	$+/-$	Hopf bifurcation possible
$(-,+)$	$+$	$+/-$	$+/-$	Hopf bifurcation possible

When $(\beta_1, \beta_2) = (+,+)$, i.e., both are positive, one sees immediately that $q_1 > 0$ and $q_0 > 0$ as well. As such, any $\mathcal{S}_L$ equilibria in this case will be stable, as noted in Table 6.1.

When $(\beta_1, \beta_2) = (-,-)$, it is possible for q_0 to be positive or negative. To understand this case we first note from (6.1) that

$$\Delta'(u) = -G'(u) + (\beta_1 + \gamma\beta_2). \tag{6.12}$$

If $q_0 > 0$, then $G'(u) < \beta_1 + \gamma\beta_2$, and we may conclude $\Delta'(u) > 0$. Rewriting q_1 in terms of $\Delta(u)$, we find

$$q_1 = \frac{\Delta'(u)}{\tau_1} + \frac{\Delta'(u)}{\tau_2} - \frac{\gamma\beta_2}{\tau_1} - \frac{\beta_1}{\tau_2}. \tag{6.13}$$

We now see that if $(\beta_1, \beta_2) = (-,-)$ and $q_0 > 0$, then $q_1 > 0$ so that $\mathbf{X}_e \in \mathcal{S}_L$ are stable.

If $(\beta_1, \beta_2) = (-,-)$ and $q_0 < 0$, then q_1 may be positive or negative. Moreover, in this case, $\sqrt{q_1^2 - 4q_2q_0} > |q_1|$ so that $\lambda_1^{\pm}$ are both real and have opposite signs. Then, $\mathbf{X}_e$ has a one-dimensional unstable manifold and a three-dimensional stable manifold. We list such unstable equilibria $\mathbf{X}_e$ of (FULL) in Table 6.1 as "saddles" since the associated equilibria $(\bar{x}, \bar{y})$ of the (SS) will be saddles.

Finally, we consider the $(\beta_1, \beta_2) = (+,-)$ and $(\beta_1, \beta_2) = (-,+)$ cases together. It is not hard to see in that restrictions on the sign of q_0 do not place similar restrictions on the sign of q_1; any sign is possible. Thus, $\lambda_1^{\pm}$ may be complex conjugate pairs, and transverse crossings of the imaginary axis could be possible. We do not discuss the details of this here but merely note that $\mathbf{X}_e$ may be stable or unstable and that Hopf bifurcations of such equilibria are possible (see [25] for a detailed treatment).

At this stage we summarize some of the previous results. First, to ensure there is a unique equilibria $\mathbf{X}_e \in \mathcal{S}_L$, slow parameters must be chosen so that $(a, b) \in D_L$ while simultaneously satisfying (6.4). The stability of such equilibria will largely depend on the sign of the activation/inactivation parameters (β_1, β_2). This dependence is summarized in Table 6.1. In Figure 6.1, the slow parameters were chosen to satisfy these conditions in the $(\beta_1, \beta_2) = (+,-)$ case. Moreover, in this case $\lambda_1^{\pm}$ were complex conjugates to create the spiral motion depicted in Figure 6.1C. We do not exclude the possibility of bistability for the other cases in Table 6.1. Even if this issue is resolved, it does not address issues concerning the simultaneous coexistence of a stable fixed point of the map ϕ needed to ensure the existence of a stable bursting solution. To address this latter issue, we examine the map domains and ranges defined in previous sections.

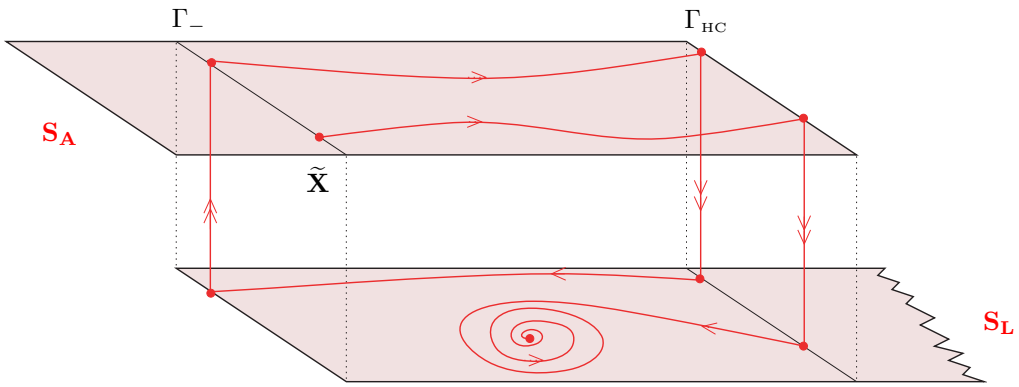


FIG. 6.2. *Illustration of bistability between bursting solutions and stable equilibria of (FULL). One solution illustrated has an initial condition $\widetilde{\mathbf{X}} \in S_A$ in the active phase but with $\hat{\Phi}(\widetilde{\mathbf{X}})$ in the basin of attraction of an equilibria of (FULL) on S_L. Also shown is a bursting cycle associated with the fixed point of Φ.*

First, we note that Figure 6.2 illustrates the simulations shown in Figure 6.1. In the following discussion we assume the domain $D(\widehat{\phi}) = \mathbb{R}$, whereas $D(\widetilde{\phi})$ is a proper subset of $\mathbb{R}$. While shortly it will become evident that the latter is necessary for bistability, we acknowledge that the former assumption is not. For example, for other parameter values the (AFS) may itself have a stable fixed point. For simplicity, we assume this is not the case in the following discussion.

For the system to exhibit the bistability illustrated in Figure 6.2, some trajectories of the (SS) originating along Γ_{HC} must be attracted to the equilibria $\mathbf{X}_e$ on $\mathbb{S}_L$. Since $D(\widetilde{\phi})$ consists only of initial x-coordinates for which the (SS) trajectories starting on Γ_{HC} reach Γ_-, we conclude that in the bistable case, $D(\widetilde{\phi})$ must be a proper subset of $\mathbb{R}$.

Now, consider an initial condition where $\widetilde{\mathbf{X}} = (x_0, y_0) \in \Gamma_-$ and the fast variables are sufficiently near the (FS) limit cycle $\Omega(t, z_0), z_0 = x_0 + \gamma y_0$. To leading order, the resulting trajectory $\gamma_{\widetilde{X}}$ will traverse the active phase as described by the (AFS). Since $x_0 \in D(\widehat{\phi})$, $\gamma_{\widetilde{X}}$ will eventually reach Γ_{HC} and exit to the silent phase. Such a trajectory will enter the silent phase at an x-coordinate $\widehat{\phi}(x_0) \in R(\widehat{\phi})$.[4] The next issue is whether $\widehat{\phi}(x_0) \in D(\widetilde{\phi})$. If $\widehat{\phi}(x_0) \in D(\widetilde{\phi})$, then the trajectory starting in the active phase will eventually traverse the entire slow manifold and make a transition back to the active phase. The set of x values on the slow manifold for which this is possible is the previously discussed set $W = R(\widehat{\phi}) \cap D(\widetilde{\phi})$. Since $D(\widetilde{\phi})$ must be a proper subset of $\mathbb{R}$ in the bistable case, it follows that W must also necessarily be a proper subset of $\mathbb{R}$. The next issue is whether W is equal to or is a proper subset of $D(\widetilde{\phi})$. Clearly, $W \neq D(\widetilde{\phi})$; if it were, no (AFS) trajectory making a transition into the silent phase would ever be in the basin of attraction of the equilibria $\mathbf{X}_e$ on $\mathbb{S}_L$, and bistability would not be possible. To summarize, two necessary conditions for the system to exhibit bistability are

$$D(\widetilde{\phi}) \subset \mathbb{R}, \quad W = R(\widehat{\phi}) \cap D(\widetilde{\phi}) \subset D(\widetilde{\phi}),$$

with the understanding that the inclusions are proper.

[4]The range $R(\widehat{\phi})$ of $\widehat{\phi}$.

To conclude this section we remark that the above conditions, though necessary, may not be sufficient for bistability. For example, even if $\widehat{\phi}(x_0) \in D(\widetilde{\phi})$, in the next iterate it may be that $(\widehat{\phi} \circ \widetilde{\phi} \circ \widehat{\phi})(x_0) \notin D(\widetilde{\phi})$—in which case the trajectory would eventually be attracted to the equilibria $\mathbf{X}_e$. To completely resolve these types of details, accurate estimations of the map domains and ranges are paramount. We do not make such estimates in this paper, but some progress toward resolving these issues is presented in [25].

7. Conclusion and discussion. In this paper we have shown how singular approximations of bursting solutions of a model with two slow variables can be identified with fixed points of a one-dimensional map. When the time constants of the slow variables are equal (the degenerate case), the map and fixed point can be computed explicitly. Such a calculation is possible because, in that case, the system decouples under a simple transformation. Furthermore, from the transformed system (4.5)–(4.8) it was deduced that the original model has singular bursting solutions even if the slow variables are activating ($\beta_1 > 0$) and inactivating ($\beta_2 < 0$). In other studies both slow variables were inhibitory [51]. It was also shown that as inactivation was increased in this degenerate case, the magnitude of the fixed point $\bar{x}$ increased. Further, given the negative slope of the curve Γ_{HC}, as $\bar{x}$ increases the associated $\bar{y}$ value decreases. Insofar as the model studied in this paper is homotopic to other two slow variable models exhibiting bursting, the $(\bar{x}, \bar{y})$ would represent the extreme values of the slow regulatory variables (i.e., calcium concentration, channel activation variable) at the start of the silent phase. Thus, experimentally, if the slow regulatory variables have similar time constants, one might expect to see these extreme values increase and decrease inversely as inactivation of one process is increased. However, independent of the levels of activation and inactivation, no bifurcations of the bursting solution through a destabilization of the map fixed point are possible in the degenerate case. Moreover, as previously noted, the singular bursting solutions are always stable in this degenerate case.

In the more generic nondegenerate case when the time constants of the slow variables are not equal, it was demonstrated in section 5 how the maps used to determine bursting solutions can be computed by solving two one-parameter families of boundary value problems. There, AUTO [19] was used to homotope from known solutions to a solution which describes trajectories of the (SS) and the (AFS) of the original model. The methods described in section 5 would be substantially faster than using multiple integrations of (FULL) to compute the Poincaré return maps. Moreover, since AUTO automatically detects a variety of bifurcations, the aforementioned methods would be far better suited for numerical studies of bifurcations of bursting solutions in systems with two slow variables.

Some of the numerical techniques presented in section 5 might be adaptable to other models of bursting. For instance, in other models, slow subsystems can often be computed explicitly so that system (5.1)–(5.2) can be coded. However, in models such as (1.1)–(1.2), saddle-node and homoclinic bifurcation points of the associated fast subsystems are not known explicitly. Thus, it may not be possible to explicitly code boundary conditions such as (5.4) and (5.6). One possible resolution to this difficulty is to augment (5.1)–(5.6) with the (FS) of the model. It should be noted that similar issues would arise when attempting to code (5.7)–(5.12) in other models. In addition to the aformentioned issues, the vector field of the (AFS) of other models is not explicitly known. However, this issue can be circumvented by first calculating averaged quantities over a grid of slow variable values. Then, intermediate values can

be computed using interpolation from this tabulated data within the AUTO code. In [25] such a method was implemented to determine $\hat{\phi}$ for the model defined in this paper.

Of additional importance is the bistability between bursting solutions and equilibria of (FULL) demonstrated in section 6. This is a fundamentally new type of bistability. In two variable neuron models, such as the FitzHugh–Nagumo model [23, 35], one variable is often slow and the other fast. The resulting (FS) exhibits bistability between equilibria. Bistability in the (FS) of three variable models exhibiting bursting is typically between stable equilibria and planar limit cycles. Such is the case in the (FS) of the model studied here. Even this type of bistability should be contrasted with bistability between stable periodic solutions such as that studied by Canavier et al. [10, 11]. Shilnikov, Calabrese, and Cymbalyuk [47] have also examined bistability between bursting solutions and (tonic) periodic solutions in a neuronal model with one slow variable. In this study, the bistability discussed in section 6 is between equilibria and (four-dimensional) bursting solutions. Regarding (FULL) as a model of neural activity, this type of bistability suggests some interesting potential neural dynamics. For instance, certain perturbations could switch the neural activity between bursting modes and quiescent modes. Since bursting typically acts on a time scale an order of magnitude larger than tonic spiking, this might have functional relevance in regulating physiological processes on time scales of the order of many bursts.

The former type of bistability is also of interest as it relates to the issue of why some isolated pancreatic β-cells burst while others do not [50]. Bursting electrical activity is observed in β-cells that are still intact in the islets [42]. At elevated temperatures (e.g., 30° C), bursting is also observed in β-cells that are isolated from the islets [48]. Yet, in other experiments at room temperature, isolated β-cells exhibit irregular spiking but do not burst [22]. Some researchers have discussed how the fundamental stochasticity of the β-cells might explain these different behaviors [42]. In other studies [37, 38], cell heterogeneity and diffusive coupling was a premise used to explain this experimental fact. There it was shown that if collections of cells with a stable equilibria were diffusively coupled to collections of bursting cells, the coupled system could burst synchronously. However, this explanation used a model with a sole slow variable. Since most recent models of β-cell electrical activity have two or more slow variables, it now appears that cell bistability might also play a role. For instance, if for some parameter sets these newer models exhibit the same sort of bistability demonstrated in section 6, then the experimental preparation could play a role. Some extracted cells may not burst merely because those preparations have slow variable values in the basin of attraction of the stable equilibria on S_L. Moreover, if the latter basin of attraction were sufficiently large, then even noise would not affect the result. Alternately, the sporadic spiking of "nonbursters" might be due to the noise (or channel stochasticity) being sufficiently large so that the (SS) trajectories traverse close to the basin of attraction of the equilibria $\mathbf{X_e}$. When properly exploring such ideas using a model, the size of the map domains become especially important. Also, in this case, the analytical and numerical tools developed in this paper may be particularly useful.

We note, however, that it is not known if the more recent β-cell models can exhibit such bistability for other parameter values. Additionally, we currently have no systematic way of determining such parameter sets. For the model used in this paper, some advances in solving this latter problem are presented in [25]. Although it is not too difficult to determine parameter sets where (FULL) has stable equilibria

on S_L, it is difficult to determine a subset of such a parameter space where the model simultaneously exhibits bursting. As mentioned previously, the latter is intimately connected with domain issues of the maps $\tilde{\phi}$ and $\hat{\phi}$, which requires knowledge of certain global information about the slow and averaged fast subsystems.

The bistability discussed in section 6 also presents some other interesting modeling possibilities. For instance, if collections of such cells are coupled, then aggregate behavior would depend greatly on initial conditions. In reaction diffusion systems where bistability is between equilibria, traveling wave phenomena are ubiquitous. If the bistability is between equilibria and bursting solutions, the question arises if wavefronts separating such behaviors are possible. Even if no such (stable) wave phenomena exist, it is not known if strongly coupled aggregates exhibit bistable synchronous solutions. For instance, synchronous (monostable) solutions are present in heterogeneous collections of cells which are diffusively coupled with strong coupling [38].

Lastly, the approximation of the average quantity $\widehat{u}(z)$ in (2.24) is of some mathematical interest. In Figure 2.2, the functions $\widehat{u}_a(z)$ and $\widehat{u}(z)$ are shown to be very close over a wide range of z. This closeness suggests that (2.24) is a two term asymptotic approximation of the average $\widehat{u}(z)$. If this is the case, we are unaware of a systematic (analytical) method for estimating the parameters p, a_0, and a_1 in that approximation. Moreover, we are not aware how robust this approximation is when fast parameter values are altered. In Figure 2.2 the fast parameters are fixed and the comparison is over different values of z. Regardless, such approximation methods would be of interest, as they might apply to the more complicated Hodgkin–Huxley based models with two fast variables. For instance, were such methods available and the dependence of a_0 on fast parameters derivable, then it would be easier to predict how things such as the active phase duration might depend on fast conductances and other biological quantities.

Appendix. Here we use a multiple scales procedure to derive the (AFS) summarized in section 2.3. We assume that the (FS) has $T(\mathbf{z})$-periodic solutions $\Omega(t, \mathbf{z})$ which satisfy

$$\frac{d\Omega}{dt} = \mathbf{F}(\Omega, \mathbf{z}) \quad \forall \mathbf{z} \in S_A\,. \tag{A.1}$$

We seek an asymptotic approximation of (2.1)–(2.2) valid to $O(\varepsilon)$ for $t = O(\frac{1}{\varepsilon})$. Toward that end, we assume an expansion of the form

$$\mathbf{u}(t) = U(s, \tau, \varepsilon) = U_0(s, \tau) + \varepsilon U_1(s, \tau) + O(\varepsilon^2)\,, \tag{A.2}$$

$$\mathbf{z}(t) = Z(s, \tau, \varepsilon) = Z_0(s, \tau) + \varepsilon Z_1(s, \tau) + O(\varepsilon^2)\,, \tag{A.3}$$

where τ is a slow time,

$$\tau = \varepsilon t, \tag{A.4}$$

and the strained (fast) time s is defined by

$$\frac{ds}{dt} = \omega(\tau) \tag{A.5}$$

for a function ω as yet to be determined.

As with other multiple scales methods, it is possible to choose an appropriate ω so that U_0 is 1-periodic in s. In addition to the aforementioned periodicity, we want

our leading-order approximation $U_0(s,\tau)$ close to $u(t)$ for times $\tau = O(1)$. With U_0 and Z_0 suitably determined, we also require $U_i(s,\tau)$ and $Z_i(s,\tau)$ to be 1-periodic in s for $i = 0, 1, 2, \ldots$. The periodicity of U_i and Z_i in s ensures the functions U, Z are bounded in s, so that, for instance, $|U - U_0| = O(\varepsilon)$ for $\tau = O(1)$.

Using these definitions, the time derivatives of $U(s,\tau,\varepsilon)$ are

$$\frac{dU}{dt} = \omega\frac{\partial U}{\partial s} + \varepsilon\frac{\partial U}{\partial \tau}\,, \tag{A.6}$$

$$\frac{d^2U}{dt^2} = \omega^2\frac{\partial^2 U}{\partial s^2} + 2\varepsilon\omega\frac{\partial^2 U}{\partial s\partial\tau} + \varepsilon\frac{\partial\omega}{\partial\tau}\frac{\partial U}{\partial s} + \varepsilon^2\frac{\partial^2 U}{\partial\tau^2}\,, \tag{A.7}$$

with similar expressions for $Z(s,\tau,\varepsilon)$.

Expanding $\mathbf{F}(\mathbf{u},\mathbf{z})$ about (U_0, Z_0) gives

$$\mathbf{F}(\mathbf{u},\mathbf{z}) = \mathbf{F}(U_0,Z_0) + \varepsilon D_{\mathbf{u}}\mathbf{F}(U_0,Z_0)U_1 + \varepsilon D_{\mathbf{z}}\mathbf{F}(U_0,Z_0)Z_1 + O(\varepsilon^2)\,, \tag{A.8}$$

where, for $\mathbf{F} = (F_1, F_2)^T$,

$$D_{\mathbf{u}}\mathbf{F} = \begin{bmatrix} \frac{\partial F_1}{\partial u} & \frac{\partial F_1}{\partial w} \\ \frac{\partial F_2}{\partial u} & \frac{\partial F_2}{\partial w} \end{bmatrix}, \tag{A.9}$$

and in a similar fashion, $D_{\mathbf{z}}\mathbf{F}$ is the Jacobian of $\mathbf{F}$ in $\mathbf{z}$.

Putting this all together, (2.1)–(2.2) become

$$\omega\frac{\partial U_0}{\partial s} + \varepsilon\left(\frac{\partial U_0}{\partial\tau} + \omega\frac{\partial U_1}{\partial s}\right) = \mathbf{F}^{(0)} + \varepsilon\left(D_{\mathbf{u}}\mathbf{F}^{(0)}U_1 + D_{\mathbf{z}}\mathbf{F}^{(0)}Z_1\right) + O(\varepsilon^2)\,, \tag{A.10}$$

$$\omega\frac{\partial Z_0}{\partial s} + \varepsilon\left(\frac{\partial Z_0}{\partial\tau} + \omega\frac{\partial Z_1}{\partial s}\right) = \varepsilon\mathbf{G}(U_0,Z_0) + O(\varepsilon^2)\,, \tag{A.11}$$

where the superscript$^{(0)}$ in (A.10) means evaluated at (U_0, Z_0). These equations can be written as a system of partial differential equations. The $O(1)$ terms from (A.10)–(A.11) are

$$\omega\frac{\partial U_0}{\partial s} = \mathbf{F}(U_0,Z_0)\,, \tag{A.12}$$

$$\frac{\partial Z_0}{\partial s} = 0\,, \tag{A.13}$$

and the $O(\varepsilon)$ terms,

$$\frac{\partial U_0}{\partial\tau} + \omega\frac{\partial U_1}{\partial s} = D_{\mathbf{u}}\mathbf{F}(U_0,Z_0)U_1 + D_{\mathbf{z}}\mathbf{F}(U_0,Z_0)Z_1\,, \tag{A.14}$$

$$\frac{\partial Z_0}{\partial\tau} + \omega\frac{\partial Z_1}{\partial s} = \mathbf{G}(U_0,Z_0)\,. \tag{A.15}$$

From (A.13) it is evident that Z_0 does not depend on s. Henceforth, we write $Z_0 = Z_0(\tau)$.

Since $\Omega(s,\mathbf{z})$ is $T(\mathbf{z})$-periodic in s, then $\Omega(Ts,\mathbf{z})$ is 1-periodic in s. Letting $\psi(\tau)$ be a slowly varying phase, then

$$U_0(s,\tau) = \Omega(sT(Z_0(\tau)) + \psi(\tau), Z_0(\tau)) \tag{A.16}$$

will also be a 1-periodic function of s. Moreover, since Ω solves (A.1), the U_0 defined in (A.16) also solves (A.12), provided we choose

$$\omega(\tau) = \frac{1}{T(Z_0(\tau))}\,. \tag{A.17}$$

With this choice of ω the strained time s is completely defined in terms of the original time t via (A.5) once $Z_0(\tau)$ has been determined. Although a subsequent determination of the dependence of U_0 on t additionally depends on the slowly varying phase ψ, it will shortly be shown that the leading-order evolution of Z_0 does not depend on ψ.

We now turn our attention to the Z equation. The requirement that Z_1 be 1-periodic in s helps us derive necessary conditions for Z_0. Integrating (A.15) over $s \in (0,1)$, we get

$$\int_0^1 \mathbf{G}(U_0(s,\tau), Z_0(\tau))ds = \int_0^1 \left(\frac{\partial Z_0}{\partial \tau} + \omega(\tau)\frac{\partial Z_1}{\partial s}(s,\tau) \right) ds \tag{A.18}$$

$$= \frac{\partial Z_0}{\partial \tau} + \omega(\tau)[Z_1(1,\tau) - Z_1(0,\tau)]\,. \tag{A.19}$$

Thus, choosing Z_0 as a solution to

$$\frac{\partial Z_0}{\partial \tau} = \int_0^1 \mathbf{G}(U_0(s,\tau), Z_0(\tau))ds = \int_0^1 \mathbf{G}\left(\Omega(sT(Z_0) + \psi(\tau)), Z_0\right) ds \tag{A.20}$$

implies $Z_1(1,\tau) = Z_1(0,\tau)$ for all τ. This in itself does not prove that Z_1 is periodic in s. However, one need only note that (A.15) is invariant under the transformation $s \to s + \bar{\psi}$ to make this inference. Using this transformation, the preceding integration of (A.15) would then imply $Z_1(1+\bar{\psi},\tau) = Z_1(\bar{\psi},\tau)$ for an arbitrary $\bar{\psi}$, from which one concludes that if Z_0 solves (A.20), then Z_1 is 1-periodic in s.

Lastly, as mentioned previously, the slowly varying phase ψ is not needed to determine the dependence of Z_0 on τ. In order to show this, we introduce the change of coordinates $\eta = sT(Z_0) + \psi(\tau)$ to rewrite (A.20) as

$$\frac{\partial Z_0}{\partial \tau} = \frac{1}{T}\int_{\psi}^{T+\psi} \mathbf{G}(\Omega(\eta, Z_0), Z_0)d\eta\,. \tag{A.21}$$

Since $\mathbf{G}(\Omega(\eta, Z_0), Z_0)$ is $T(Z_0)$-periodic in η, we have

$$\frac{\partial Z_0}{\partial \tau} = \frac{1}{T}\int_0^T \mathbf{G}(\Omega(\eta, Z_0), Z_0)d\eta \equiv \widehat{\mathbf{G}}(Z_0)\,, \tag{A.22}$$

showing that the value of the integral does not depend on $\psi(\tau)$. The right-hand equality in (A.22) has been added to emphasize that there is no explicit dependence on τ in $\widehat{G}$. Also, in summary, we may now conclude that if

$$\frac{dZ_0}{d\tau} = \widehat{\mathbf{G}}(Z_0)\,, \tag{A.23}$$

then Z_1 is 1-periodic in s so that $|Z - Z_0| = O(\varepsilon)$ for $\tau = O(1)$. Given the definition of $\mathbf{G}$ and the linearity of $h_i(u)$ in u, we see that (A.23) written out in component form is precisely (2.22)–(2.23).

REFERENCES

[1] B. Alving, *Spontaneous activity in isolated somata of Aplysia pacemaker neurons*, J. Gen. Physiol., 51 (1968), pp. 29–45.

[2] I. Atwater, C. Dawson, A. Scott, G. Eddlestone, and E. Rojas, *The nature of the oscillatory behavior in electrical activity from pancreatic β-cell*, Hormone and Metabolic Research Supplement, 10 (1980), pp. 100–107.

[3] S. Baer, J. Rinzel, and H. Carrillo, *Analysis of an autonomous phase model for neuronal parabolic bursting*, J. Math. Biol., 33 (1995), pp. 309–333.

[4] R. Bertram, M. Butte, T. Kiemel, and A. Sherman, *Topological and phenomenological classification of bursting oscillations*, Bull. Math. Biol., 57 (1995), pp. 413–439.

[5] R. Bertram, J. Previte, A. Sherman, T. Kinard, and L. Satin, *The phantom burster model for pancreatic β-cells*, Biophys. J., 79 (2000), pp. 2880–2892.

[6] R. Bertram, P. Smolen, A. Sherman, D. Mears, I. Atwater, F. Martin, and B. Soria, *A role for calcium release-activated current (CRAC) in cholinergic modulation of electrical activity in pancreatic β-cells*, Biophys. J., 68 (1995), pp. 2323–2332.

[7] W. Beyer, ed., *CRC Standard Mathematical Tables and Formulae*, 29th ed., CRC Press, Boca Raton, FL, 1991.

[8] A. Bose, N. Kopell, and D. Terman, *Almost-synchronous solutions for mutually coupled excitatory neurons*, Phys. D, 140 (2000), pp. 69–94.

[9] R. Butera, *Multirhythmic bursting*, Chaos, 8 (1998), pp. 274–284.

[10] C. C. Canavier, D. A. Baxter, J. W. Clark, and J. H. Byrne, *Nonlinear dynamics in a model neuron provide a novel mechanism for transient synaptic inputs to produce long-term alterations of postsynaptic activity*, J. Neurophysiology, 69 (1993), pp. 2252–2257.

[11] C. C. Canavier, D. A. Baxter, J. W. Clark, and J. H. Byrne, *Multiple modes of activity in a model neuron suggest a novel mechanism for the effects of neuromodulatros*, J. Neurophysiology, 72 (1994), pp. 872–882.

[12] T. Chay, *Chaos in a three-variable model of an excitable cell*, Phys. D, 16 (1985), pp. 233–242.

[13] T. Chay, *Effect of compartmentalized Ca^{2+} ions on electrical bursting activity of pancreatic β-cells*, Am. J. Physiol., 258 (1990), pp. C955–C965.

[14] T. Chay and D. Cook, *Endogenous bursting patterns in excitable cells*, Math. Biosci., 90 (1988), pp. 139–153.

[15] T. Chay and J. Keizer, *Minimal model for membrane oscillations in the pancreatic β-cell*, Biophys. J., 42 (1983), pp. 181–190.

[16] G. de Vries, *Multiple bifurcations in a polynomial model of bursting oscillations*, J. Nonlinear Sci., 8 (1998), pp. 281–316.

[17] G. de Vries, A. Sherman, and H. Zhu, *Diffusively coupled bursters: Effects of cell heterogeneity*, Bull. Math. Biol., 60 (1998), pp. 1167–1200.

[18] M. Deschênes, J. Roy, and M. Steriade, *Thalamic bursting mechanism: An inward slow current revealed by membrane hyperpolarization*, Brain Res., 239 (1982), pp. 289–293.

[19] E. J. Doedel, *AUTO: A program for the automatic bifurcation analysis of autonomous systems*, Congr. Numer., 30 (1981), pp. 265–284.

[20] E. J. Doedel and J. P. Kernevez, *AUTO: Software for Continuation and Bifurcation Problems in Ordinary Differential Equations with Applications*, Report, Department of Applied Mathematics, Caltech, Pasadena, CA, 1986.

[21] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*, SIAM, Philadelphia, 2002.

[22] L. Falke, K. Gillis, D. Pressel, and S. Misler, *Perforated patch recording allows long-term monitoring of metabolite-induced electrical activity and voltage-dependent ca^{2+} currents in pancreatic b cells*, FEBS (Fed. Eur. Biochem. Soc.) Lett., 251 (1989), pp. 167–172.

[23] R. Fitzhugh, *Impulses and physiological states in theoretical models of nerve membrane*, Biophys. J., 1 (1961), pp. 445–466.

[24] T. Gedeon, H. Kokubu, K. Mischaikow, H. Oka, and J. F. Reineck, *The Conley index for fast-slow systems. I. One-dimensional slow variable*, J. Dynam. Differential Equations, 11 (1999), pp. 427–470.

[25] R. E. Griffiths, *Return Map Characterizations of Singular Solutions for a Model of Bursting with Two Slow Variables*, Ph.D. thesis, Montana State University, Bozeman, MT, 2003.

[26] R. Harris-Warrick and R. Flamm, *Multiple mechanism of bursting in a conditioned bursting neuron*, J. Neurosci., 7 (1987), pp. 2113–2128.

[27] D. Himmel and T. Chay, *Theoretical studies on the electrical activity of pancreatic β-cells as a function of glucose*, Biophys. J., 51 (1987), pp. 89–107.

[28] J. Hindmarsh and R. Rose, *A model of neuronal bursting using three coupled first order differential equations*, Proc. Roy. Soc. Lond. Ser. B, 221 (1984), pp. 87–102.

[29] A. Hodgkin and A. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, J. Physiol. (Lond.), 117 (1952), pp. 500–544.
[30] E. Izhikevich, *Neural excitability, spiking, and bursting*, Inter. J. Bifur. Chaos Appl. Sci. Engrg., 10 (2000), pp. 1171–1266.
[31] J. Keizer and G. Magnus, *Atp-sensitive potassium channel and bursting in the pancreatic β cell*, Biophys. J., 56 (1989), pp. 229–242.
[32] J. Keizer and P. Smolen, *Bursting electrical activity in pancreatic β-cells caused by Ca^{2+}- and voltage-inactivated Ca^{2+} channels*, Proc. Nat. Acad. Sci. USA, 88 (1991), pp. 3897–3901.
[33] A. Kepecs and X. Wang, *Analysis of complex bursting in cortical pyramidal neuron models*, Neurocomputing, 32/33 (2000), pp. 181–187.
[34] G. S. Medvedev, *Reduction of a model of an excitable cell to a one-dimensional map*, Phys. D, 202 (2005), pp. 37–59.
[35] M. Nagumo, *A theory of degree of mapping based on infinitesimal analysis*, Amer. J. Math., 73 (1951), pp. 485–496.
[36] M. Pernarowski, *Fast subsystem bifurcations in a slowly varying Liénard system exhibiting bursting*, SIAM J. Appl. Math., 54 (1994), pp. 814–832.
[37] M. Pernarowski, *Fast and slow subsystems for a continuum model of bursting activity in the pancreatic islet*, SIAM J. Appl. Math., 58 (1998), pp. 1667–1687.
[38] M. Pernarowski, *Fast subsystem bifurcations in strongly coupled heterogeneous collections of excitable cells*, Bull. Math. Biol., 62 (2000), pp. 101–120.
[39] M. Pernarowski, R. M. Miura, and J. Kevorkian, *Perturbation techniques for models of bursting electrical activity in the pancreatic β-cell*, SIAM J. Appl. Math., 52 (1992), pp. 1627–1650.
[40] R. Plant and M. Kim, *Mathematical description of a bursting pacemaker neuron by a modification of the Hodgkin-Huxley equations*, Biophys. J., 16 (1976), pp. 227–244.
[41] J. Rinzel, *A formal classification of bursting mechanisms in excitable systems*, in Mathematical Topics in Population Biology, Morphogenesis, and Neurosciences, E. Teramato and M. Yamaguti, eds., Lecture Notes in Biomath. 71, Springer-Verlag, Berlin, 1987, pp. 267–281.
[42] P. Rorsman and G. Trube, *Calcium and delayed potassium currents in mouse pancreatic β-cells under voltage-clamp conditions*, J. Physiol. (Lond.), 374 (1986), pp. 531–550.
[43] A. Sherman, *Anti-phase, asymmetric, and aperiodic oscillations in excitable cells—*I. *Coupled bursters*, Bull. Math. Biol., 56 (1994), pp. 811–835.
[44] A. Sherman, *Case Studies in Mathematical Modeling—Ecology, Physiology, and Cell Biology*, Prentice-Hall, Upper Saddle River, NJ, 1997.
[45] A. Sherman and J. Rinzel, *Model for synchronization of pancreatic β-cells by gap junctions*, Biophys. J., 59 (1991), pp. 547–559.
[46] A. Sherman, J. Rinzel, and J. Keizer, *Emergence of organized bursting in clusters of pancreatic β-cells by channel sharing*, Biophys. J., 54 (1988), pp. 411–425.
[47] A. Shilnikov, R. Calabrese, and G. Cymbalyuk, *Mechanism of bistability: Tonic spiking and bursting in a neuron model*, Phys. Rev. E, 71 (2005), p. 056214.
[48] P. Smith, F. Aschroft, and P. Rorsman, *Simultaneous recordings of glucose dependent electrical activity and atp-regulated k^+-currents in isolated mouse pancreatic β-cells*, FEBS (Fed. Eur. Biochem. Soc.) Lett., 261 (1990), pp. 187–190.
[49] P. Smolen and J. Keizer, *Slow voltage-inactivation of Ca^{2+} currents and bursting mechanisms for the mouse pancreatic β-cell*, J. Memb. Biol., 127 (1992), pp. 9–19.
[50] P. Smolen, J. Rinzel, and A. Sherman, *Why pancreatic islets burst but single β-cells do not: The heterogeneity hypothesis*, Biophys. J., 64 (1993), pp. 1668–1680.
[51] P. Smolen, D. Terman, and J. Rinzel, *Properties of a bursting model with two slow inhibitory variables*, SIAM J. Appl. Math., 53 (1993), pp. 861–892.
[52] F. Strumwasser, *Types of information stored in single neurons*, in Invertebrate Nervous Systems: Their Significance for Mammalian Neurophysiology, C. Wiersma, ed., The University of Chicago Press, Chicago, 1967, pp. 290–319.
[53] D. Terman, *Chaotic spikes arising from a model of bursting in excitable membranes*, SIAM J. Appl. Math., 51 (1991), pp. 1418–1450.
[54] D. Terman, *The transition from bursting to continuous spiking in excitable membrane models*, J. Nonlinear Sci., 2 (1992), pp. 135–182.
[55] X. Wang, *Fast burst firing and short-term synaptic plasticity*: *A model of neocortical chattering neurons*, Neurosci., 89 (1999), pp. 347–362.
[56] R. Wong and D. Prince, *Afterpotential generation in hippocampal pyramidal cells*, J. Neurophysiol., 45 (1981), pp. 86–97.

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 1949–1977

SURFACE TENSION-DRIVEN FLOW IN A SLENDER WEDGE*

J. BILLINGHAM†

Abstract. We consider an inviscid fluid, initially at rest inside a wedge, bounded by one free surface and one solid surface. When $t = 0$, we allow the contact angle to change discontinuously, which leads the free surface to recoil under the action of surface tension. As noted by Keller and Miksis [*SIAM J. Appl. Math.*, 43 (1983), pp. 268–277], a similarity scaling is available, with lengths scaling like $t^{2/3}$. We consider the situation when the wedge is slender, with angle $\epsilon \ll 1$, and the contact angle changes from ϵ to $\lambda\epsilon$. The leading order asymptotic problem for $\lambda = O(1)$, a pair of nonlinear ordinary differential equations, was considered by King [*Quart. J. Mech. Appl. Math.*, 44 (1991), pp. 173–192], numerically for $\lambda = O(1)$ and asymptotically for $|\lambda - 1| \ll 1$. In this paper, we begin by considering this system when $1 \ll \lambda \ll \epsilon^{-1}$, and use Kuzmak's method to construct the asymptotic solution. When $\lambda = O(\epsilon^{-1})$, the slope of the free surface becomes of $O(1)$, and it is no longer possible to reduce the problem to ordinary differential equations alone. However, we can approach this problem in a similar manner, even though the underlying oscillator is the solution of a nonlinear boundary value problem for Laplace's equation, and construct an asymptotic solution. In fact, the solution takes the form of a modulated set of waves on fluid of finite depth, with the underlying analytical solution given by Kinnersley [*J. Fluid Mech.*, 77 (1976), pp. 229–241]. The case $\lambda\epsilon = 90^\circ$ is the solution for the inviscid recoil of a wedge of fluid with two free surfaces and semiangle $\epsilon \ll 1$, which was discussed by Billingham and King [*J. Fluid Mech.*, 533 (2005), pp. 193–221]. We also show that no non-self-intersecting solution is available for $\lambda\epsilon > 90^\circ$ as $\epsilon \to 0$, and compare our asymptotic solutions with numerical, boundary integral solutions of the full, nonlinear free boundary problem.

Key words. fluid mechanics, surface tension, moving contact line, similarity solution, Kuzmak's method, boundary integral method

AMS subject classifications. 35C20, 76B07, 76B45

DOI. 10.1137/05064655X

1. Introduction.

In this paper, we study the response of a wedge of inviscid fluid, initially at rest, bounded by a free surface and a solid surface, to a discontinuous change in the contact angle that it makes with the solid surface. Such a change could be brought about by, for example, an abrupt change in temperature or chemical composition. For a viscous fluid, modelling the motion of this contact line is complicated by the force singularity at the contact line associated with the no slip boundary condition (Dussan and Davis (1974)). Although it is possible to make progress by modifying the no slip boundary condition (Cox (1986)), some unresolved issues remain (see Shikhmurzaev (1997) for a review). However, at sufficiently large times, the solution is likely to asymptote to the inviscid similarity form discussed in this paper, as has been demonstrated in comparable problems (see, for example, Billingham (1999), Billingham and King (2005)).

Inviscid, surface tension-driven flow in a fluid wedge was first studied by Keller and Miksis (1983), who noted that a similarity solution, with lengths scaling like $t^{2/3}$, is available. These scalings have since been used by many other authors to study related problems, for example, Lawrie (1990), Lawrie and King (1994), King

*Received by the editors December 2, 2005; accepted for publication May 30, 2006; published electronically September 21, 2006. This work was supported by the Engineering and Physical Sciences Research Council through an Advanced Research Fellowship.

http://www.siam.org/journals/siap/66-6/64655.html

†School of Mathematical Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK (John.Billingham@Nottingham.ac.uk).

(1991), Billingham and King (1995), King, Billingham, and Popple (1999), Decent and King (2001), Keller, Milewski, and Vanden-Broeck (2000, 2002), and Sierou and Lister (2004). Such flows are also relevant to situations where bodies of fluid rupture and recoil under the action of surface tension. Indeed, for a contact angle of 90°, by symmetry, the problem that we study here is equivalent to the recoil of a wedge of fluid with two free surfaces.

We begin our analysis in section 2, where we derive the nonlinear free boundary problem that governs similarity solutions of the initial value problem. In section 3, we discuss the one-dimensional approximation to the solution that is possible when the new contact angle is sufficiently small. This was first studied by King (1991). We show that if the new contact angle is small, but much larger than the wedge angle, it is possible to construct a solution using Kuzmak's method. The free surface is then a slowly varying nonlinear oscillator. In section 4, we consider the solution when the new contact angle is of $O(1)$. In this case, we can still use Kuzmak's method to solve the problem, but the underlying nonlinear oscillator is the solution of a two-dimensional, nonlinear free boundary problem. We are able to make progress because this free boundary problem has a family of analytical solutions, which was first studied in the context of capillary waves on fluid of finite depth by Kinnersley (1976). We also show that, for ϵ sufficiently small, non-self-intersecting solutions exist only if the contact angle is less than 90°. In section 5, we make a comparison between the asymptotic solution and numerical solutions of the full problem obtained using the boundary integral method.

2. Similarity solution of the initial value problem. We consider the two-dimensional flow of an inviscid fluid, initially at rest inside a wedge of angle ϵ, as shown in Figure 2.1. The fluid is bounded by a solid surface at $y = 0$, while its other surface is free and subject to a constant, uniform surface tension, σ. We denote by D the domain that contains the fluid. Since the flow is initially irrotational, it remains irrotational, and we can describe the flow using a velocity potential ϕ, with the fluid velocity given by $\mathbf{u} = \nabla\phi$. The potential satisfies Laplace's equation

$$\nabla^2\phi = 0 \quad \text{in } D. \tag{2.1}$$

The kinematic and dynamic boundary conditions are

$$\frac{\partial y_s}{\partial t} = \frac{\partial \phi}{\partial y} - \frac{\partial \phi}{\partial x}\frac{\partial y_s}{\partial x} \quad \text{at } y = y_s(x,t), \tag{2.2}$$

$$\frac{\partial \phi}{\partial t} + \frac{1}{2}\left|\nabla\phi\right|^2 = \frac{\sigma}{\rho}\frac{\partial^2 y_s}{\partial x^2}\left\{1 + \left(\frac{\partial y_s}{\partial x}\right)^2\right\}^{3/2} \quad \text{at } y = y_s(x,t). \tag{2.3}$$

There is no normal flux through the solid surface so that

$$\frac{\partial \phi}{\partial y} = 0 \quad \text{at } y = 0. \tag{2.4}$$

The contact angle condition is

$$\frac{\partial y_s}{\partial x} = \tan(\lambda\epsilon) \quad \text{at } y = 0,\ x = x_c(t). \tag{2.5}$$

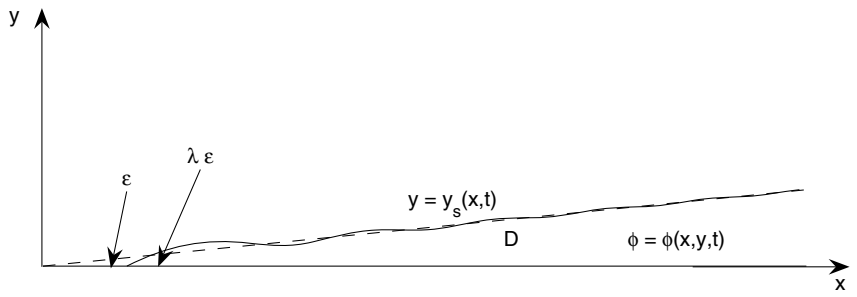


FIG. 2.1. *The initial geometry (broken line) and subsequent recoil.*

The fluid is at rest in the far field so that

$$\phi(x,y,t) \to 0 \quad \text{as } x^2+y^2 \to \infty. \tag{2.6}$$

The initial conditions are

$$\phi = 0, \quad y_s = x\tan\epsilon \quad \text{when } t = 0. \tag{2.7}$$

There is no geometrical lengthscale in this problem, and the only dimensional quantities involved are ρ and σ. As noted by Keller and Miksis (1983), dimensional analysis shows that we can define the independent similarity variables

$$\bar{x} = \frac{\rho^{1/3}x}{\sigma^{1/3}t^{2/3}}, \quad \bar{y} = \frac{\rho^{1/3}y}{\sigma^{1/3}t^{2/3}}, \tag{2.8}$$

and look for a solution of the form

$$\bar{\phi} = \frac{\rho^{2/3}\phi}{\sigma^{2/3}t^{1/3}} \equiv \bar{\phi}(\bar{x},\bar{y}), \quad \bar{y}_s = \frac{\rho^{1/3}y_s}{\sigma^{1/3}t^{2/3}} \equiv \bar{y}_s(\bar{x}). \tag{2.9}$$

In terms of these similarity variables, the initial value problem (2.1)–(2.6) becomes the boundary value problem

$$\nabla^2\bar{\phi} = 0 \quad \text{for } 0 < \bar{y} < \bar{y}_s(\bar{x}),\ \bar{x} > \bar{x}_c, \tag{2.10}$$

$$\frac{1}{3}\bar{\phi} - \frac{2}{3}\left(\bar{x}\frac{\partial\bar{\phi}}{\partial\bar{x}} + \bar{y}\frac{\partial\bar{\phi}}{\partial\bar{y}}\right) + \frac{1}{2}\left|\nabla\bar{\phi}\right|^2 = \frac{\bar{y}_s''}{(1+\bar{y}_s'^2)^{3/2}} \quad \text{at } \bar{y} = \bar{y}_s, \tag{2.11}$$

$$\frac{\partial\bar{\phi}}{\partial\bar{y}} + \frac{2}{3}\bar{x}\bar{y}_s' - \frac{2}{3}\bar{y}_s - \frac{\partial\bar{\phi}}{\partial\bar{x}}\bar{y}_s' = 0 \quad \text{at } \bar{y} = \bar{y}_s, \tag{2.12}$$

$$\frac{\partial\bar{\phi}}{\partial\bar{y}} = 0 \quad \text{at } \bar{y} = 0 \tag{2.13}$$

subject to

$$\bar{y}_s' = \tan(\lambda\epsilon) \quad \text{at } \bar{x} = \bar{x}_c, \tag{2.14}$$

$$\bar{y}_s - \bar{x}\tan\epsilon \to 0 \quad \text{as } \bar{x} \to \infty, \tag{2.15}$$

$$\bar{\phi} \to 0 \quad \text{as } \bar{x}^2+\bar{y}^2 \to \infty, \tag{2.16}$$

where a prime denotes $d/d\bar{x}$ and $\bar{x}_c$ is a constant to be determined.

3. The one-dimensional slender wedge limit, $\epsilon \ll 1$, $\lambda = O(1)$. King (1991) showed that an appropriate scaling in the limit of a slender wedge, $\epsilon \ll 1$, with a small change in the contact angle, $\lambda = O(1)$, is

$$\bar{x} = \epsilon^{1/3}\xi, \quad \bar{y} = \epsilon^{4/3}\eta, \quad \bar{\phi} = \epsilon^{2/3}\Phi, \quad \bar{y}_s = \epsilon^{4/3}Y. \tag{3.1}$$

In terms of these variables, (2.10)–(2.16) become

$$\epsilon^2 \frac{\partial^2 \Phi}{\partial \xi^2} + \frac{\partial^2 \Phi}{\partial \eta^2} = 0 \quad \text{for } 0 < \eta < Y(\xi),\ \xi > \xi_c, \tag{3.2}$$

$$\frac{1}{3}\Phi - \frac{2}{3}\left(\xi\frac{\partial \Phi}{\partial \xi} + \eta\frac{\partial \Phi}{\partial \eta}\right) + \frac{1}{2}\left\{\left(\frac{\partial \Phi}{\partial \xi}\right)^2 + \frac{1}{\epsilon^2}\left(\frac{\partial \Phi}{\partial \eta}\right)^2\right\} = \frac{Y''}{(1+\epsilon^2 Y'^2)^{3/2}} \quad \text{at } \eta = Y, \tag{3.3}$$

$$\frac{1}{\epsilon^2}\frac{\partial \Phi}{\partial \eta} + \frac{2}{3}\xi Y' - \frac{2}{3}Y - \frac{\partial \Phi}{\partial \xi}Y' = 0 \quad \text{at } \eta = Y, \tag{3.4}$$

$$\frac{\partial \Phi}{\partial \eta} = 0 \quad \text{at } \eta = 0 \tag{3.5}$$

subject to

$$Y' = \frac{1}{\epsilon}\tan(\lambda\epsilon) \quad \text{at } \xi = \xi_c, \tag{3.6}$$

$$Y - \xi\frac{\tan\epsilon}{\epsilon} \to 0 \quad \text{as } \xi \to \infty, \tag{3.7}$$

$$\Phi \to 0 \quad \text{as } \xi^2 + \epsilon^2\eta^2 \to \infty. \tag{3.8}$$

We make the expansions

$$\Phi = \Phi_0 + \epsilon^2\Phi_2 + O(\epsilon^4), \quad Y = Y_0 + O(\epsilon^2)$$

and substitute into (3.2)–(3.8). King (1991) showed that $\Phi_0 \equiv \Phi_0(\xi)$ and $\Phi_2 = -\frac{1}{2}\eta^2\Phi_0'' + D(\xi)$. The leading order problem is then a nonlinear boundary value problem for $\Phi_0(\xi)$ and $Y_0(\xi)$. In order to formulate this problem on a known domain, we define $\hat{\xi} = \xi - \xi_c$, and the resulting coupled, nonlinear ordinary differential equations are

$$\frac{1}{3}\Phi_0 - \frac{2}{3}\left(\hat{\xi} + \xi_c\right)\Phi_0' + \frac{1}{2}\Phi_0'^2 - Y_0'' = 0, \tag{3.9}$$

$$-\left(Y_0\Phi_0'\right)' + \frac{2}{3}\left(\hat{\xi} + \xi_c\right)Y_0' - \frac{2}{3}Y_0 = 0, \tag{3.10}$$

to be solved for $\hat{\xi} > 0$ subject to

$$Y_0(0) = 0, \quad Y_0'(0) = \lambda, \tag{3.11}$$

$$Y_0 - \hat{\xi} - \xi_c \to 0, \quad \Phi_0 \to 0 \quad \text{as } \hat{\xi} \to \infty, \tag{3.12}$$

where a prime now denotes $d/d\hat{\xi}$.

King (1991) examined the far field solution of (3.9) and (3.10) and showed that

$$\Phi_0 \sim \frac{F}{\hat{\xi}^2} \exp\left(\frac{4i\hat{\xi}^{3/2}}{9C^{1/2}}\right), \quad Y_0 \sim C\hat{\xi} + \xi_c + \frac{G}{\hat{\xi}^{3/2}} \exp\left(\frac{4i\hat{\xi}^{3/2}}{9C^{1/2}}\right) \quad \text{as } \hat{\xi} \to \infty, \tag{3.13}$$

and, to satisfy (3.12), $C = 1$. The far field solution therefore consists of decaying capillary waves. King (1991) demonstrated that there is a weak nonuniformity in the far field solution when $\eta\hat{\xi}^{1/2} = O(\epsilon^{-1})$ and the flow becomes two-dimensional. Rescaling into a region with $\hat{\xi}$, $\eta = O(\epsilon^{-2/3})$ then completes the solution structure. We will not consider the detailed structure of the two-dimensional flow here but refer the interested reader to King (1991).

3.1. Numerical solution. We can solve (3.9) and (3.10) subject to (3.11) and (3.12) numerically using the MATLAB routine bvp4c, which uses an adaptive gridding method. Since we know that $\Phi_0 = 0$, $Y_0 = \hat{\xi}$, and $\xi_c = 0$ when $\lambda = 1$, it is straightforward to use continuation to find the solution for $\lambda > 1$. However, we did find it convenient to first make the transformation $Y_0 \mapsto \lambda^{4/3} Y_0$, $\Phi_0 \mapsto \lambda^{2/3}\Phi_0$, $\hat{\xi} \mapsto \lambda^{1/3}\hat{\xi}$, $\xi_c \mapsto \lambda^{1/3}\xi_c$, which leaves (3.9)–(3.12) unchanged, except that the first relation in (3.12) becomes $Y_0 - \lambda^{-1}(\hat{\xi} + \xi_c) \to 0$ as $\hat{\xi} \to \infty$. This allowed us to find a numerical solution up to $\lambda \approx 105$, a somewhat larger value than was possible for the system in its original form. Note that it is tempting to suggest that, when $\lambda \gg 1$, we simply need to look for a solution with $Y_0 \to 0$ as $\hat{\xi} \to \infty$. However, the far field behavior, (3.13), shows that no such solution exists, since we cannot take $C = 0$.

Figure 3.1 shows the solution for various values of λ. As λ increases, both the amplitude and the frequency of the oscillations of the free surface increase, as does the size of the potential, Φ_0. Figure 3.2 shows the numerical solution for $\lambda = 100$, along with the asymptotic solution for its envelope, which we will determine in section 3.2. Figure 3.3 shows the numerically calculated behavior of the position of the contact line, ξ_c, as a function of λ. Also shown is the asymptotic behavior for $|\lambda - 1| \ll 1$, $\xi_c \sim 0.80755(\lambda - 1)$, determined by King (1991), and the asymptotic behavior for $\lambda \gg 1$, which we shall determine in section 3.2.

3.2. Asymptotic solution for $\lambda \gg 1$. The numerical solutions presented above suggest that the asymptotic solution for $\lambda \gg 1$ takes the form of a large amplitude oscillation of the free surface, varying slowly over a long lengthscale. This suggests that we can find the solution using Kuzmak's method (see, for example, Bourland and Haberman (1988), and King, Billingham, and Otto (2003)), which is a version of the method of multiple scales that works for nonlinear oscillators.

By looking for an asymptotic balance, we find that appropriate scaled variables are

$$Y_0 = \lambda^{2/3}\bar{Y}(\psi, X), \quad \Phi_0 = \lambda^{4/3}\hat{\Phi}(X) + \lambda^{1/3}\bar{\Phi}(\psi, X), \quad \hat{\xi} = \lambda^{-1/3}\bar{\xi}, \quad \xi_c = \lambda^{2/3}\bar{\xi}_c, \tag{3.14}$$

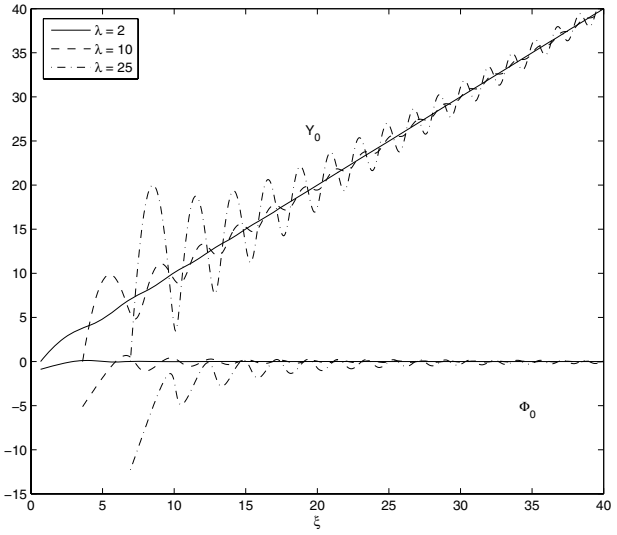


FIG. 3.1. *The numerical solution for* $\lambda = 2$, 10, *and* 25.

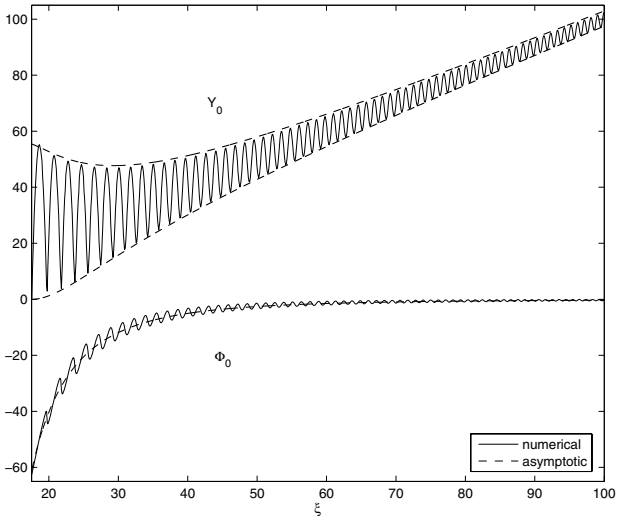


FIG. 3.2. *The numerical solution for* $\lambda = 100$, *along with the asymptotic solution for the envelope of* Y_0 *and the leading order behavior of* Φ_0.

where $X = \lambda^{-1}\bar{\xi}$ is a slow space variable and $\psi = \lambda\theta(X) + p(X)$ is a fast space variable, with $\theta(0) = p(0) = 0$. Note that

$$\frac{d}{d\bar{\xi}} = \left\{\omega(X) + \lambda^{-1}p'(X)\right\}\frac{\partial}{\partial\psi} + \lambda^{-1}\frac{\partial}{\partial X},$$

where $\omega(X) \equiv \theta'(X) = O(1)$ is the frequency of the underlying oscillatory solution, which we will choose so that the solution has unit period in terms of the fast variable,

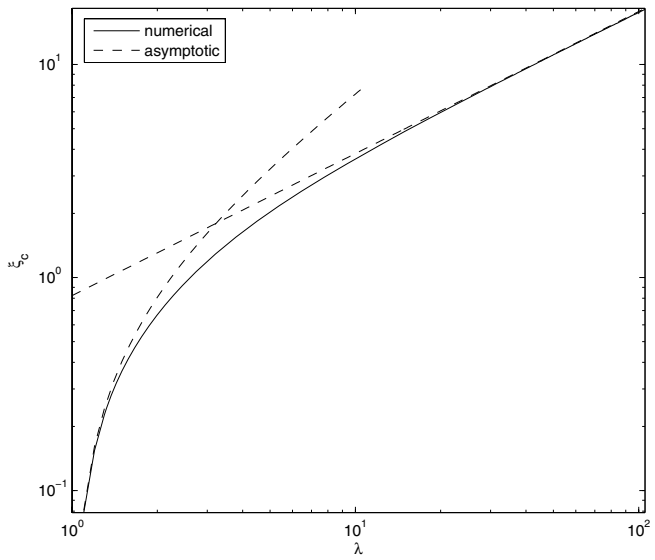


FIG. 3.3. *The position of the contact line, ξ_c, determined numerically for $\lambda > 1$, along with the asymptotic estimates for $|\lambda - 1| \ll 1$ and $\lambda \gg 1$.*

ψ. The function $p(X) = O(1)$ is the phase. In terms of these variables, (3.9)–(3.12) become

$$\frac{1}{3}\left(\hat{\Phi} + \lambda^{-1}\bar{\Phi}\right) - \frac{2}{3}\left(\bar{\xi}_c + X\right)\left\{\left(\omega + \lambda^{-1}p'\right)\bar{\Phi}_\psi + \hat{\Phi}' + \lambda^{-1}\bar{\Phi}_X\right\}$$

$$+\frac{1}{2}\left\{\left(\omega + \lambda^{-1}p'\right)\bar{\Phi}_\psi + \hat{\Phi}' + \lambda^{-1}\bar{\Phi}_X\right\}^2 \tag{3.15}$$

$$-\left\{\left(\omega + \lambda^{-1}p'\right)^2\bar{Y}_{\psi\psi} + 2\lambda^{-1}\left(\omega + \lambda^{-1}p'\right)\bar{Y}_{\psi X} + \lambda^{-2}\bar{Y}_{XX} + \lambda^{-1}\left(\omega' + \lambda^{-1}p''\right)\bar{Y}_\psi\right\} = 0,$$

$$-\bar{Y}\left\{\left(\omega + \lambda^{-1}p'\right)^2\bar{\Phi}_{\psi\psi} + 2\lambda^{-1}\left(\omega + \lambda^{-1}p'\right)\bar{\Phi}_{\psi X}\right.$$

$$\left.+\lambda^{-1}\left(\hat{\Phi}'' + \lambda^{-1}\bar{\Phi}_{XX}\right) + \lambda^{-1}\left(\omega' + \lambda^{-1}p''\right)\bar{\Phi}_\psi\right\} \tag{3.16}$$

$$-\left\{\left(\omega + \lambda^{-1}p'\right)\bar{Y}_\psi + \lambda^{-1}\bar{Y}_X\right\}\left\{\left(\omega + \lambda^{-1}p'\right)\bar{\Phi}_\psi + \hat{\Phi}' + \lambda^{-1}\bar{\Phi}_X\right\}$$

$$+\frac{2}{3}\left(\bar{\xi}_c + X\right)\left\{\left(\omega + \lambda^{-1}p'\right)\bar{Y}_\psi + \lambda^{-1}\bar{Y}_X\right\} - \frac{2}{3}\lambda^{-1}\bar{Y} = 0,$$

to be solved for $\bar{\xi} > 0$, $X > 0$ subject to

$$\bar{Y}(0,0) = 0, \quad (\omega(0) + \lambda^{-1}p'(0))\bar{Y}_\psi(0,0) + \lambda^{-1}\bar{Y}_X(0,0) = 1, \tag{3.17}$$

$$\bar{Y} - \bar{\xi} - \bar{\xi}_c \to 0, \quad \hat{\Phi} + \lambda^{-1}\bar{\Phi} \to 0 \ \text{ as } \bar{\xi} \to \infty. \tag{3.18}$$

We now expand

$$\bar{Y} = \bar{Y}_0 + \lambda^{-1}\bar{Y}_1 + O(\lambda^{-2}), \quad \bar{\Phi} = \bar{\Phi}_0 + \lambda^{-1}\bar{\Phi}_1 + O(\lambda^{-2}), \quad \bar{\xi}_c = \bar{\xi}_{c0} + \lambda^{-1}\bar{\xi}_{c1} + O(\lambda^{-2}).$$

As we shall see, we can find the leading order solution up to some unknown functions of X by solving the equations at leading order but must consider secularity conditions for the equations at $\mathrm{O}(\lambda^{-1})$ to determine the slow time behavior of the leading order solution, with the exception of the phase, $p(X)$. In order to determine the slow drift in the phase, we would have to solve at $O(\lambda^{-1})$ and determine a secularity condition from the equations at $O(\lambda^{-2})$. This proves to be intractable, and we will not address this issue here. The strategy for finding $p(X)$ is described in King, Billingham, and Otto (2003) for a simple model problem.

3.2.1. Solution at leading order. At leading order, (3.15) and (3.16) give

$$-\omega^2\bar{Y}_0\bar{\Phi}_{0\psi\psi} - \omega\bar{Y}_{0\psi}\left(\omega\bar{\Phi}_{0\psi} + \hat{\Phi}'\right) + \frac{2}{3}\left(\bar{\xi}_{c0} + X\right)\omega\bar{Y}_{0\psi} = 0, \tag{3.19}$$

$$\frac{1}{3}\hat{\Phi} - \frac{2}{3}\left(\bar{\xi}_{c0} + X\right)\left(\omega\bar{\Phi}_{0\psi} + \hat{\Phi}'\right) + \frac{1}{2}\left(\omega\bar{\Phi}_{0\psi} + \hat{\Phi}'\right)^2 - \omega^2\bar{Y}_{0\psi\psi} = 0. \tag{3.20}$$

We can integrate (3.19) once to give

$$\omega\bar{\Phi}_{0\psi} = -\hat{\Phi}' + \frac{2}{3}\left(\bar{\xi}_{c0} + X\right) - \frac{d_0(X)}{\bar{Y}_0}, \tag{3.21}$$

where $d_0(X)$ is to be determined. As we shall see, $\bar{Y}_0$ is a periodic function of ψ, whose period we normalize to unity by an appropriate choice of $\omega(X)$. In order that $\bar{\Phi}_0$ remains bounded as $\psi \to \infty$ (the secularity condition), we must choose

$$\hat{\Phi}' = \frac{2}{3}\left(\bar{\xi}_{c0} + X\right) - d_0(X)\overline{\left(\frac{1}{\bar{Y}_0}\right)}, \tag{3.22}$$

where an overbar indicates the mean value over the unit period of the oscillation,

$$\overline{f(\psi)} \equiv \int_0^1 f(\psi)\, d\psi.$$

Substituting (3.22) into (3.21) gives

$$\omega\bar{\Phi}_{0\psi} = d_0(X)\left\{\overline{\left(\frac{1}{\bar{Y}_0}\right)} - \frac{1}{\bar{Y}_0}\right\}. \tag{3.23}$$

Now that we have an expression for $\bar{\Phi}_0$, we can substitute this into (3.20) to give

$$\omega^2\bar{Y}_{0\psi\psi} - \frac{1}{3}\hat{\Phi} + \frac{2}{9}\left(\bar{\xi}_{c0} + X\right)^2 - \frac{d_0^2}{2\bar{Y}_0^2} = 0. \tag{3.24}$$

We can integrate this equation once and write the result as

$$\bar{Y}_{0\psi}^2 = \frac{1}{\omega^2}\left\{\frac{4}{9}\left(\bar{\xi}_{c0} + X\right)^2 - \frac{2}{3}\hat{\Phi}\right\}\frac{\left(\bar{Y}_+ - \bar{Y}_0\right)\left(\bar{Y}_0 - \bar{Y}_-\right)}{\bar{Y}_0}, \tag{3.25}$$

where $\bar{Y}_+(X)$ and $\bar{Y}_-(X)$ are related by

$$\bar{Y}_+\bar{Y}_- = \frac{d_0^2}{\frac{4}{9}\left(\bar{\xi}_{c0}+X\right)^2 - \frac{2}{3}\hat{\Phi}}. \tag{3.26}$$

It is clear from the form of (3.25) that $\bar{Y}_0$ is an oscillatory function of ψ and varies between $\bar{Y}_-$ and $\bar{Y}_+ > \bar{Y}_-$. In fact, we can integrate (3.25) and write the solution for the first half-period of the oscillation in implicit form as (see Byrd and Friedman (1954))

$$\frac{1}{\omega}\left\{\frac{4}{9}\left(\bar{\xi}_{c0}+X\right)^2 - \frac{2}{3}\hat{\Phi}\right\}^{1/2}\psi$$

$$= 2\sqrt{\bar{Y}_+}\left\{\mathbf{E}\left(\sqrt{1-\frac{\bar{Y}_-}{\bar{Y}_+}}\right) - E\left(\sin^{-1}\left(\sqrt{\frac{\bar{Y}_+-\bar{Y}_0}{\bar{Y}_+-\bar{Y}_-}}\right);\sqrt{1-\frac{\bar{Y}_-}{\bar{Y}_+}}\right)\right\}, \tag{3.27}$$

where E is the incomplete elliptic integral of the second kind, and $\mathbf{E}$ is the complete elliptic integral of the second kind. We can now choose $\omega(X)$ so that the period of this oscillation is unity, which gives

$$\omega(X) = \frac{\left\{\frac{4}{9}\left(\bar{\xi}_{c0}+X\right)^2 - \frac{2}{3}\hat{\Phi}\right\}^{1/2}}{4\bar{Y}_+^{1/2}\mathbf{E}\left((1-R)^{1/2}\right)}, \tag{3.28}$$

where $R \equiv \bar{Y}_-/\bar{Y}_+$. We also note that, now that we know $\bar{Y}_0$, exact integral formulas given in Byrd and Friedman (1954) show that

$$\overline{\left(\frac{1}{\bar{Y}_0}\right)} = \frac{1}{\bar{Y}_+}\frac{\mathbf{K}\left((1-R)^{1/2}\right)}{\mathbf{E}\left((1-R)^{1/2}\right)}, \tag{3.29}$$

$$\overline{\bar{Y}_0} = \frac{1}{3}\bar{Y}_+\left\{2\left(1+R\right) - R\frac{\mathbf{K}\left((1-R)^{1/2}\right)}{\mathbf{E}\left((1-R)^{1/2}\right)}\right\}, \tag{3.30}$$

where $\mathbf{K}$ is the complete elliptic integral of the first kind.

Note that, for a given value of the slow time variable, X, $\bar{Y}_0$ is an even, periodic function with unit period and $\bar{Y}_{0\psi} = 0$ at $\psi = 0$. At first sight, this makes it impossible to satisfy the leading order initial condition, given by (3.17) as $\bar{Y}_0 = 0$, $\omega(0)\bar{Y}_{0\psi} = 1$ at $\psi = X = 0$. However, this is only the case if $d_0(0) > 0$. We can see from (3.24) that if $d_0(0) = 0$, $\bar{Y}_0$ is quadratic in ψ, and we can satisfy these initial conditions. A closer examination of the behavior as $X \to 0$ leads to the conditions

$$d_0 \sim \frac{8X}{9\left\{\frac{4}{9}\bar{\xi}_{c0}^2 - \frac{2}{3}\hat{\Phi}(0)\right\}}, \quad \bar{Y}_- \sim d_0^2, \quad \bar{Y}_+ \to \frac{1}{\frac{4}{9}\bar{\xi}_{c0}^2 - \frac{2}{3}\hat{\Phi}(0)} \quad \text{as } X \to 0. \tag{3.31}$$

We should also consider what happens as $X \to \infty$. Since (3.18) shows that $\bar{Y}_0 \sim X + \bar{\xi}_{c0}$, we must have $\bar{Y}_+ \sim \bar{Y}_- \sim X + \bar{\xi}_{c0}$. Then (3.26) and (3.28) show that

$$d_0 \sim \frac{2}{3}\left(X+\bar{\xi}_{c0}\right)^2, \quad \omega \sim \frac{1}{3\pi}\left(X+\bar{\xi}_{c0}\right)^{1/2} \quad \text{as } X \to \infty. \tag{3.32}$$

This means that $\theta(X) \sim 2\left(X+\bar{\xi}_{c0}\right)^{3/2}/9\pi$ and hence that

$$\bar{Y}_0 \sim \bar{\xi}_{c0} + X + \frac{1}{2}\left(\bar{Y}_+ - \bar{Y}_-\right)\cos\left\{\frac{4\lambda}{9}\left(X+\bar{\xi}_{c0}\right)^{3/2} + p'\right\}.$$

This is consistent with (3.13) and shows that we should find

$$\bar{Y}_+ - \bar{Y}_- = O((\bar{\xi}_c + X)^{-3/2}) \quad \text{for } X \gg 1. \tag{3.33}$$

We will confirm this later.

3.2.2. Secularity conditions at $O(\lambda^{-1})$. At $O(\lambda^{-1})$, (3.16) again leads to an equation that we can integrate once, to find that

$$\omega\bar{\Phi}_{1\psi} = -\bar{\Phi}_{0X} + \frac{d_0\bar{Y}_1}{\bar{Y}_0^2} + \frac{d_1}{\bar{Y}_0}$$

$$+\frac{1}{\omega\bar{Y}_0}\left(d_0'\psi - \frac{4}{3}\bar{Z}_0\right) - \frac{p'}{\omega}\left\{\frac{2}{3}\left(\bar{\xi}_{c0}+X\right) - \hat{\Phi}'\right\} + \frac{2}{3}\bar{\xi}_{c1}, \tag{3.34}$$

where $d_1(X)$ is a function of integration and

$$\bar{Z}_0 = \int_0^\psi \bar{Y}_0(\hat{\psi}, X)\, d\hat{\psi}.$$

In order that $\bar{\Phi}_1$ remains bounded as $\psi \to \infty$, we require

$$d_0' = \frac{4}{3}\overline{\bar{Y}_0}. \tag{3.35}$$

Equation (3.15) at $O(\lambda^{-1})$, after using (3.34) to eliminate $\bar{\Phi}_1$, gives

$$\omega^2\bar{Y}_{1\psi\psi} + \frac{d_0^2\bar{Y}_1}{\bar{Y}_0^3} = -2\omega p'\bar{Y}_{0\psi\psi} - 2\omega\bar{Y}_{0\psi X} - \omega'\bar{Y}_{0\psi} + \frac{1}{3}\bar{\Phi}_0$$

$$+\frac{d_0}{\omega\bar{Y}_0^2}\left(d_0p' - \omega d_1 - \frac{4}{3}\overline{\bar{Y}_0}\psi + \frac{4}{3}\bar{Z}_0\right) - \frac{4}{9}\bar{\xi}_{c1}\left(\bar{\xi}_{c0}+X\right). \tag{3.36}$$

Since $\bar{Y}_1 = \bar{Y}_{0\psi}$ is a solution of this equation, the secularity condition is that the integral over one period of $\bar{Y}_{0\psi}$ times the right-hand side should be zero. Taking into account the parity of the various terms on the right-hand side, this means that

$$\int_0^1 \bar{Y}_{0\psi}\left\{-2\omega\bar{Y}_{0\psi X} - \omega'\bar{Y}_{0\psi} + \frac{1}{3}\bar{\Phi}_0 - \frac{4d_0}{3\omega\bar{Y}_0^2}\left(\overline{\bar{Y}_0}\psi - \bar{Z}_0\right)\right\} d\psi = 0. \tag{3.37}$$

After integration by parts, making use of (3.35), we find that

$$\frac{d}{dX}\int_0^1 \omega\bar{Y}_{0\psi}^2\, d\psi = \frac{5d_0}{3\omega}\left\{1 - \overline{\bar{Y}_0}\overline{\left(\frac{1}{\bar{Y}_0}\right)}\right\}. \tag{3.38}$$

From (3.27), we have

$$\int_0^1 \omega \bar{Y}_{0\psi}^2 \, d\psi = \frac{2^{5/2}\bar{Y}_+^{3/2}}{3} \left\{ \frac{2}{9} \left(\bar{\xi}_{c0} + X \right)^2 - \frac{1}{3}\hat{\Phi} \right\}^{1/2} \left\{ (1+R)\, \mathbf{E}\left((1-R)^{1/2} \right) \right.$$

$$\left. -2R\, \mathbf{K}\left((1-R)^{1/2} \right) \right\}. \tag{3.39}$$

Now, making use of (3.22), (3.26), (3.28), (3.35), (3.38), and (3.39), we find, after considerable manipulation, that we can obtain ordinary differential equations for $\hat{\Phi}$, R, and d_0:

$$\hat{\Phi}' = \frac{2}{3}\left(\bar{\xi}_{c0} + X \right) - \sqrt{R} \frac{\mathbf{K}((1-R)^{1/2})}{\mathbf{E}((1-R)^{1/2})} \left\{ \frac{4}{9}\left(\bar{\xi}_{c0} + X \right)^2 - \frac{2}{3}\hat{\Phi} \right\}^{1/2}, \tag{3.40}$$

$$R' = -\frac{4\left(\bar{\xi}_{c0} + X \right)}{27\left\{ \frac{4}{9}\left(\bar{\xi}_{c0} + X \right)^2 - \frac{2}{3}\hat{\Phi} \right\}} \frac{R}{1-R} \left\{ 1 + R - 2R \frac{\mathbf{K}((1-R)^{1/2})}{\mathbf{E}((1-R)^{1/2})} \right\}$$

$$- \frac{2\sqrt{R}}{9\left\{ \frac{4}{9}\left(\bar{\xi}_{c0} + X \right)^2 - \frac{2}{3}\hat{\Phi} \right\}^{1/2}} \frac{1}{1-R} \left\{ R\,(1+R)\, \frac{\mathbf{K}((1-R)^{1/2})}{\mathbf{E}((1-R)^{1/2})} - 2\left(4R^2 - 7R + 4 \right) \right\}, \tag{3.41}$$

$$d_0' = \frac{4 d_0}{9R^{1/2}\left\{ \frac{4}{9}\left(\bar{\xi}_{c0} + X \right)^2 - \frac{2}{3}\hat{\Phi} \right\}^{1/2}} \left\{ 2\,(1+R) - R \frac{\mathbf{K}((1-R)^{1/2})}{\mathbf{E}((1-R)^{1/2})} \right\}. \tag{3.42}$$

Note that, although (3.42) is not coupled to (3.40) and (3.41), it is convenient to consider all three equations together. Equations (3.40)–(3.42) are to be solved subject to

$$R \sim \frac{64X^2}{81\left\{ \frac{4}{9}\bar{\xi}_{c0}^2 - \frac{2}{3}\hat{\Phi}(0) \right\}}, \quad d_0 \sim \frac{8X}{9\left\{ \frac{4}{9}\bar{\xi}_{c0}^2 - \frac{2}{3}\hat{\Phi}(0) \right\}} \quad \text{as } X \to 0, \tag{3.43}$$

$$R \to 1, \quad \hat{\Phi} \to 0, \quad d_0 \sim \frac{2}{3}\left(\bar{\xi}_{c0} + X \right)^2 \quad \text{as } X \to \infty. \tag{3.44}$$

It is now helpful to make this system autonomous by the transformation

$$\hat{\Phi} = \left(\bar{\xi}_{c0} + X \right)^2 \tilde{\Phi}(s), \quad R = \tilde{R}(s), \quad d_0 = \tilde{d}_0(s), \quad s = \log\left(1 + \frac{X}{\bar{\xi}_{c0}} \right),$$

which gives

$$\frac{d\tilde{\Phi}}{ds} = -2\tilde{\Phi} + \frac{2}{3} - \sqrt{\tilde{R}} \frac{\mathbf{K}((1-\tilde{R})^{1/2})}{\mathbf{E}((1-\tilde{R})^{1/2})} \left(\frac{4}{9} - \frac{2}{3}\tilde{\Phi} \right)^{1/2}, \tag{3.45}$$

$$\frac{d\tilde{R}}{ds} = -\frac{4}{27\left(\frac{4}{9}-\frac{2}{3}\tilde{\Phi}\right)}\frac{\tilde{R}}{1-\tilde{R}}\left\{1+\tilde{R}-2\tilde{R}\frac{\mathbf{K}((1-\tilde{R})^{1/2})}{\mathbf{E}((1-\tilde{R})^{1/2})}\right\}$$

$$-\frac{2\sqrt{\tilde{R}}}{9\left(\frac{4}{9}-\frac{2}{3}\tilde{\Phi}\right)^{1/2}}\frac{1}{1-\tilde{R}}\left\{\tilde{R}\left(1+\tilde{R}\right)\frac{\mathbf{K}((1-\tilde{R})^{1/2})}{\mathbf{E}((1-\tilde{R})^{1/2})}-2\left(4\tilde{R}^2-7\tilde{R}+4\right)\right\}, \tag{3.46}$$

$$\frac{d\tilde{d}_0}{ds} = \frac{4\tilde{d}_0}{9\tilde{R}^{1/2}\left(\frac{4}{9}-\frac{2}{3}\tilde{\Phi}\right)^{1/2}}\left\{2\left(1+\tilde{R}\right)-\tilde{R}\frac{\mathbf{K}((1-\tilde{R})^{1/2})}{\mathbf{E}((1-\tilde{R})^{1/2})}\right\}, \tag{3.47}$$

$$\tilde{R}\sim\frac{64s^2}{81\left\{\frac{4}{9}-\frac{2}{3}\tilde{\Phi}(0)\right\}},\quad \tilde{d}_0\sim\frac{8s}{9\bar{\xi}_{c0}\left\{\frac{4}{9}-\frac{2}{3}\tilde{\Phi}(0)\right\}}\ \text{ as } s\to 0, \tag{3.48}$$

$$\tilde{R}\to 1,\quad \tilde{\Phi}=o(e^{-2s}),\quad \tilde{d}_0\sim\frac{2}{3}\bar{\xi}_{c0}^2e^{2s}\ \text{ as } s\to\infty. \tag{3.49}$$

As $\tilde{R}\to 1$ and $\tilde{\Phi}\to 0$, the linearized equations for $\tilde{R}$ and $\tilde{\Phi}$ are

$$\frac{d\tilde{R}}{ds}\sim-\frac{5}{2}\left(\tilde{R}-1\right),\quad \frac{d\tilde{\Phi}}{ds}\sim-\frac{3}{2}\tilde{\Phi}.$$

First, we note that this means that $R=O((\bar{\xi}_{c0}+X)^{-5/2})$ for $X\gg 1$, consistent with (3.33). Second, the point $(0,1)$ in the $(\tilde{\Phi},\tilde{R})$-phase plane is a stable node, and locally all solutions have $\tilde{\Phi}=O(e^{-3s/2})$ as $s\to\infty$, with the exception of the solution associated with the eigenvector in the $\tilde{R}$-direction. Condition (3.49) shows that this is the solution we require. It is straightforward to determine this solution numerically and find where the solution meets the $\tilde{\Phi}$-axis. We used the routine ode45 in MATLAB, which is an implementation of the fifth order Runge–Kutta method with adaptive stepping. We obtain $\tilde{\Phi}(0)\approx -0.1939$. If we then define $\bar{d}_0=\bar{\xi}_{c0}\tilde{d}_0$, we find that $\bar{d}_0$ also satisfies (3.47) and that

$$\bar{d}_0\sim\frac{8s}{9\left\{\frac{4}{9}-\frac{2}{3}\tilde{\Phi}(0)\right\}}\ \text{ as } s\to 0,\ \ \bar{d}_0\sim\frac{2}{3}\bar{\xi}_{c0}^3e^{2s}\ \text{ as } s\to\infty. \tag{3.50}$$

It is again straightforward to solve numerically for $\tilde{R}$, $\tilde{\Phi}$, and $\bar{d}_0$ subject to (3.48) and (3.50), integrating forward until $\bar{d}_0$ is large, and then determine from the behavior of $\bar{d}_0$ as $s\to\infty$ that $\bar{\xi}_{c0}\approx 0.8231$ from (3.50) when $s\gg 1$. Figure 3.3 confirms that $\xi_c\sim\bar{\xi}_{c0}\lambda^{1/3}$ as $\lambda\to\infty$. Figure 3.4 shows how R, $\hat{\Phi}$, and d_0 vary with s. It is now straightforward to calculate $\bar{Y}_+$ and $\bar{Y}_-$ from (3.26). These provide an envelope for the solution and are shown in Figure 3.2 to be in excellent agreement with the numerical solution of the full problem when $\lambda=100$. Figure 3.2 also shows that the asymptotic solution for $\hat{\Phi}(X)$ is in good agreement with the numerical solution. Note that the oscillatory part of Φ_0 has amplitude of $O(\lambda^{-1})$ smaller than the slowly varying part and is given by $\lambda^{-1}\bar{\Phi}$. Figure 3.5 shows a direct comparison of the numerical and asymptotic solutions for $\lambda=100$. The agreement is good in terms of the amplitude, as we would expect from the results shown in Figure 3.2, but we can see a slow drift of the phase as ξ increases. This is because we have taken $p(X)=0$ in the definition of the fast space variable. As discussed earlier, to determine $p(X)$ we would need to solve completely at $O(\lambda^{-1})$ and examine the secularity condition at $O(\lambda^{-2})$, which is not tractable.

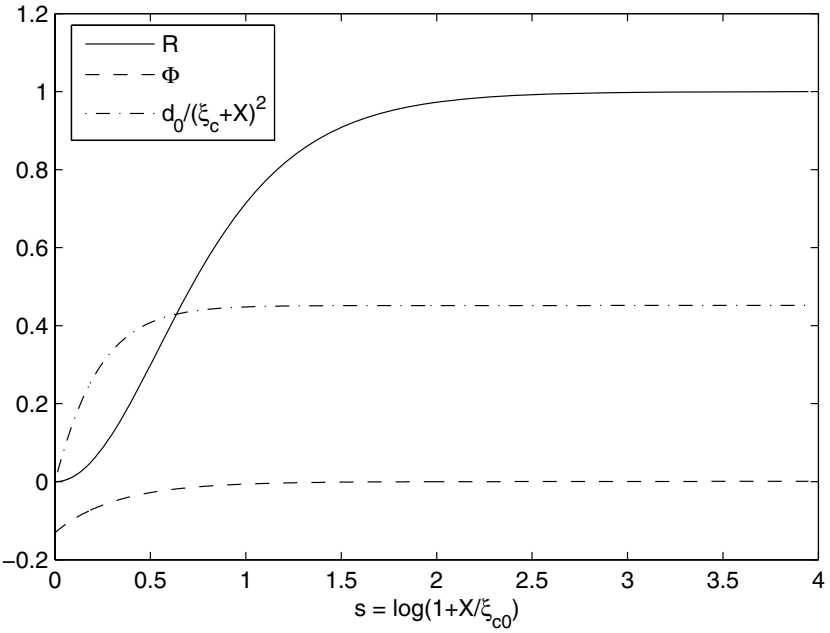


FIG. 3.4. *The variation of R, $\hat{\Phi}$, and d_0 with s.*

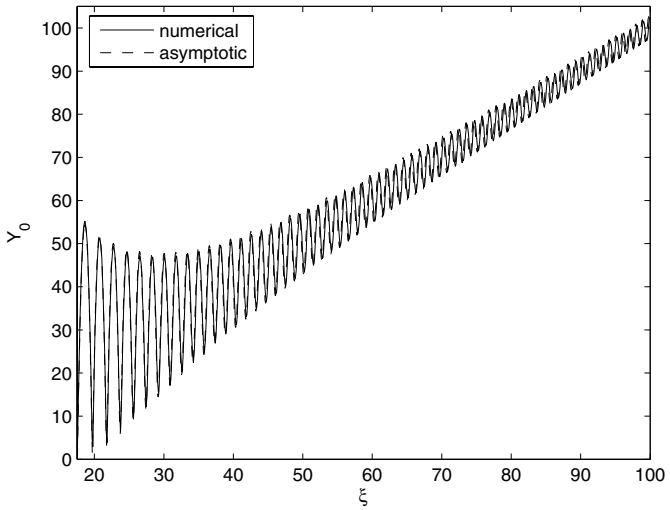


FIG. 3.5. *The numerical and asymptotic solutions for $\lambda = 100$.*

4. The fully nonlinear slender wedge limit, $\epsilon \ll 1$, $\lambda = O(\epsilon^{-1})$. We expect the asymptotic solution that we constructed in the previous section to remain valid for $\lambda = o(\epsilon^{-1})$. When $\lambda = O(\epsilon^{-1})$, the slope of the free surface is $\lambda\epsilon = O(1)$ at the contact line, so we would not expect to be able to reduce the problem to a set

of ordinary differential equations. We define $\bar{\lambda} = \lambda\epsilon = O(1)$ for $\epsilon \ll 1$. The scaled variables (3.1) and (3.14) then suggest that we define new scaled variables

$$\bar{x} = \epsilon^{-1/3}\tilde{\xi}_c + \epsilon^{2/3}\tilde{\xi}, \quad \bar{y} = \epsilon^{2/3}\tilde{\eta}, \quad \bar{y}_s = \epsilon^{2/3}\tilde{Y}, \quad \bar{\phi} = \epsilon^{-2/3}\Phi_0 + \epsilon^{1/3}\tilde{\Phi}. \tag{4.1}$$

Apart from the large shift of the origin, the spatial scalings are those given by King (1991) for the far field of the problem with $\lambda = O(1)$. Indeed, we can think of $\lambda = O(\epsilon^{-1})$ as the limit in which the two-dimensional flow in the far field becomes comparable with the one-dimensional flow in the near field and the new scalings (4.1) emerge.

4.1. The underlying periodic solution. Although we will be using Kuzmak's method to solve the problem in terms of the variables (4.1) for $\epsilon \ll 1$, it will make things clearer if we study the underlying periodic solution first. For the moment, we assume that Φ_0 is a constant, but we will see later that it is a function of the slow variable.

At leading order, a solution of wavelength L in the $\tilde{\xi}$-direction satisfies

$$\nabla^2\tilde{\Phi} = 0 \quad \text{for } 0 < \tilde{\eta} < \tilde{Y},\ 0 < \tilde{\xi} < L, \tag{4.2}$$

$$\frac{1}{3}\Phi_0 - \frac{2}{3}\tilde{\xi}_c\frac{\partial\tilde{\Phi}}{\partial\tilde{\xi}} + \frac{1}{2}|\nabla\tilde{\Phi}|^2 = \frac{\tilde{Y}''}{(1+\tilde{Y}'^2)^{3/2}} \quad \text{at } \tilde{\eta} = \tilde{Y} \text{ for } 0 < \tilde{\xi} < L, \tag{4.3}$$

$$\frac{\partial\tilde{\Phi}}{\partial\tilde{\eta}} + \frac{2}{3}\tilde{\xi}_c\tilde{Y}' - \frac{\partial\tilde{\Phi}}{\partial\tilde{\xi}}\tilde{Y}' = 0 \quad \text{at } \tilde{\eta} = \tilde{Y} \text{ for } 0 < \tilde{\xi} < L, \tag{4.4}$$

$$\frac{\partial\tilde{\Phi}}{\partial\tilde{\eta}} = 0 \quad \text{at } \tilde{\eta} = 0 \text{ for } 0 < \tilde{\xi} < L. \tag{4.5}$$

If we now subtract off a uniform flow in the $\tilde{\xi}$-direction by defining

$$\tilde{\Phi} = \bar{\Phi} + \frac{2}{3}\tilde{\xi}_c\tilde{\xi} + P, \tag{4.6}$$

where P is an arbitrary constant (later, a function of the slow variable), (4.2)–(4.5) become

$$\nabla^2\bar{\Phi} = 0 \quad \text{for } 0 < \tilde{\eta} < \tilde{Y},\ 0 < \tilde{\xi} < L, \tag{4.7}$$

$$\frac{1}{2}|\nabla\bar{\Phi}|^2 = \frac{1}{2}c^2 + \frac{\tilde{Y}''}{(1+\tilde{Y}'^2)^{3/2}} \quad \text{at } \tilde{\eta} = \tilde{Y} \text{ for } 0 < \tilde{\xi} < L, \tag{4.8}$$

$$\frac{\partial\bar{\Phi}}{\partial\tilde{\eta}} - \frac{\partial\bar{\Phi}}{\partial\tilde{\xi}}\tilde{Y}' = 0 \quad \text{at } \tilde{\eta} = \tilde{Y} \text{ for } 0 < \tilde{\xi} < L, \tag{4.9}$$

$$\frac{\partial\bar{\Phi}}{\partial\tilde{\eta}} = 0 \quad \text{at } \tilde{\eta} = 0 \text{ for } 0 < \tilde{\xi} < L, \tag{4.10}$$

where $c = \sqrt{\frac{4}{9}\tilde{\xi}_c^2 - \frac{2}{3}\Phi_0}$. Equation (4.8) is the standard Bernoulli equation, and (4.9) is the equation for no flux through the free surface. We have therefore reduced the system to that of periodic capillary waves on a finite layer of fluid with wavespeed c. A similar approach was used by Billingham and King (2005) for the related problem of flow external to a thin, wedge-shaped void, although the resulting equations had no solution. We have more success here, since there exists a remarkable analytical solution describing capillary waves on fluid of finite depth first elucidated by Kinnersley (1976), building on the work of Crapper (1957), and put into a more systematic framework using complex variable theory by Crowdy (1999). As we shall see when we use Kuzmak's method, the solution of the fully nonlinear wedge problem can therefore be described in terms of capillary waves on fluid of finite depth, modulated in wavelength and amplitude on a long lengthscale.

Case I described by Kinnersley (1976) is the solution of relevance to us. The detailed investigation of this case given in Kinnersley (1976) spares us a lot of hard work, since each of the three limiting cases that he studied is one that we need to understand for our purposes. The solution is given implicitly in terms of Jacobian elliptic functions as

$$\tilde{\xi} = \frac{\mathrm{sn}B\,\mathrm{cd}B}{c^2}\left\{\frac{2k(1-k^2)\,\mathrm{sd}\phi\,\mathrm{nd}\phi}{\mathrm{dn}\psi + k\,\mathrm{cd}\phi} + (1-k^2)\phi - 2E(\phi) + 2k^2\,\mathrm{sn}\phi\,\mathrm{cd}\phi\right\}, \tag{4.11}$$

$$\tilde{\eta} = \frac{\mathrm{sn}B\,\mathrm{cd}B}{c^2}\left\{\frac{2(1-k^2)\,\mathrm{sn}\psi\,\mathrm{cn}\psi}{\mathrm{dn}\psi + k\,\mathrm{cd}\phi} + (1+k^2)\psi - 2E(\psi)\right\}, \tag{4.12}$$

where $\phi = A\bar{\Phi}$, $\psi = A\Psi + B$, Ψ is the streamfunction, $A \equiv c/(1-k^2)\,\mathrm{sn}B\,\mathrm{cd}B$, and B and k parameterize the solution. Note that we have adopted the convention used by Kinnersley (1976), that $\mathrm{sn}\phi \equiv \mathrm{sn}(\phi, k)$, $\mathrm{sn}\psi \equiv \mathrm{sn}(\psi, \sqrt{1-k^2})$, $\mathrm{sn}B \equiv \mathrm{sn}(B, \sqrt{1-k^2})$, and similarly for the other Jacobian elliptic functions, and

$$E(\phi) \equiv E\left(\sin^{-1}(\,\mathrm{sn}\phi)\,; k\right), \quad E(B) \equiv E\left(\sin^{-1}(\,\mathrm{sn}B)\,; \sqrt{1-k^2}\right).$$

It is a feature of the solution that each streamline represents a possible position of the free surface. We will take the free surface to be given by the streamline with $\psi = B$. At the solid surface, $\tilde{\eta} = 0$, we have $\psi = 0$. This solution has period $4\mathbf{K}(k)$ in terms of ϕ, which gives the wavelength as

$$L = \frac{4\,\mathrm{sn}B\,\mathrm{cd}B}{c^2}\left\{2\mathbf{E}(k) - (1-k^2)\mathbf{K}(k)\right\}.$$

We also note that the peak-to-trough amplitude of the solution is

$$a = \frac{4k\,\mathrm{sn}B\,\mathrm{sd}B}{c^2}, \tag{4.13}$$

and the height of the wave, given by the depth of water at the trough, is

$$h = \frac{\mathrm{sn}B\,\mathrm{cd}B}{c^2}\left\{2\,\mathrm{sc}B(\,\mathrm{dn}B - k) + (1+k^2)B - 2E(B)\right\}. \tag{4.14}$$

Figure 4.1(a) shows the streamlines for $B = 2$ and $k = 0.25$ over a single wavelength ($-4\mathbf{K}(k) \le \phi \le 0$). Note that for larger values of B, the free surface becomes self-intersecting and does not represent a valid solution.

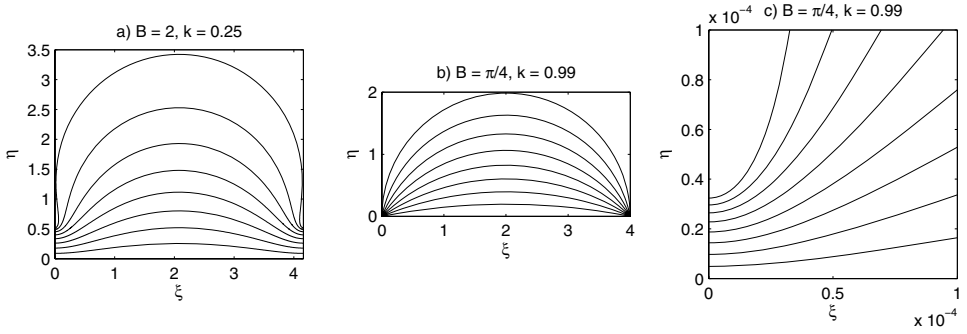


FIG. 4.1. *Some solutions given by* (4.11) *and* (4.12)*, all with* $c = 1$*. The free surface and bulk streamlines, each of which is also a possible position of the free surface, are shown.*

4.2. Limiting cases. We need to understand the behavior of the solution in each of the three limits $k \to 0$, $k \to 1$, and $B \to 0$. Fortunately, these are precisely the limits studied by Kinnersley (1976), so we will briefly summarize his results here.

4.2.1. The linear capillary wave limit, $k \to 0$. As $k \to 0$ with $B = O(1)$, the solution takes the form of small amplitude capillary waves. The free surface is given by

$$Y \sim \frac{\tanh B}{c^2}\left\{B - 2k \sinh B \cos\left(\frac{c^2\tilde{\xi}}{\tanh B}\right)\right\},$$

with

$$L \sim \frac{2\pi \tanh B}{c^2}, \quad a \sim \frac{4k \tanh B \sinh B}{c^2}, \quad h \sim \frac{B \tanh B}{c^2}.$$

This form of the solution will appear in the far field when we use Kumak's method.

4.2.2. The "string of beads" limit, $k \to 1$. As $k \to 1$, the solution takes the form of a sequence of "beads" of fluid bounded by segments of ellipses, as shown in Figure 4.1(b), connected by small inner "neck" regions, with size of $O((1-k)^2)$, as shown in Figure 4.1(c). We will not repeat the analysis of these two regions presented by Kinnersley (1976), but note that, at leading order, the free surface in the outer region (the "beads") is given by

$$\left(\frac{2\tilde{\xi}}{L}\right)^2 = \left(1 - \frac{\tilde{\eta}}{a}\right)\left(1 + \frac{\tilde{\eta}}{a}\tan^2 B\right),$$

with

$$a \sim \frac{4\sin^2 B}{c^2}, \quad L \sim \frac{8 \sin B \cos B}{c^2}.$$

This form of solution allows us to satisfy the contact angle boundary condition when we use Kuzmak's method. In particular, the free surface makes an angle $2B$ with the $\tilde{\xi}$-axis, and a non-self-intersecting solution exists only for $B \leq \pi/4$, with the free surface becoming semicircular when $B = \pi/4$. This is a point of some significance, to which we shall return later.

4.2.3. The shallow water limit, $B \to 0$. As $B \to 0$,

$$L \sim \frac{4B}{c^2}\left(2\mathbf{E}(k) - (1-k^2)\mathbf{K}(k)\right), \quad a \sim \frac{4B^2 k}{c^2}, \quad h \sim \frac{B^2(1-k)^2}{c^2},$$

and the free surface is given by

$$\tilde{\xi} \sim \frac{B}{c^2}\left\{(1-k^2)\phi - 2E(\phi) + 2k\,\mathrm{sn}\phi\right\}, \tag{4.15}$$

$$Y \sim \frac{B^2}{c^2}\left(\,\mathrm{dn}\phi - k\,\mathrm{cn}\phi\right)^2. \tag{4.16}$$

We can see that, although a and h tend to zero as $B \to 0$, $a/h \sim 4k/(1-k)^2 = O(1)$, so that the slope is finite. This is the limit corresponding to shallow water theory. More importantly for us, this is the limit in which we should recover the solution (3.27) for $\lambda = O(1)$ that we constructed in section 3, which takes the form of a single equation. This suggests that we should be able to reduce (4.15) and (4.16) to the same form. We can do this using a Landen transformation, as suggested by Crowdy (1999). The identities

$$\mathrm{dn}\phi - k\,\mathrm{cn}\phi \equiv (1-k)\,\mathrm{nd}\left(\frac{1}{2}(1+k)\phi; \frac{2k^{1/2}}{1+k}\right),$$

$$\mathrm{sn}\phi \equiv \frac{2}{1+k}\,\mathrm{sn}\left(\frac{1}{2}(1+k)\phi; \frac{2k^{1/2}}{1+k}\right)\,\mathrm{cn}\left(\frac{1}{2}(1+k)\phi; \frac{2k^{1/2}}{1+k}\right)\,\mathrm{nd}\left(\frac{1}{2}(1+k)\phi; \frac{2k^{1/2}}{1+k}\right),$$

$$2E(\phi) - (1-k^2)\phi + 2k\,\mathrm{sn}\phi \equiv 2(1+k)E\left(\sin^{-1}\left\{\,\mathrm{sn}\left(\frac{1}{2}(1+k)\phi; \frac{2k^{1/2}}{1+k}\right)\right\}; \frac{2k^{1/2}}{1+k}\right),$$

which follow from the Landen transformation (see Byrd and Friedman (1954)), show that

$$\tilde{\xi} \sim \frac{B}{c^2}\left\{\frac{8k}{1+k}\,\mathrm{sn}\left(\frac{1}{2}(1+k)\phi; \frac{2k^{1/2}}{1+k}\right)\,\mathrm{cn}\left(\frac{1}{2}(1+k)\phi; \frac{2k^{1/2}}{1+k}\right)\,\mathrm{nd}\left(\frac{1}{2}(1+k)\phi; \frac{2k^{1/2}}{1+k}\right)\right.$$

$$\left. -2(1+k)E\left(\sin^{-1}\left\{\,\mathrm{sn}\left(\frac{1}{2}(1+k)\phi; \frac{2k^{1/2}}{1+k}\right)\right\}; \frac{2k^{1/2}}{1+k}\right)\right\},$$

$$\tilde{Y} \sim \frac{B^2}{c^2}(1-k)^2\,\mathrm{nd}^2\left(\frac{1}{2}(1+k)\phi; \frac{2k^{1/2}}{1+k}\right).$$

This allows us to eliminate ϕ and obtain

$$\tilde{\xi} \sim \frac{2}{c}\left\{\sqrt{\frac{(\tilde{Y}-\tilde{Y}_-)(\tilde{Y}_+-\tilde{Y})}{\tilde{Y}}} - \tilde{Y}_+^{1/2}E\left\{\sin^{-1}\left(\frac{\tilde{Y}_+^{1/2}(\tilde{Y}-\tilde{Y}_-)^{1/2}}{\tilde{Y}^{1/2}(\tilde{Y}_+-\tilde{Y}_-)^{1/2}}\right); \sqrt{1-\frac{\tilde{Y}_-}{\tilde{Y}_+}}\right\}\right\},$$

where

$$\tilde{Y}_+ = \frac{B^2}{c^2}(1+k)^2, \quad \tilde{Y}_- = \frac{B^2}{c^2}(1-k)^2$$

are the maximum and minimum values of $\tilde{Y}$. Finally, the addition formula (117.0.1) in Byrd and Friedman (1954) shows that

$$\tilde{\xi} \sim -\frac{2\tilde{Y}_+^{1/2}}{c^2}\left[\mathbf{E}\left(\sqrt{1-\frac{\tilde{Y}_-}{\tilde{Y}_+}}\right) - E\left\{\sin^{-1}\left(\sqrt{\frac{\tilde{Y}_+-\tilde{Y}}{\tilde{Y}_+-\tilde{Y}_-}}\right); \sqrt{1-\frac{\tilde{Y}_-}{\tilde{Y}_+}}\right\}\right].$$

This has the same form as (3.27), the solution for $\lambda = O(1)$, with the minus sign arising because we have taken $\phi < 0$ here.

4.3. Solution using Kuzmak's method. We define slow and fast space variables as

$$X = \epsilon\tilde{\xi}, \quad x = \epsilon^{-1}\theta(X) + p(X), \quad y = \tilde{\eta}, \quad p(0) = \theta(0) = 0.$$

We have used x and y to tidy up our notation, and these should not be confused with the original variables x and y. We seek a solution of the form

$$Y = Y(x, X), \quad \Phi_0 = \Phi_0(X), \quad \tilde{\Phi} = \tilde{\Phi}(x, y, X),$$

where we have dropped the tildes from the spatial variables, again for notational convenience. In terms of these variables, (2.10)–(2.16) become, after first using (4.1),

$$(\omega+\epsilon p')^2\frac{\partial^2\tilde{\Phi}}{\partial x^2} + \frac{\partial^2\tilde{\Phi}}{\partial y^2} + 2\epsilon(\omega+\epsilon p')\frac{\partial^2\tilde{\Phi}}{\partial x\partial X}$$

$$+\epsilon(\omega'+\epsilon p'')\frac{\partial\tilde{\Phi}}{\partial x} + \epsilon^2\frac{\partial^2\tilde{\Phi}}{\partial X^2} + \epsilon\Phi_0'' = 0 \quad \text{for } x > 0,\ 0 < y < Y, \tag{4.17}$$

$$\frac{1}{3}(\Phi_0+\epsilon\tilde{\Phi}) - \frac{2}{3}\left[(\xi_c+X)\left\{(\omega+\epsilon p')\frac{\partial\tilde{\Phi}}{\partial x} + \epsilon\frac{\partial\tilde{\Phi}}{\partial X} + \Phi_0'\right\} + \epsilon y\frac{\partial\tilde{\Phi}}{\partial y}\right]$$

$$+\frac{1}{2}\left[\left(\frac{\partial\tilde{\Phi}}{\partial y}\right)^2 + \left\{(\omega+\epsilon p')\frac{\partial\tilde{\Phi}}{\partial x} + \epsilon\frac{\partial\tilde{\Phi}}{\partial X} + \Phi_0'\right\}^2\right]$$

$$= \left[1 + \left\{(\omega+\epsilon p')\frac{\partial Y}{\partial x} + \epsilon\frac{\partial Y}{\partial X}\right\}^2\right]^{-3/2} \tag{4.18}$$

$$\times\left[(\omega+\epsilon p')^2\frac{\partial^2 Y}{\partial x^2} + 2\epsilon(\omega+\epsilon p')\frac{\partial^2 Y}{\partial x\partial X} + \epsilon(\omega'+\epsilon p'')\frac{\partial Y}{\partial x} + \epsilon^2\frac{\partial^2 Y}{\partial X^2}\right] \text{ at } y = Y \text{ for } x > 0,$$

$$\frac{\partial\tilde{\Phi}}{\partial y} + \frac{2}{3}(\xi_c+X)\left\{(\omega+\epsilon p')\frac{\partial Y}{\partial x} + \epsilon\frac{\partial Y}{\partial X}\right\} - \frac{2}{3}\epsilon Y$$

(4.19)
$$-\left\{(\omega+\epsilon p')\frac{\partial Y}{\partial x}+\epsilon\frac{\partial Y}{\partial X}\right\}\left\{(\omega+\epsilon p')\frac{\partial\tilde{\Phi}}{\partial x}+\epsilon\frac{\partial\tilde{\Phi}}{\partial X}+\Phi_0'\right\}=0 \quad \text{at } y=Y \text{ for } x>0,$$

$$\frac{\partial\tilde{\Phi}}{\partial y}=0 \quad \text{at } y=0 \text{ for } x>0, \tag{4.20}$$

where $\omega(X)\equiv\theta'(X)$, subject to

$$Y(0,0)=0, \quad (\omega(0)+\epsilon p'(0))\frac{\partial Y}{\partial x}(0,0)+\epsilon\frac{\partial Y}{\partial X}(0,0)=\tan\bar{\lambda}, \tag{4.21}$$

$$Y-\frac{\tan\epsilon}{\epsilon}(\xi_c+X)\to 0, \quad \Phi_0+\epsilon\tilde{\Phi}\to 0 \quad \text{as } X\to\infty. \tag{4.22}$$

We now expand

$$\tilde{\Phi}=\tilde{\Phi}_1(x,y,X)+\epsilon\tilde{\Phi}_2(x,y,X)+O(\epsilon^2), \quad Y=Y_1(x,X)+\epsilon Y_2(x,X)+O(\epsilon^2),$$

$$\xi_c=\xi_{c0}+\epsilon\xi_{c1}+O(\epsilon^2), \quad P=P_1(X)+\epsilon P_2(X)+O(\epsilon^2),$$

where $P(X)$ appears in the transformation (4.6). At leading order, we obtain equations governing the underlying periodic solution that we discussed in the previous section, but here x is scaled with ω. The solution is

$$\frac{x}{\omega}=\frac{\mathrm{sn}B\,\mathrm{cd}B}{c^2}\left\{\frac{2k(1-k^2)\,\mathrm{sd}\phi\,\mathrm{nd}\phi}{\mathrm{dn}\psi+k\,\mathrm{cd}\phi}+(1-k^2)\phi-2E(\phi)+2k^2\,\mathrm{sn}\phi\,\mathrm{cd}\phi\right\}, \tag{4.23}$$

$$y=\frac{\mathrm{sn}B\,\mathrm{cd}B}{c^2}\left\{\frac{2(1-k^2)\,\mathrm{sn}\psi\,\mathrm{cn}\psi}{\mathrm{dn}\psi+k\,\mathrm{cd}\phi}+(1+k^2)\psi-2E(\psi)\right\}, \tag{4.24}$$

where $\phi=A\bar{\Phi}_1$, $\psi=A\Psi+B$, $A=c/(1-k^2)\,\mathrm{sn}B\,\mathrm{cd}B$, $c(X)=\sqrt{\frac{4}{9}(\xi_{c0}+X)^2-\frac{2}{3}\Phi_0}$, and $B(X)$ and $k(X)$ vary on the slow lengthscale. We have also used (4.6) so that

$$\tilde{\Phi}_1=\bar{\Phi}_1+\frac{1}{\omega}\left\{\frac{2}{3}(\xi_{c0}+X)-\Phi_0'\right\}x+P_1(X). \tag{4.25}$$

In order that this solution has unit wavelength, we choose

$$\omega(X)=\frac{\frac{4}{9}(\xi_{c0}+X)^2-\frac{2}{3}\Phi_0}{4\,\mathrm{sn}B\,\mathrm{cd}B\,\{2\mathbf{E}(k)-(1-k^2)\mathbf{K}(k)\}}. \tag{4.26}$$

In order to satisfy the contact angle condition, the solution must take the "string of beads" form discussed in section 4.2.2 so that

$$k\to 1, \quad B\to\frac{1}{2}\bar{\lambda} \quad \text{as } X\to 0. \tag{4.27}$$

In the far field, the amplitude of the disturbance of the free surface must decay to zero; thus we need the linear capillary wave form of the solution, discussed in section 4.2.1, so that

$$\Phi_0\to 0, \quad k\to 0, \quad B\sim\frac{4}{9}(\xi_{c0}+X)^3 \quad \text{as } X\to\infty. \tag{4.28}$$

4.3.1. Secularity conditions. We now have the functional form of the leading order solution but still need to find the ordinary differential equations that determine $\Phi_0(X)$, $k(X)$, $B(X)$, and the eigenvalue, ξ_{c0}. This is the same situation that arose in section 3, and we proceed in the same manner to determine secularity conditions.

Now, $\tilde{\Phi}_1$ must be periodic and not grow linearly with x. Since $\bar{\Phi}_1$ changes by $-4\mathbf{K}(k)/A$ over a period, (4.25) shows that the first secularity condition is

$$\frac{1}{\omega}\left\{\frac{2}{3}(\xi_{c0}+X)-\Phi_0'\right\}-\frac{4\mathbf{K}(k)(1-k^2)\,\mathrm{sn}B\,\mathrm{cd}B}{c}=0.$$

From the definitions of $c(X)$ and $\omega(X)$, this gives the ordinary differential equation for Φ_0,

$$\Phi_0'=\frac{2}{3}(\xi_{c0}+X)-\frac{\mathbf{K}(k)(1-k^2)\sqrt{\frac{4}{9}(\xi_{c0}+X)^2-\frac{2}{3}\Phi_0}}{2\mathbf{E}(k)-(1-k^2)\mathbf{K}(k)}. \tag{4.29}$$

In the limit $B\to 0$, we can show that (4.29) reduces to (3.40) by using a Gauss transformation (Byrd and Friedman (1954)).

In order to obtain the two remaining secularity conditions, we need to consider the problem at $O(\epsilon)$. Over one wavelength of the leading order solution, the field equation and boundary conditions can be written in the form

$$\nabla^2\tilde{\Phi}_2=F(x,y,X)\ \text{ for } 0<\bar{x}<\frac{1}{\omega},0<y<Y_1, \tag{4.30}$$

$$L_1(\tilde{\Phi}_2,\bar{Y}_2)\equiv\frac{\partial\tilde{\Phi}_2}{\partial s}-\frac{1}{\partial\bar{\Phi}_1/\partial\bar{x}}\frac{\partial}{\partial s}\left(\frac{\partial Y_2/\partial s}{1+(\partial Y_1/\partial\bar{x})^2}\right)$$

$$-\frac{\partial}{\partial y}\left\{\sqrt{\left(\frac{\partial\bar{\Phi}_1}{\partial\bar{x}}\right)^2+\left(\frac{\partial\bar{\Phi}_1}{\partial y}\right)^2}\right\}Y_2=g(s,X)\ \text{ at } y=Y_1 \text{ for } 0<\bar{x}<\frac{1}{\omega}, \tag{4.31}$$

$$L_2(\tilde{\Phi}_2,\bar{Y}_2)\equiv\mathbf{n}\cdot\nabla\tilde{\Phi}_2-\frac{\partial}{\partial s}\left(Y_2\frac{\partial\bar{\Phi}_1}{\partial\bar{x}}\right)=f(s,X)\ \text{ at } y=Y_1 \text{ for } 0<\bar{x}<\frac{1}{\omega}, \tag{4.32}$$

$$L_3(\tilde{\Phi}_2,\bar{Y}_2)\equiv\frac{\partial\tilde{\Phi}_2}{\partial y}=0\ \text{ at } y=0 \text{ for } 0<\bar{x}<\frac{1}{\omega}, \tag{4.33}$$

where $\bar{x}\equiv x/\omega$, $\mathbf{n}$ is the outward unit normal at the boundary, and s measures arc length along the free surface. The forcing functions F, f, and g are, recalling the definition (4.25) of $\tilde{\Phi}_1$,

$$F(\bar{x},y,X)\equiv-\frac{2p'}{\omega}\frac{\partial^2\tilde{\Phi}_1}{\partial\bar{x}^2}-2\frac{\partial^2\tilde{\Phi}_1}{\partial\bar{x}\partial X}-\frac{\omega'}{\omega}\frac{\partial\tilde{\Phi}_1}{\partial\bar{x}}-\Phi_0'', \tag{4.34}$$

$$f(s,X)\equiv\left\{1+\left(\frac{\partial Y_1}{\partial\bar{x}}\right)^2\right\}^{-1/2}\left\{\frac{\partial\bar{\Phi}_1}{\partial\bar{x}}\left(\frac{p'}{\omega}\frac{\partial Y_1}{\partial\bar{x}}+\frac{\partial Y_1}{\partial X}\right)\right.$$

$$-\frac{2}{3}\xi_{c1}\frac{\partial Y_1}{\partial \bar{x}}+\frac{2}{3}Y_1+\frac{\partial Y_1}{\partial \bar{x}}\left(\frac{p'}{\omega}\frac{\partial \tilde{\Phi}_1}{\partial \bar{x}}+\frac{\partial \tilde{\Phi}_1}{\partial X}\right)\Bigg\}, \tag{4.35}$$

$$g(s,X)\equiv -\frac{1}{\partial\bar{\Phi}_1/\partial\bar{x}}\left\{1+\left(\frac{\partial Y_1}{\partial \bar{x}}\right)^2\right\}^{-1/2}\Bigg[-\frac{1}{3}\tilde{\Phi}_1-\frac{\partial\bar{\Phi}_1}{\partial\bar{x}}\left(\frac{p'}{\omega}\frac{\partial\tilde{\Phi}_1}{\partial\bar{x}}+\frac{\partial\tilde{\Phi}_1}{\partial X}\right)$$

$$+\frac{2}{3}\xi_{c1}\left(\frac{\partial\tilde{\Phi}_1}{\partial\bar{x}}+\Phi_0'\right)+\frac{2}{3}Y_1\frac{\partial\tilde{\Phi}_1}{\partial y}-3\frac{\partial Y_1}{\partial\bar{x}}\frac{\partial^2 Y_1}{\partial\bar{x}^2}\left(\frac{p'}{\omega}\frac{\partial Y_1}{\partial\bar{x}}+\frac{\partial Y_1}{\partial X}\right)\left\{1+\left(\frac{\partial Y_1}{\partial\bar{x}}\right)^2\right\}^{-5/2} \tag{4.36}$$

$$+\left(2\frac{p'}{\omega}\frac{\partial^2 Y_1}{\partial\bar{x}^2}+2\frac{\partial^2 Y_1}{\partial\bar{x}\partial X}+\frac{\omega'}{\omega}\frac{\partial Y_1}{\partial\bar{x}}\right)\left\{1+\left(\frac{\partial Y_1}{\partial\bar{x}}\right)^2\right\}^{-3/2}\Bigg].$$

Now, let $\tilde{\Phi}_2=G(\bar{x},y,X)$, $Y_2=H(\bar{x},y,X)$ be a solution of the unforced problem, $\nabla^2 G=0$, $L_1(G,H)=L_2(G,H)=L_3(G,H)=0$. Green's theorem shows that

$$\int\int_D\left(G\nabla^2\tilde{\Phi}_2-\tilde{\Phi}_2\nabla^2 G\right)dA\equiv\int_{\partial D}\left(G\,\mathbf{n}\cdot\nabla\tilde{\Phi}_2-\tilde{\Phi}_2\,\mathbf{n}\cdot\nabla G\right)ds,$$

where D is the domain of solution and ∂D its boundary. Using (4.30)–(4.33) and integrating by parts twice, we arrive at

$$\int\int_D G\,F(\bar{x},y,X)\,d\bar{x}\,dy=\int_0^l\left\{G\,f(s,X)-H\frac{\partial\bar{\Phi}_1}{\partial\bar{x}}g(s,X)\right\}ds,$$

where l is the length of the free surface.

One obvious solution of the unforced problem is $G=1$, $H=0$, which gives

$$\int\int_D F(\bar{x},y,X)\,d\bar{x}\,dy=\int_0^l f(s,X)\,ds. \tag{4.37}$$

Another solution, as we would expect since these equations govern the correction to the leading order solution, is $G=\partial\bar{\Phi}_1/\partial\bar{x}$, $H=\partial Y_1/\partial\bar{x}$, which gives

$$\int\int_D\frac{\partial\bar{\Phi}_1}{\partial\bar{x}}F(\bar{x},y,X)\,d\bar{x}\,dy=\int_0^l\left\{\frac{\partial\bar{\Phi}_1}{\partial\bar{x}}f(s,X)-\frac{\partial Y_1}{\partial\bar{x}}\frac{\partial\bar{\Phi}_1}{\partial\bar{x}}g(s,X)\right\}ds. \tag{4.38}$$

As we shall see, (4.37) and (4.38) provide us with the two remaining ordinary differential equations for $k(X)$ and $B(X)$.

If we substitute (4.34) and (4.35) into (4.37) and use the fact that $\tilde{\Phi}_1$ is odd and Y_1 even in $\bar{x}$, noting that we are able to reduce the double integral to a surface integral, we arrive at

$$\frac{d}{dX}\int_0^1\omega Y_1\frac{\partial\bar{\Phi}_1}{\partial x}dx=-\frac{4}{3}\int_0^1 Y_1\,dx. \tag{4.39}$$

This is consistent with the solution that we constructed in section 3, since, using the current notation, $d_0\equiv-\omega Y_1\partial\bar{\Phi}_1/\partial x$, which we recall is a function of X alone for $\bar{\lambda}\ll 1$, so that (4.39) is equivalent to (3.35).

Similarly, (4.38) leads to

$$\frac{d}{dX}\int\int_D \omega\left(\frac{\partial\bar\Phi_1}{\partial x}\right)^2 dx\,dy = \left[\frac{\omega'}{\omega}\left\{\frac{2}{3}(\xi_{c0}+X)-\Phi_0'\right\}-\frac{7}{3}+\Phi_0''\right]\int_0^1 Y_1\frac{\partial\bar\Phi_1}{\partial x}dx$$

$$-\int_0^1 \frac{\partial Y_1}{\partial x}\left[\frac{2}{3}\omega^2 Y_1\frac{\partial\bar\Phi_1}{\partial x}\frac{\partial Y_1}{\partial x}-3\omega^3\frac{\partial Y_1}{\partial x}\frac{\partial^2 Y_1}{\partial x^2}\frac{\partial Y_1}{\partial X}\left\{1+\omega^2\left(\frac{\partial Y_1}{\partial x}\right)^2\right\}^{-5/2}\right. \tag{4.40}$$

$$\left.+\left(2\omega\frac{\partial^2 Y_1}{\partial x\partial X}+\omega'\frac{\partial Y_1}{\partial x}\right)\left\{1+\omega^2\left(\frac{\partial Y_1}{\partial x}\right)^2\right\}^{-3/2}\right]dx-\frac{1}{3\omega}\left\{\frac{2}{3}(\xi_{c0}+X)-\Phi_0'\right\}\int_0^1 Y_1\,dx.$$

We can show that this is consistent with (3.38) by first noting that when $B\ll 1$ (which corresponds to $\bar\lambda\ll 1$), $Y_1=O(B^2)$, $\bar\Phi_1=O(B)$, and $\omega=O(B^{-1})$, so that (4.40) becomes

$$\frac{d}{dX}\int\int_D \omega\left(\frac{\partial\bar\Phi_1}{\partial x}\right)^2 dx\,dy \sim \left[\frac{\omega'}{\omega}\left\{\frac{2}{3}(\xi_{c0}+X)-\Phi_0'\right\}-\frac{7}{3}+\Phi_0''\right]\int_0^1 Y_1\frac{\partial\bar\Phi_1}{\partial x}dx$$

$$-\int_0^1 \frac{\partial Y_1}{\partial x}\left[2\omega\frac{\partial^2 Y_1}{\partial x\partial X}+\omega'\frac{\partial Y_1}{\partial x}\right]dx-\frac{1}{3\omega}\left\{\frac{2}{3}(\xi_{c0}+X)-\Phi_0'\right\}\int_0^1 Y_1\,dx.$$

From the definition of d_0, we find that

$$\frac{d}{dX}\int\int_D \omega\left(\frac{\partial\bar\Phi_1}{\partial x}\right)^2 dx\,dy \sim \frac{d}{dX}\int_0^1 \omega Y_1\left(\frac{\partial\bar\Phi_1}{\partial x}\right)^2 dx \sim -\frac{d}{dX}\left(d_0\int_0^1\frac{\partial\bar\Phi_1}{\partial x}\,dx\right)$$

$$\sim\frac{4}{3\omega}\left\{\frac{2}{3}(\xi_{c0}+X)-\Phi_0'\right\}\int_0^1 Y_1\,dx+\frac{\omega'}{\omega}Y_1\frac{\partial\bar\Phi_1}{\partial x}\left\{\frac{2}{3}(\xi_{c0}+X)-\Phi_0'\right\}-Y_1\frac{\partial\bar\Phi_1}{\partial x}\left(\frac{2}{3}-\Phi_0''\right),$$

and combining these equations gives

$$\frac{d}{dX}\int_0^1\left\{\omega\left(\frac{\partial Y_1}{\partial x}\right)^2\right\}dx\sim\frac{5d_0}{3\omega}\left[1-\frac{1}{d_0}\left\{\frac{2}{3}(\xi_{c0}+X)-\Phi_0'\right\}\int_0^1 Y_1\,dx\right].$$

Equation (3.22) then gives us (3.38), as required.

In order to determine k' and B' from (4.39) and (4.40), we need to use the solution (4.23) and (4.24) for $\bar\Phi_1$ and Y_1. The resulting integrals can be evaluated only numerically. We must also be very careful in evaluating the derivatives, since (4.23) and (4.24) give x and y as functions of $\bar\Phi$ and Ψ. For example, after integrating (4.39) by parts and noting that $\phi=-4\mathbf{K}(k)$ at $x=1$ and $\phi=0$ at $x=0$, we can write

$$-\int_{-4\mathbf{K}(k)}^0\left\{\frac{\omega}{A}\left.\frac{\partial Y_1}{\partial X}\right|_x-\frac{d}{dX}\left(\frac{\omega}{A}\right)\phi\left.\frac{\partial Y_1}{\partial\phi}\right|_X-\frac{\omega}{A}\left.\frac{\partial\phi}{\partial X}\right|_x\left.\frac{\partial Y_1}{\partial\phi}\right|_X+\frac{4}{3}Y_1\left.\frac{\partial x_1}{\partial\phi}\right|_X\right\}d\phi$$

$$=4\frac{d}{dX}\left(\frac{\mathbf{K}(k)\omega}{A}\right),$$

where $x_1(\phi, X)$ is the value of x on the free surface and the subscripts to the derivative indicate the variable to be held constant. We then use

$$\left.\frac{\partial \phi}{\partial X}\right|_x = -\left.\frac{\partial x_1}{\partial X}\right|_\phi \Big/ \left.\frac{\partial x_1}{\partial \phi}\right|_X, \quad \left.\frac{\partial Y_1}{\partial X}\right|_x = \left.\frac{\partial Y_1}{\partial X}\right|_\phi - \left.\frac{\partial x_1}{\partial X}\right|_\phi \left.\frac{\partial Y_1}{\partial \phi}\right|_X \Big/ \left.\frac{\partial x_1}{\partial \phi}\right|_X,$$

$$\left.\frac{\partial \phi}{\partial x}\right|_X = \left.\frac{\partial x_1}{\partial \phi}\right|_X \Big/ \left\{\omega^2 \left(\left.\frac{\partial Y_1}{\partial \phi}\right|_X\right)^2 + \left(\left.\frac{\partial x_1}{\partial \phi}\right|_X\right)^2\right\}$$

to write the derivatives in a form for which we can use (4.23) and (4.24), and rewrite the X-derivatives using

$$\left.\frac{\partial}{\partial X}\right|_\phi = k' \left.\frac{\partial}{\partial k}\right|_{\phi,B,c} + B' \left.\frac{\partial}{\partial B}\right|_{\phi,k,c} + c' \left.\frac{\partial}{\partial c}\right|_{\phi,k,B},$$

noting that

$$c' = c^{-1}\left\{\frac{2}{9}(\xi_{c0} + X) + \frac{4\mathbf{K}(k)\omega}{3A}\right\}.$$

In this way, we can write (4.39) as a linear equation in k' and B'. A similar but algebraically more complicated procedure reduces (4.40) to a linear equation, and we can thereby extract k' and B'.

We used Mathematica to determine the integrands and simplify them as far as possible. It is worth noting that Mathematica is not very good at simplifying expressions involving Jacobian elliptic functions. We found that an effective strategy was to delete the definitions of all the Jacobian elliptic functions except sn and cn, and teach Mathematica that $\mathrm{sn}^2(u;k) + \mathrm{cn}^2(u;k) \equiv 1$. Once the integrands had been simplified, we then replaced obvious expressions (e.g., $\sqrt{1 - k^2\,\mathrm{sn}^2(u;k)} = \mathrm{dn}(u;k)$), and cut and pasted the integrands, which run to several printed pages, into MATLAB. In this way, we could write a routine to evaluate Φ_0', k', and B' as functions of Φ_0, k, B, and X, using quadl and dblquad to evaluate the necessary single and double integrals. Note that our knowledge of the asymptotic behavior of (4.39) and (4.40) as $B \to 0$ proved invaluable in debugging the code.

In order to find the solution, we must solve a nonlinear eigenvalue problem, with $\Phi_0(0)$ and ξ_{c0} the two eigenvalues. In this case, there is no convenient transformation that will make the equations autonomous. We solve the three ordinary differential equations for Φ_0, k, and B with the MATLAB routine ode45. The initial conditions are given by (4.27), and the conditions (4.28) must be satisfied as $X \to \infty$. We find that the solution automatically satisfies $k \to 0$ as $X \to \infty$. However, there is a problem when $k \ll 1$, since both (4.39) and (4.40) give the leading order equation

$$B' = \frac{B \tanh B \left(c' + \frac{4}{3}\right)}{c\,(\tanh B + B\,\mathrm{sech}^2 B)}, \tag{4.41}$$

with the equation for k' given by the $O(k^2)$ correction. This means that our method of calculating k' becomes very inaccurate when k is small. However, when $k \ll 1$, the solution takes the form of small amplitude capillary waves on a flat free surface. We do not, therefore, need to calculate the amplitude of these waves, for which we would need to know k, in order to determine whether the solution satisfies (4.28). We simply need to solve (4.29) and (4.41) once k becomes sufficiently small (we used $k < 10^{-3}$).

We determined the eigenvalues $\Phi_0(0)$ and ξ_{c0} using a shooting method and Newtonian iteration, which typically converged after one cycle, calculating the Jacobian by finite differences. We used the asymptotic solution

$$\Phi_0(0) \sim -0.1314\bar{\lambda}^{4/3}, \quad \xi_{c0} \sim 0.8231\bar{\lambda}^{2/3} \quad \text{as } \bar{\lambda} \to 0 \tag{4.42}$$

as the initial guess for $\bar{\lambda} \ll 1$, and then continuation for larger values of $\bar{\lambda}$, scaling the previous values with $\bar{\lambda}^{4/3}$ and $\bar{\lambda}^{2/3}$, respectively, to improve the rate of convergence.

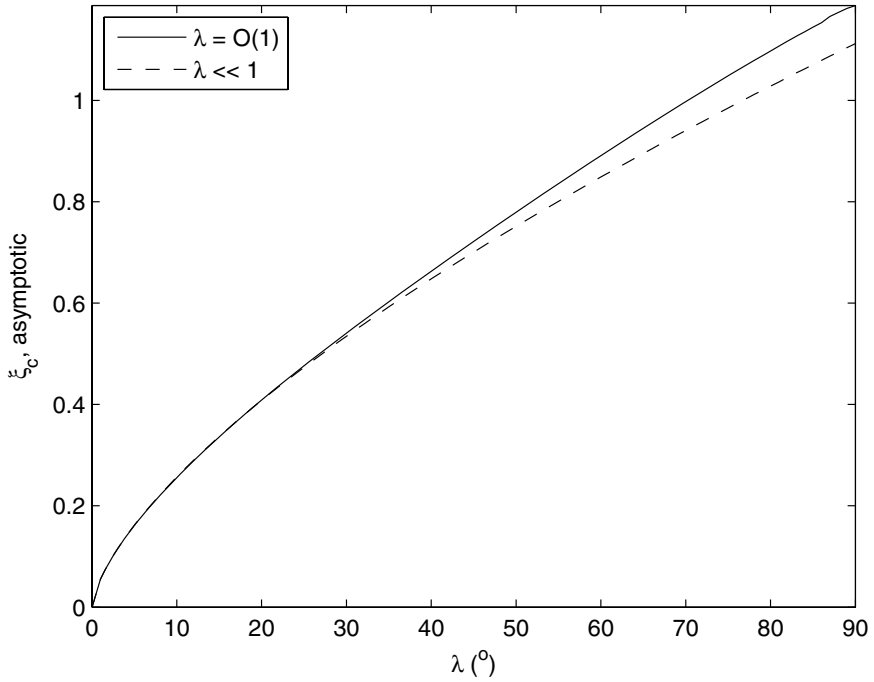


FIG. 4.2. *The position of the tip of the wedge for* $\epsilon \ll 1$ *and* $\bar{\lambda} = O(1)$ *(solid line) and* $\bar{\lambda} \ll 1$ *(broken line).*

Figure 4.2 shows the leading order asymptotic estimate of the position of the tip of the wedge, ξ_{c0}, both for $\bar{\lambda} = O(1)$, as calculated using Kuzmak's method described above, and for $\bar{\lambda} \ll 1$, as given by (4.42). We can see that there is little difference between these two curves and that they are in excellent agreement with each other for contact angles less than about 30°.

5. Boundary integral solutions. We would now like to compare the asymptotic solution that we have constructed for $\bar{\lambda} = O(1)$ and $\epsilon \ll 1$ with the solution of the full problem. In order to solve the full nonlinear boundary value problem given by (2.10)–(2.16) numerically, we use the boundary integral method.

5.1. Numerical method. The approach that we have used is a development of the method described in Billingham and King (2005), and we refer the interested reader to this paper for full details. The free surface was discretized using straight line elements and lies at $(x, y) = (X(s), Y(s))$, where s is arc length. For the boundary

integral equation, the potential, $\bar{\phi}$, was assumed to vary linearly along each element, and the integral along each element then calculated analytically. The arc length equation, $X_s^2 + Y_s^2 = 1$, was evaluated on each element using central differences. Derivatives in the dynamic boundary condition (2.11) were evaluated using a four-point finite difference approximation. The resulting system of nonlinear algebraic equations was solved using Newton's method and continuation, reusing the Jacobian as often as possible. We started with $\epsilon = \bar{\lambda}$, for which the solution is $Y = X \tan\epsilon$, $\bar{\phi} = 0$, and gradually decreased ϵ, using the known solution as the initial estimate of the new solution.

In contrast to the method used by Billingham and King (2005), we have introduced two new features. First, we have been able to calculate the Jacobian used in Newton's method analytically, which speeds up our calculations to the point where the LU decomposition of the Jacobian is the most costly step of the algorithm. Second, we used adaptive gridding, increasing the density of grid points close to the high curvature, "neck" regions of the solution. For each new value of ϵ, we changed the length of the element at the first five local minima of the free surface to be small enough to resolve the local curvature, and allowed the element length to grow slowly with distance from each "neck," up to a maximum of 0.01. For $s > 10$, we allowed the length of the elements to gradually increase to 0.25, and extended the domain of solution to $s = 15$.

In this manner we were able to compute solutions until the maximum curvature was about 200. We were unable to resolve more highly curved solutions with the computing resources available to us. However, we note from our asymptotic solution that for $1 - k \ll 1$, the "neck" region (see Figure 4.1) has size of $O((1-k)^2)$ and therefore curvature of $O((1-k)^{-2})$. At the first "neck," $1-k = O(\epsilon)$ for $\epsilon \ll 1$, so that the curvature is of $O(\epsilon^{-2})$ and therefore, in terms of the original variable, which we use to calculate the numerical solution, the curvature is of $O(\epsilon^{-8/3})$. The minimum element length must therefore scale with $\epsilon^{8/3}$ as $\epsilon \to 0$. Moreover, as ϵ decreases, the number of "necks" at which the curvature is high and scales in the same manner is of $O(\epsilon^{-1})$. That we struggle to resolve this highly multiscale solution numerically is therefore not surprising and suggests that our asymptotic method is the correct way to attack the problem.

5.2. Comparison of numerical and asymptotic solutions. Figure 5.1 shows the position of the free surface and the potential at the free surface for $\bar{\lambda} = 45°$ and $\epsilon = 0.0105$. This is the smallest value of ϵ for which we could obtain a numerical solution. Recall that the leading order asymptotic solution for $\bar{\phi}$, shown here, is a smoothly increasing function; the oscillatory part is given by the next term in the asymptotic expansion. We note that the agreement between the asymptotic prediction and numerical calculation of the position of the tip of the wedge and the free surface and potential close to the tip is not perfect, although the period and amplitude are reasonably well predicted.

Figure 5.2 shows the same functions for $\bar{\lambda} = 90°$. Since this corresponds to the case of a recoiling wedge with two free surfaces, we have also plotted the reflection of y in the x-axis. It is easy to see that the region next to the tip is approximately circular and that successive "beads" are approximately elliptical, as predicted. However, the disagreement between numerical and asymptotic solutions is more marked in this case, although, again, the amplitude and period of the oscillations are reasonably well predicted.

We can go some way to explaining this discrepancy by noting that ϵ is unlikely to

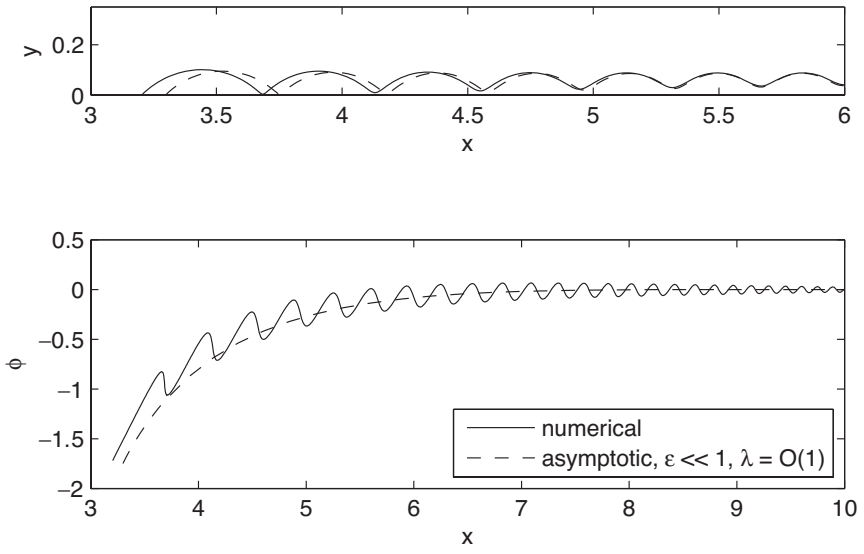


FIG. 5.1. *The position of the free surface and the potential at the free surface calculated numerically for* $\bar{\lambda} = 45^\circ$ *and* $\epsilon = 0.0105$. *Also shown is the asymptotic prediction for* $\epsilon \ll 1$ *and* $\bar{\lambda} = O(1)$.

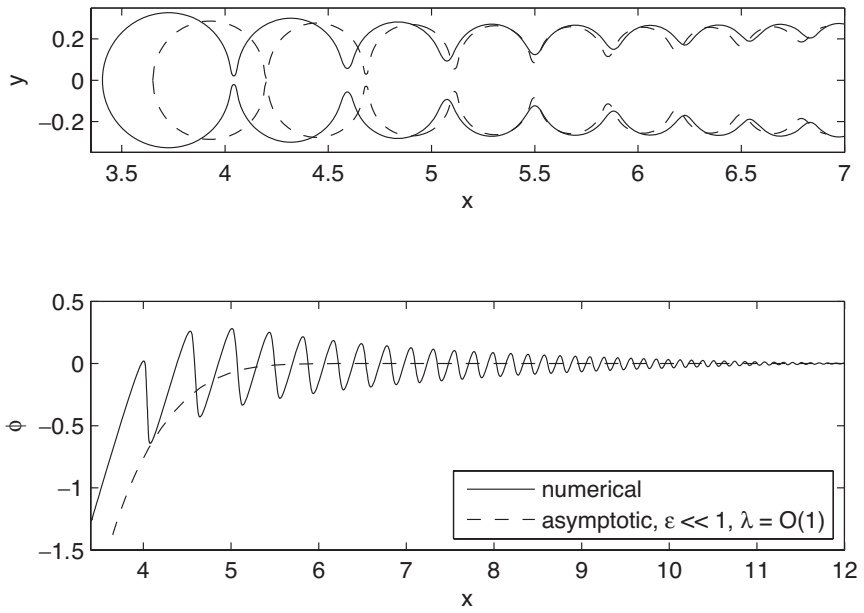


FIG. 5.2. *The position of the free surface and the potential at the free surface calculated numerically for* $\bar{\lambda} = 90^\circ$ *and* $\epsilon = 0.0344$. *Also shown is the asymptotic prediction for* $\epsilon \ll 1$ *and* $\bar{\lambda} = O(1)$.

be small enough to produce good agreement. Note, in particular, that the oscillations in the potential at the free surface shown in Figure 5.2 are of an amplitude comparable to the mean value—an indication that the asymptotic form has yet to be reached. However, there is a very curious feature of these solutions that is rather harder to explain. Figure 5.3 shows the position of the tip of the wedge, $X(0)$, as a function

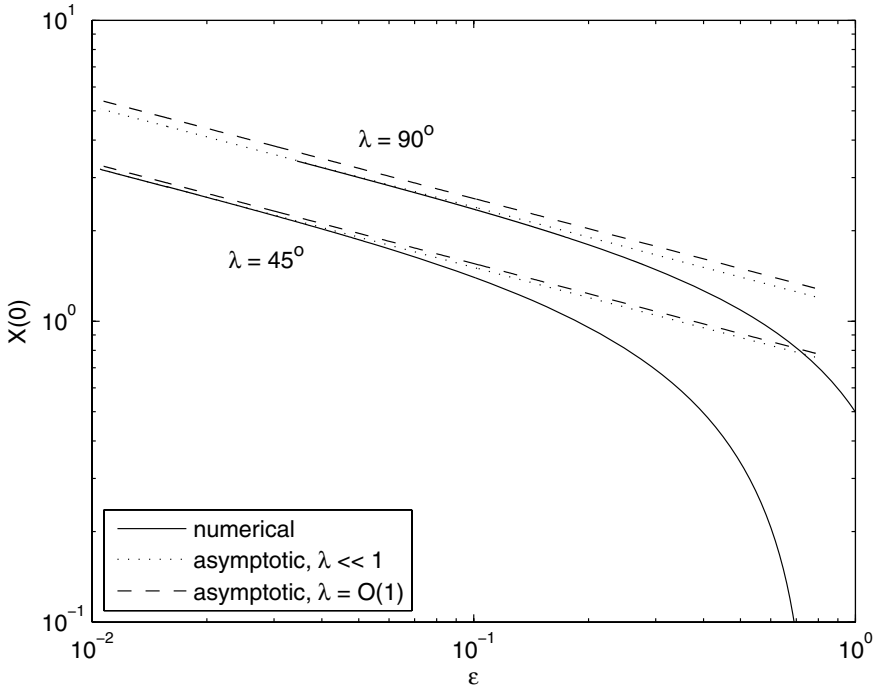


FIG. 5.3. *The position of the tip of the wedge calculated numerically for* $\bar{\lambda} = 45°$ *and* $90°$ *(solid lines). Also shown are the asymptotic predictions for* $\epsilon \ll 1$ *and* $\bar{\lambda} = O(1)$ *(broken lines) and* $\bar{\lambda} \ll 1$ *(dotted lines).*

of ϵ for $\bar{\lambda} = 45°$ and $90°$. Although $X(0)$ appears to asymptote to a multiple of $\epsilon^{-1/3}$ as expected, it is the leading order estimate that comes from the asymptotic analysis presented in section 3, valid for $\bar{\lambda} \ll 1$, namely $0.8231\bar{\lambda}^{2/3}\epsilon^{-1/3}$, that is in best agreement with the numerical solution. As we can see from Figure 4.2, the leading order estimate of the coefficient valid for $\bar{\lambda} = O(1)$ is slightly larger than $0.8231\bar{\lambda}^{2/3}$. This is all the more puzzling because the asymptotic solution for $\bar{\lambda} \ll 1$, although we have not shown it here, bears no resemblance to the numerical solution, except that the position of its tip accurately predicts the numerically calculated position. If our numerical solution of the full problem is correct and there is some error in our asymptotic analysis when $\bar{\lambda} = O(1)$, it is hard to see how the correct asymptotic position of the tip of the wedge could be of the similarity form, scaling with $\bar{\lambda}$, for all $\bar{\lambda} \leq 90°$. Conversely, it is hard to see how an error in our implementation of the numerical solution of the full problem could result in this similarity form for the position of the tip of the wedge. For the moment, our best explanation is that, for smaller ϵ than we can at present access numerically, the numerical solution will start to approach the asymptotic solution, although we cannot say that this is a fully satisfactory explanation.

Finally, we note that another prediction of our analysis is that there is no asymptotic solution for any $\bar{\lambda} > 90°$ for sufficiently small ϵ. This can be confirmed by solving the full problem numerically for a moderately small value of ϵ. Figure 5.4 shows some numerical solutions of the full problem with $\epsilon = 5°$. As $\bar{\lambda}$ increases, the first two

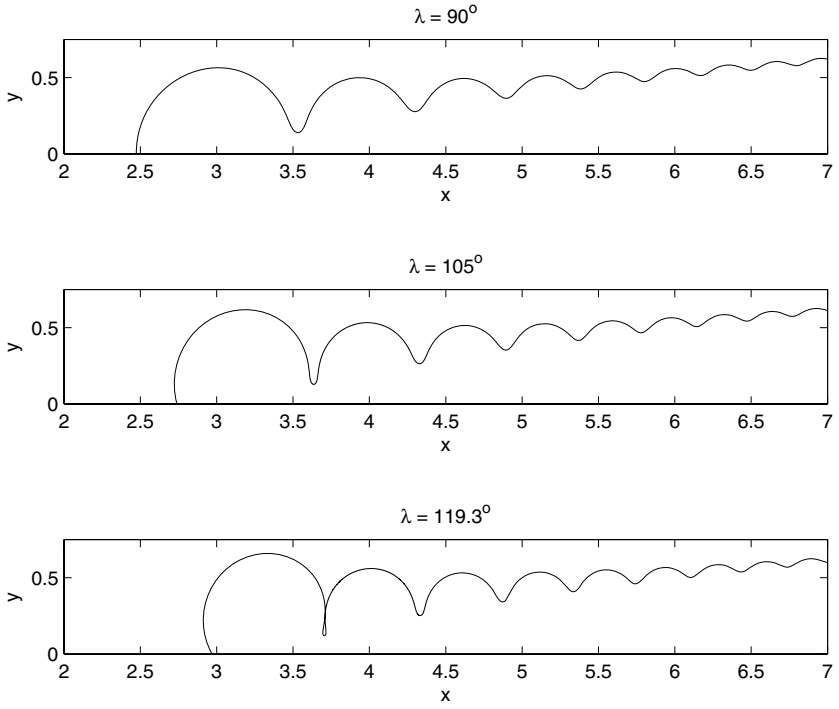


FIG. 5.4. *The position of the free surface calculated numerically for various* $\bar{\lambda}$ *and* $\epsilon = 5°$.

"beads" approach each other and meet when $\bar{\lambda} = \bar{\lambda}_c \approx 119.3°$, so that the free surface becomes self-intersecting. This raises the interesting question of what happens in the initial value problem when $\bar{\lambda} > \bar{\lambda}_c$, since no self-similar solution is available. The likely answer is that the problem can be regularized by the inclusion of the effect of viscosity at small times, and that a sequence of pinch-off events will occur as the flow develops. A similar problem is investigated in Billingham and King (2005), where the arguments that lead to this conclusion are presented.

6. Conclusions. In this paper, we have shown how Kuzmak's method can be used to analyze the response of a slender wedge of inviscid fluid to an abrupt change in contact angle. When the contact angle is of $O(1)$, although the underlying nonlinear oscillator is a nonlinear free boundary problem, we found that we could still make analytical progress. There are two related problems that may be amenable to this type of analysis.

First, the axisymmetric recoil of a slender cone of inviscid fluid bears many resemblances to the equivalent two-dimensional problem. However, there are two important differences. The first is that in the two-dimensional problem, the contact angle is a natural continuation parameter to move from a simple known solution to the required solution. No such parameter exists for the axisymmetric problem, since there is no equivalent of the moving contact line problem. The second difference is that, as far as we know, there is no axisymmetric equivalent of Kinnersley's analytical solution for the underlying nonlinear oscillator. The solution of the underlying axisymmetric nonlinear problem would therefore have to be obtained numerically, along with all of

the derivatives required for the secularity conditions. Although these solutions can be determined using the boundary integral method, this significantly increases the computational complexity of the problem. Some progress on the axisymmetric problem has been made using a one-dimensional approximation (Decent and King (2001)).

Second, it would be of interest to study the two-fluid version of the two-dimensional problem. In common with the axisymmetric problem, there is the difficulty that, again as far as we know, there exists no two-fluid version of Kinnersley's analytical solution. Moreover, it is not obvious how to modify the asymptotic scalings used here to accomodate the presence of an outer fluid. The two-fluid problem can be thought of as a combination of a recoiling slender void, described by Billingham and King (2005), and the problem we have studied here. Both the axisymmetric and two-fluid problems represent significant challenges in asymptotic analysis.

REFERENCES

J. Billingham (1999), *Surface-tension-driven flow in fat fluid wedges and cones*, J. Fluid Mech., 397, pp. 45–71.

J. Billingham and A. C. King (1995), *The interaction of a moving fluid/fluid interface with a flat plate*, J. Fluid Mech., 296, pp. 325–351.

J. Billingham and A. C. King (2005), *Surface-tension-driven flow outside a slender wedge with an application to the inviscid coalescence of drops*, J. Fluid Mech., 533, pp. 193–221.

F. J. Bourland and R. Haberman (1988), *The modulated phase shift for strongly nonlinear, slowly varying, and weakly damped oscillators*, SIAM J. Appl. Math., 48, pp. 737–748.

P. F. Byrd and M. D. Friedman (1954), *Handbook of Elliptic Integrals for Engineers and Physicists*, Springer-Verlag, Berlin.

R. G. Cox (1986), *The dynamics of the spreading of liquids on a solid surface.* I. *Viscous flow*, J. Fluid Mech., 168, pp. 169–194.

G. D. Crapper (1957), *An exact solution for progressive capillary waves of arbitrary amplitude*, J. Fluid Mech., 2, pp. 532–540.

D. Crowdy (1999), *Exact solutions for steady capillary waves on a fluid annulus*, J. Nonlinear Sci., 9, pp. 615–640.

S. P. Decent and A. C. King (2001), *The recoil of a broken liquid bridge*, in Proceedings of the IUTAM Symposium on Free Surface Flows, A. C. King and Y. D. Shikhmurzaev, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 81–88.

E. B. Dussan V and S. H. Davis (1974), *On the motion of a fluid-fluid interface along a solid surface*, J. Fluid Mech., 65, pp. 71–95.

J. B. Keller and M. J. Miksis (1983), *Surface tension driven flows*, SIAM J. Appl. Math., 43, pp. 268–277.

J. B. Keller, P. A. Milewski, and J. Vanden-Broeck (2000), *Wetting and merging driven by surface tension*, Eur. J. Mech. B Fluids, 19, pp. 491–502.

J. B. Keller, P. A. Milewski, and J.-M. Vanden-Broeck (2002), *Breaking and merging of liquid sheets and filaments*, J. Engrg. Math., 42, pp. 283–290.

A. C. King (1991), *Moving contact lines in slender fluid wedges*, Quart. J. Mech. Appl. Math., 44, pp. 173–192.

A. C. King, J. Billingham, and S. R. Otto (2003), *Differential Equations. Linear, Nonlinear, Ordinary, Partial*, Cambridge University Press, Cambridge, UK.

A. C. King, J. Billingham, and D. F. Popple (1999), *The moving contact line between two wedges of fluid on a flat plate*, Quart. J. Mech. Appl. Math., 53, pp. 453–468.

W. Kinnersley (1976), *Exact large amplitude capillary waves on sheets of fluid*, J. Fluid Mech., 77, pp. 229–241.

J. B. Lawrie (1990), *Surface-tension-driven flow in a wedge*, Quart. J. Mech. Appl. Math., 43, pp. 251–273.

J. B. Lawrie and A. C. King (1994), *Exact solutions to a class of functional difference equations with application to a moving contact line flow*, European J. Appl. Math., 5, pp. 141–157.

Y. D. Shikhmurzaev (1997), *Moving contact lines in liquid/liquid/solid systems*, J. Fluid Mech., 334, pp. 211–249.

A. Sierou and J. R. Lister (2004), *Self-similar recoil of inviscid drops*, Phys. Fluids, 16, pp. 1379–1394.

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 1978–1997

CONVECTION EFFECTS IN THIN REACTION ZONES: APPLICATIONS TO BIACORE*

DAVID A. EDWARDS†

Abstract. Surface-volume reactions occur in many physical systems such as biological and industrial processes. Though traditionally modeled as a surface, the reaction zone is usually a thin layer (often a gel) abutting a flowing fluid or gas. Therefore, one would expect a more realistic model for the reacting zone to include the effects of transport in the gel. In this paper we examine the BIAcore, a device for measuring rate constants which has this geometry. To explain anomalous measurements from the device, it has been proposed that some flow penetrates into the dextran (gel) layer, thus enhancing transport. To analyze the reversible kinetics, asymptotic and singular perturbation techniques are used, yielding linear and nonlinear integrodifferential equations. Explicit and asymptotic solutions are constructed for cases motivated by experimental design. The results indicate that such flow penetration effects are bound to be negligible in surface-volume reactions, regardless of the flow model used.

Key words. asymptotic expansions, BIAcore, biochemical reactions, integrodifferential equations, reaction zone, singular perturbations

AMS subject classifications. 35B25, 35C20, 35K57, 45J05, 80A30, 92C45

DOI. 10.1137/040621831

1. Introduction.

In many physical systems, so-called "surface-volume" reactions occur. In the simplest model, one reactant (herein called the *receptor*) is confined to a two-dimensional surface, while the other (the *ligand*) floats free in (a possibly stirred) solution, and the reaction occurs only when the ligand interacts with the surface. However, since the receptors are three-dimensional molecules, they either form or are embedded in a thin *reaction zone* (such as a gel) near the surface. Given that many of these systems occur in the presence of an active flow, it is natural to inquire into the effects of the flow on the reaction zones.

Models of this type are applicable to various industries. The creation of alginate gel in the food industry is enhanced by the addition of a convective flow of reactant [27]. In bubble reactors, gas reacts with the liquid which impinges on the bubble surfaces [19]. Corrosive processes occur in such geometries [13]. Inorganic material synthesis can be enhanced if the templates are immersed in flow, rather than fixed-batch, reactors [20]. In high-pressure continuous-flow fixed-bed reactors, gels are introduced at the reaction surface to minimize hydrodynamics effects [14]. Harmful blood clots form when platelets adhere to foreign objects in the presence of blood flow [12]. Various biological processes ensue when ligands floating in the bloodstream bind to cell receptors which occupy a thin reaction zone about the cell membrane [11].

1.1. The BIAcore.

For the purposes of this paper, we focus on the BIAcore, which is a surface plasmon resonance (SPR) device for measuring rate constants. The configuration of the BIAcore is described in great detail elsewhere [16], [17], [18], [26]. For the purposes of this manuscript, we may consider the BIAcore to consist

*Received by the editors December 31, 2004; accepted for publication (in revised form) June 6, 2006; published electronically September 26, 2006. This work was supported by National Institute for General Medical Sciences grant 1R01GM067244-01.
http://www.siam.org/journals/siap/66-6/62183.html
†Department of Mathematical Sciences, University of Delaware, Newark, DE 19716-2553 (edwards@math.udel.edu).

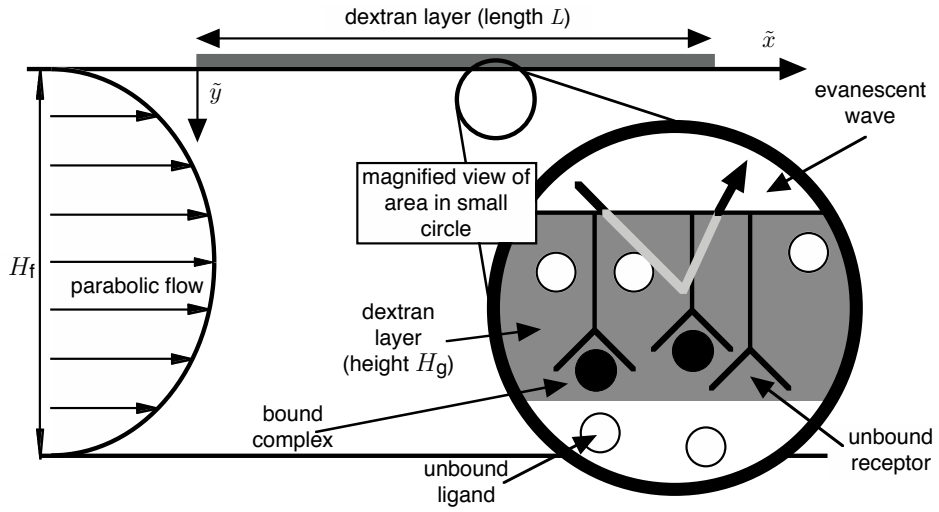


FIG. 1.1. *Schematic of BIAcore device. The coordinate system has its origin at the intersection of the $\tilde{y}$-axis and the dextran-flow interface.*

of a rectangular channel through which the ligand is convected in standard two-dimensional Poiseuille flow from $\tilde{x} = 0$, the inlet position (see Figure 1.1). Receptors are embedded in a thin dextran gel attached to the ceiling of the channel. Hence this device can serve as a representative of many physical systems of the type described above.

As the ligand diffuses to receptor sites, the binding process is measured by an evanescent wave that tracks mass changes in the dextran gel, as described more fully in section 6. This *sensogram* data is then transferred to a regression program which estimates the rate constants. During recent years, mathematical models of the BIAcore have become increasingly more sophisticated, treating many facets of its transport processes, including depletion of the free-flowing ligand along the channel [2], [3], [8], [21], [22]; diffusion in the gel [5], [25], [30], [31]; and signal decay associated with the measuring wave [6], [18], [25]. However, discrepancies still occur between measurements and simulations using the most sophisticated models [16], [24], [29].

To explain some anomalous observations, Witz [29] proposed that some of the buffer flow in the channel penetrates into the dextran gel, thus enhancing transport. In [7], Edwards formulated a mathematical model for this flow and analyzed it in the case where the reacting zone is treated as a surface. Now we shall treat the reaction zone as a layer.

As a first approximation, we model the dextran gel as a viscous fluid; others have treated it as a polymer brush [29]. We show that the physical parameter measuring penetration is $H_r = H_g/H_f$, the ratio of the heights of the gel and bulk flow regions. To leading order the flow adds a local depletion term to the mass action law for the bound state. When the Damköhler number Da is small, we obtain detailed expressions for the effect of penetration on the measurements. When $\mathrm{Da} = O(1)$ a nonlinear integral equation results, but the rate constants can easily be estimated using short-time asymptotics. We consider not only association, but also dissociation experiments. We also include the effect of evanescent wave decay in the measurement device.

All of our results indicate that flow penetration effects are very small. Such a conclusion arises from the geometry of the device, rather than from the model chosen for the dextran. Since the gel layer is so narrow compared to the rest of the device, velocities there will be small, no matter the actual transport model used. Clearly this result can be extended to the other physical systems described above.

2. Preliminaries. We consider the BIAcore to be divided into two regions, as shown in Figure 1.1: the open channel (the region $0 \le \tilde{y} \le H_{\rm f}$, where the subscript "f" stands for "flow") and the dextran gel layer (the region $-H_{\rm g} \le \tilde{y} \le 0$, where the subscript "g" stands for "gel"). We are interested in only that portion of the dextran layer which has length L, and so we take both regions to have $0 \le \tilde{x} \le L$.

2.1. Velocity profiles. The Reynolds number is small [7], so the flow in the channel is described by a simple one-dimensional laminar model. On the other hand, the dextran is a gel, and any true description of the flow therein would be quite complicated. For instance, Witz [29] considers the gel to be a polymer brush. For the purposes of this paper, we treat it merely as a very viscous fluid. This will necessarily misstate some quantitative features of the flow, but we shall show that such errors are negligible when analyzing sensogram data.

For simplicity, we introduce the following scalings for $\tilde{y}$ and the velocity field $\tilde{v}$:

$$y_{\rm f} = \frac{\tilde{y}}{H_{\rm f}}, \quad v_{\rm f}(y_{\rm f}) = \frac{\tilde{v}_{\rm f}(\tilde{y})}{V_{\rm f}}, \quad V_{\rm f} = \frac{\Delta p H_{\rm f}^2}{2\mu_{\rm f} L}, \quad y_{\rm g} = \frac{\tilde{y}}{H_{\rm g}}, \quad v_{\rm g}(y_{\rm g}) = \frac{\tilde{v}_{\rm g}(\tilde{y})}{V_{\rm g}}, \tag{2.1a}$$

$$V_{\rm g} = \frac{H_{\rm r}}{\mu_{\rm r}} V_{\rm f}, \quad H_{\rm r} = \frac{H_{\rm g}}{H_{\rm f}}, \quad \mu_{\rm r} = \frac{\mu_{\rm g}}{\mu_{\rm f}}, \tag{2.1b}$$

where μ is the bulk viscosity, V is the characteristic velocity in each region, and Δp is the (constant) pressure differential, which can be related to the known flow rate. Here (and throughout), if the same symbol appears both with and without tildes, the symbol with a tilde has dimensions, while the symbol without a tilde is dimensionless.

In (2.1b) the subscript "r" refers to "ratio," and we will use it in the same way (gel to flow) throughout. Using these scalings, it can be shown [7] that with suitable boundary and interface conditions, the velocity profiles are given by

$$v_{\rm f}(y_{\rm f}) = 1 - y_{\rm f}^2 + \frac{(y_{\rm f} - 1)(\mu_{\rm r} - H_{\rm r}^2)}{H_{\rm r} + \mu_{\rm r}}, \tag{2.2a}$$

$$v_{\rm g}(y_{\rm g}) = H_{\rm r}(1 - y_{\rm g}^2) + \frac{(y_{\rm g} + 1)(\mu_{\rm r} - H_{\rm r}^2)}{H_{\rm r} + \mu_{\rm r}}. \tag{2.2b}$$

Since solid dextran corresponds to $\mu_{\rm r} = \infty$, we might consider $\mu_{\rm r}$ as a large parameter to use in a perturbation approach. However, we can solve our problem for *any* $\mu_{\rm r}$ if we choose $H_{\rm r} \ll 1$, as motivated by its value in Table 4.1 below. Since $H_{\rm r}$ is simply a geometric parameter, such a choice will extend our results to other physical systems with thin reaction zones.

In the limit of small $H_{\rm r}$, (2.2b) becomes a nearly linear profile, corresponding to flow driven largely by shear from the bulk interface. Though using a more complicated polymer brush model for the gel leads to exponential and Bessel-function velocity profiles, these also reduce to linear profiles for small $H_{\rm r}$ [29]. Thus the two approaches are equivalent with a proper choice of $\mu_{\rm r}$.

2.2. Transport in the flow. In the flow, the ligand (concentration $\tilde{C}_{\rm f}$) travels both by convection and diffusion. However, the Peclét number in the flow, defined as

$$\text{Pe}_{\rm f} = \frac{H_{\rm f}^2/\tilde{D}_{\rm f}}{L/V_{\rm f}} = \frac{\text{characteristic diffusion time in flow}}{\text{characteristic convection time in flow}}, \tag{2.3}$$

is large (for a typical value, see Table 4.1 below). Here $\tilde{D}_{\rm f}$ is the molecular diffusion coefficient of the ligand in the flow. Hence one needs consider only the thin Lévêque boundary layer near $\tilde{y} = 0$ [3], which motivates the following scalings:

$$x = \frac{\tilde{x}}{L}, \quad y = \text{Pe}_{\rm f}^{1/3} y_{\rm f}, \quad t = \tilde{k}_{\rm on} C_{\rm u} \tilde{t}, \quad \tilde{C}_{\rm f}(\tilde{x}, \tilde{y}, \tilde{t}) = C_{\rm u}[1 - \text{Da} C_{\rm f}(x, y, t)], \tag{2.4}$$

where $\tilde{k}_{\rm on}$ is the association rate constant and $C_{\rm u}$ is the ligand concentration entering the device, which is used to create the dimensionless ligand concentration $C_{\rm f}$. Note that with our choice of scalings, the reaction time scale is the one of interest.

Here Da is the *Damköhler number*

$$\text{Da} = \frac{\tilde{k}_{\rm on}\tilde{R}_{\rm T}}{\tilde{D}_{\rm f}/(H_{\rm f}\text{Pe}_{\rm f}^{-1/3})} = \frac{\text{reaction ``velocity''}}{\text{diffusion ``velocity'' in boundary layer}}, \tag{2.5}$$

where $\tilde{R}_{\rm T}$ is the area density of receptor sites in the device. Da, which measures the effect of transport on the chemical reaction, characterizes the size of ligand depletion induced by the reaction, as shown in (2.4). Since $\text{Pe}_{\rm f} \propto V_{\rm f}$, $\text{Da} = 0$ corresponds to the case of infinitely fast flow where no depletion occurs. Most experiments are designed so that Da is small; hence the choice of scaling in (2.4) makes $C_{\rm f}$ a perturbation.

With these scalings, it can be shown [7] that the governing equations for $C_{\rm f}$ are

$$\frac{\partial^2 C_{\rm f}}{\partial y^2} = (v_0 + v_1 y)\frac{\partial C_{\rm f}}{\partial x}, \qquad C_{\rm f}(0, y, t) = 0, \quad C_{\rm f}(x, \infty, t) = 0, \tag{2.6a}$$

$$v_0 \equiv v_{\rm f}(0)\text{Pe}_{\rm f}^{1/3} = \frac{H_{\rm r}\text{Pe}_{\rm f}^{1/3}(H_{\rm r}+1)}{H_{\rm r}+\mu_{\rm r}}, \qquad v_1 \equiv v_{\rm f}'(0) = \frac{\mu_{\rm r} - H_{\rm r}^2}{H_{\rm r}+\mu_{\rm r}}. \tag{2.6b}$$

The scaling of v_0 in (2.6b) is chosen so that our results transition smoothly to the solid dextran case in the limit of large $\mu_{\rm r}$.

For the size of the transport processes to be comparable, the length scale in the Lévêque boundary layer in the fluid (where convection and diffusion balance) should be on the order of that in the gel. This implies that

$$H_{\rm g} = O(H_{\rm f}\text{Pe}_{\rm f}^{-1/3}) \Longrightarrow H_{\rm r} = O(\text{Pe}_{\rm f}^{-1/3}). \tag{2.7}$$

Such a scaling makes v_0 into an $O(1)$ quantity as long as we treat $\mu_{\rm r}$ as $O(1)$. Equation (2.7) also motivates the choice of $H_{\rm r}$ as a small parameter, since $\text{Pe}_{\rm f} \gg 1$. Unfortunately, (2.7) is rather a weak bound. From Table 4.1 below we have that $H_{\rm r} \ll \text{Pe}_{\rm f}^{-1/3}$, and so velocities in the gel will be comparatively small. Again, this result will hold for other systems with similar geometries.

To solve for $C_{\rm f}$, we use Laplace transforms (denoted with a hat) in the x-direction. To understand the gel dynamics, we need the value of $\hat{C}_{\rm f}$ only at the flow-gel interface $y = 0$. In particular, it can be shown [7] that $\hat{C}_{\rm f}$ satisfies

$$\hat{C}_{\rm f}(0, t) = \frac{\text{Ai}(s^{1/3}v_0/v_1^{2/3})}{(sv_1)^{1/3}\,\text{Ai}'(s^{1/3}v_0/v_1^{2/3})}\frac{\partial \hat{C}_{\rm f}}{\partial y}(0, t). \tag{2.8}$$

3. Dynamics in the dextran layer.

3.1. Transport. In the dextran gel, the reaction occurs only inside the pores, so it is the concentration of ligand *per fluid volume* that is important. To convert to this quantity, we simply divide $\tilde{C}_{\mathrm{g}}$, the concentration of ligand in the gel matrix, by the volume fraction ϕ of pores in the gel. (ϕ is also called the *partition coefficient*.) Motivated by this reasoning and (2.4), we choose the following scaling for $\tilde{C}_{\mathrm{g}}$:

$$\tilde{C}_{\mathrm{g}}(\tilde{x},\tilde{y},\tilde{t}) = \phi C_{\mathrm{u}}[1 - \mathrm{Da} C_{\mathrm{g}}(x, y_{\mathrm{g}}, t)]. \tag{3.1a}$$

Because the dextran layer is often treated as a surface, $\tilde{R}_{\mathrm{T}}$ is usually quoted as an *area* concentration. To convert this to a volume concentration, we simply divide by the width of the layer H_{g}, and hence we have an appropriate scaling for $\tilde{B}$:

$$B_{\mathrm{g}}(x, y_{\mathrm{g}}, t) = \frac{H_{\mathrm{g}}}{\tilde{R}_{\mathrm{T}}} \tilde{B}_{\mathrm{g}}(\tilde{x},\tilde{y},\tilde{t}). \tag{3.1b}$$

Though the receptor density may initially be nonuniform [15], [23], for now we take it to be uniform, since the error introduced from such an assumption is small [9].

In the dextran gel, the ligand travels both by convection and diffusion. Its evolution is also affected by binding. Since the concentration of available receptors is so much greater than the ligand concentration [5], the $\partial C_{\mathrm{g}}/\partial t$ term in the transport equation may be neglected. Moreover, $H_{\mathrm{g}} \ll L$, and so diffusion in the x-direction can be neglected.

Lastly, the Peclet number in the gel is given by

$$\mathrm{Pe}_{\mathrm{g}} = \frac{H_{\mathrm{g}}^2/\tilde{D}_{\mathrm{g}}}{L/V_{\mathrm{g}}} = \frac{H_{\mathrm{r}}^3}{\mu_{\mathrm{r}}} \frac{\tilde{D}_{\mathrm{f}}}{\tilde{D}_{\mathrm{g}}} \mathrm{Pe}_{\mathrm{f}} = \frac{1}{\mu_{\mathrm{r}}} \frac{\tilde{D}_{\mathrm{f}}}{\tilde{D}_{\mathrm{g}}} O(v_0^3), \tag{3.2}$$

where we have used (2.1) and (2.7). Here $\tilde{D}_{\mathrm{g}}$ is the diffusion constant in the dextran gel. The analysis in [7], which considers the case of a surface reaction, contains terms only up to $O(v_0)$. The same sort of analysis will hold here (as shown below), and thus we may ignore convection in the ligand transport equation. Physically, the small size of Pe_{g} in (3.2) shows that diffusion is the dominant transport process in the layer, not convection. Thus the dominant effect of flow penetration is a slip condition on the bulk flow.

Hence the leading-order dimensionless ligand transport equation is given by

$$\frac{\partial^2 C_{\mathrm{g}}}{\partial y_{\mathrm{g}}^2} = -D \frac{\partial B}{\partial t}, \tag{3.3a}$$

$$D = \frac{\tilde{D}_{\mathrm{f}}/(H_{\mathrm{f}} \mathrm{Pe}_{\mathrm{f}}^{-1/3})}{\phi \tilde{D}_{\mathrm{g}}/H_{\mathrm{g}}} = \frac{\text{diffusion velocity in diffusive boundary layer}}{\text{diffusion velocity in dextran}}. \tag{3.3b}$$

We solve (3.3a) by writing our solution as the sum of a particular solution A_{p} and a homogeneous solution A_{h}, as follows:

$$C_{\mathrm{g}}(x, y_{\mathrm{g}}, t) = -D A_{\mathrm{p}}(x, y_{\mathrm{g}}, t) + A_{\mathrm{h}}(x, y_{\mathrm{g}}, t), \tag{3.4}$$

where A_{p} satisfies

$$\frac{\partial^2 A_{\mathrm{p}}}{\partial y_{\mathrm{g}}^2} = \frac{\partial B}{\partial t}, \qquad \frac{\partial A_{\mathrm{p}}}{\partial y_{\mathrm{g}}}(x, -1, t) = 0, \qquad A_{\mathrm{p}}(x, 0, t) = 0. \tag{3.5}$$

The homogeneous problem is most easily solved in Laplace transform space. Solving the homogeneous form of the operator in (3.3a) subject to a no-flux condition at the channel wall $y = -1$, we determine that $\hat{A}_\mathrm{h}$ is a function of t only. At the flow-gel interface, the flux and ligand concentration per fluid volume must be continuous. Combining these conditions with (2.8) and using the transform of (3.4), we obtain

$$\hat{A}_\mathrm{h}(t) = -\frac{\mathrm{Ai}(s^{1/3}v_0/v_1^{2/3})}{(sv_1)^{1/3}\,\mathrm{Ai}'(s^{1/3}v_0/v_1^{2/3})}\frac{\partial \hat{A}_\mathrm{p}}{\partial y_\mathrm{g}}(0,t). \tag{3.6}$$

Since we cannot invert (3.6) in closed form, we use the fact that we consider the dextran to be a very viscous fluid, so $v_0 \to 0$. Formally, there are two ways to justify this from (2.6b). The first, physically intuitive, reasoning is to say that $\mu_\mathrm{r} \to \infty$. The second, more consistent from a mathematical point of view, is to take $H_\mathrm{r} \to 0$. Then expanding (3.6) to leading two orders in v_0 and inverting, we obtain the following:

$$A_\mathrm{h}(x,t) = \frac{1}{(3v_1)^{1/3}\Gamma(2/3)}\int_0^x \frac{\partial A_\mathrm{p}}{\partial y_\mathrm{g}}(x-\xi,0,t)\frac{d\xi}{\xi^{2/3}} - \frac{v_0}{v_1}\frac{\partial A_\mathrm{p}}{\partial y_\mathrm{g}}(x,0,t) + O(v_0^2). \tag{3.7}$$

Note from (3.5) that

$$\frac{\partial A_\mathrm{p}}{\partial y_\mathrm{g}}(x,0,t) = \int_{-1}^0 \frac{\partial B}{\partial t}\,dy_\mathrm{g};$$

in other words, the derivative is simply the average rate of binding in the layer at fixed x. Thus the integral term in (3.7) has an elegant physical interpretation, namely that the deficit in the ligand concentration at position x is the accumulation of the reaction that has occurred upstream. The effect of the slip velocity is to introduce the *local* reaction into the computation of the ligand deficit through the second term in (3.7).

When expanding (3.6) to obtain the expansion in (3.7), we tacitly assumed that $s^{1/3}v_0 \ll 1$. However, Laplace transform theory states that small x corresponds to large s, so this assumption does not hold in the limit of small x. Fortunately, the BIAcore returns measurements not of B, but of its average over the entire layer and some scanning range $x_\mathrm{min} \le x \le x_\mathrm{max}$:

$$\bar{B}(t) = \frac{1}{x_\mathrm{max} - x_\mathrm{min}}\int_{x_\mathrm{min}}^{x_\mathrm{max}}\int_{-1}^0 B(x,y_\mathrm{g},t)\,dy_\mathrm{g}\,dx, \tag{3.8}$$

where x_min is bounded away from zero. Since $x = 0$ is out of the scanning range, we may confidently use our results to analyze sensogram data.

3.2. Reaction. The bound state evolves according to a standard bimolecular mass action law. Using the scalings in (2.4) and (3.1) leads to the dimensionless form

$$\frac{\partial B}{\partial t} = (1-B)(1-\mathrm{Da}C_\mathrm{g}) - KB, \quad K = \frac{\tilde{k}_\mathrm{off}}{\tilde{k}_\mathrm{on}C_\mathrm{u}}, \tag{3.9a}$$

$$B(x,y_\mathrm{g},0) = B_\mathrm{i}, \tag{3.9b}$$

where $\tilde{k}_\mathrm{off}$ is the dissociation rate constant and K is the dimensionless affinity constant. Though the theory can handle general initial conditions for B, in practice the initial condition is always spatially uniform. For an association experiment, initially there is no bound state. For a dissociation experiment, we start with the steady state of (3.9a), which will be shown to be a constant.

Since A_{p} depends on $\partial B/\partial t$, substitution of (3.4) and (3.7) into (3.9a) would yield a nonlinear integrodifferential equation, and an exact solution would have to be obtained numerically. However, asymptotic results can be derived in physically relevant regimes.

4. Small Da results. Equation (3.9a) shows that $\mathrm{Da} = 0$ corresponds to the well-mixed case where there is no depletion. When designing experiments, scientists strive to keep Da as small as possible to minimize transport effects [28]. Therefore, we now specialize to the case of small Da by introducing the following expansion:

$$B(x, y_{\mathrm{g}}, t) = B_0(x, y_{\mathrm{g}}, t) + \mathrm{Da} B_1(x, y_{\mathrm{g}}, t) + o(\mathrm{Da}). \tag{4.1}$$

4.1. Association experiment. We begin by considering an association experiment as described in section 3. If we substitute (4.1) into (3.9), we find to leading order that the ligand concentration does not contribute. Hence we are in the well-mixed case, the solution of which is given by

$$B_0(x, y_{\mathrm{g}}, t) = \frac{1 - e^{-\alpha t}}{\alpha} + B_{\mathrm{i}} e^{-\alpha t} = \bar{B}_0(t), \tag{4.2}$$

which leads to the following expression for A_{p}:

$$A_{\mathrm{p}} = \frac{dB_0}{dt} \frac{y_{\mathrm{g}}(y_{\mathrm{g}} + 2)}{2}. \tag{4.3}$$

Since A_{p} is independent of x, it is simple to use (4.3) in (3.7) and compute that

$$A_{\mathrm{h}}(x, t) = \frac{dB_0}{dt} h(x), \qquad h(x) = \frac{3^{2/3} x^{1/3}}{v_1^{1/3} \Gamma(2/3)} - \frac{v_0}{v_1}. \tag{4.4}$$

The value of $A_{\mathrm{h}}(x, t)$ in (4.4) is exactly the value of $C_{\mathrm{f}}(x, 0, t)$ obtained if the reacting zone is treated as (instead of a layer) a two-dimensional surface at $x = 0$. In that case, $D = 0$, and hence there is no contribution from A_{p} in (3.4).

Substituting (4.3) and (4.4) into (3.4), we obtain

$$C_{\mathrm{g}}(x, y_{\mathrm{g}}, t) = \frac{dB_0}{dt} h_{\mathrm{g}}(x, y_{\mathrm{g}}), \qquad h_{\mathrm{g}}(x, y_{\mathrm{g}}) = \left[-D \frac{y_{\mathrm{g}}(y_{\mathrm{g}} + 2)}{2} + h(x) \right]. \tag{4.5}$$

Thus, as in [5], the effects of the variables x and y decouple. We also note that (4.3) is exactly the same as in [5], which considered the no-penetration case. Hence the effect of flow penetration appears only in the homogeneous part. Since the velocity in the layer is negligible at this order, the flow simply provides a slip condition for the bulk, which then couples to the receptor layer through the flow-gel interface conditions.

Substituting (4.2) and (4.5) into the next order of our expansion of (3.9) and solving, we have the following:

$$B_1 = \left[\frac{(e^{-\alpha t} - 1)\chi}{\alpha} - Kt \right] \frac{\chi e^{-\alpha t}}{\alpha} h_{\mathrm{g}}(x, y_{\mathrm{g}}). \tag{4.6}$$

Then averaging, we obtain

$$\bar{B}_1 = \frac{\chi e^{-\alpha t}}{\alpha} \left[\frac{(e^{-\alpha t} - 1)\chi}{\alpha} - Kt \right] \bar{h}_{\mathrm{g}}, \qquad \bar{h}_{\mathrm{g}} = \frac{D}{3} + \bar{h}. \tag{4.7}$$

TABLE 4.1
Parameter values for Figures 4.1 *and* 4.2.

Parameter	Value	Parameter	Value
B_{i}	0	Pe_{f}	3.72×10^{2}
C_{T} (mol/cm^3)	10^{-11}	t	$10^{-3}\,\tilde{t}$/s
D	1.20×10^{-1}	$x_{\max}$	7.92×10^{-1}
Da	10^{-1}	$x_{\min}$	2.08×10^{-1}
H_{r}	2×10^{-3}	α	2
K	1	χ	1
$\tilde{k}_{\mathrm{on}}$(cm^{3}mol^{-1}s^{-1})	10^{8}		

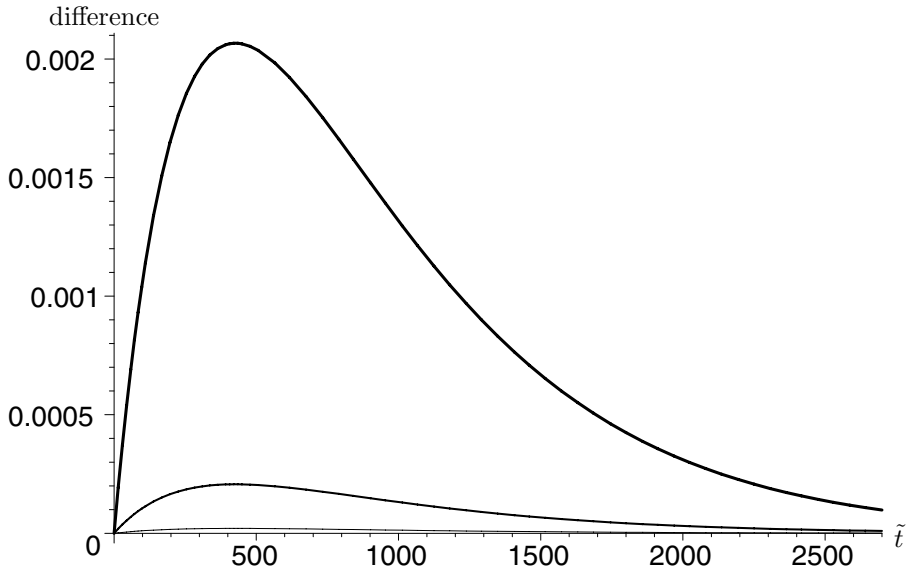


FIG. 4.1. *Absolute difference between* (4.7) *with* $\mu_{\mathrm{r}} = \infty$ (*solid dextran*) *and* μ_{r} *finite for* (*in decreasing order of thickness*) $\mu_{\mathrm{r}} = 1,\ 10,\ 100$. *Relative difference is about* 1%. *Association experiment.*

Aficionados of perturbation theory will note the term in (4.7) proportional to $te^{-\alpha t}$, similar to a secularity in a two-timing exercise. As $t \to \infty$, $B_0 = O(1)$ and $\mathrm{Da}B_1 \ll B_0$, so from an experimental standpoint, this is not a problem. However, it can be shown [3], [8] that a multiple-scale expansion is formally required. Though we could construct such an expansion for this case, it will not be illuminating.

We use the parameters listed in Table 4.1 to plot our solutions. The parameter values are from [7], with the exception of the value for D, which is from [6].

Figure 4.1 shows the effect of μ_{r} on $\bar{B}_1$ plotted against the *dimensional* time $\tilde{t}$ (in seconds) for various values of μ_{r}. We use the dimensional time in order to compare better with sensogram data. Note that in every case the difference is quite small due to the low value of H_{r}. In particular, even the error for $\mu_{\mathrm{r}} = 1$ (corresponding to the absence of a dextran layer) is only $O(H_{\mathrm{r}})$. In addition, the difference is positive; that is, allowing the flow to penetrate into the dextran layer enhances the association process. Since H_{r} is a geometrical parameter, this order estimate holds for other physical systems of this type.

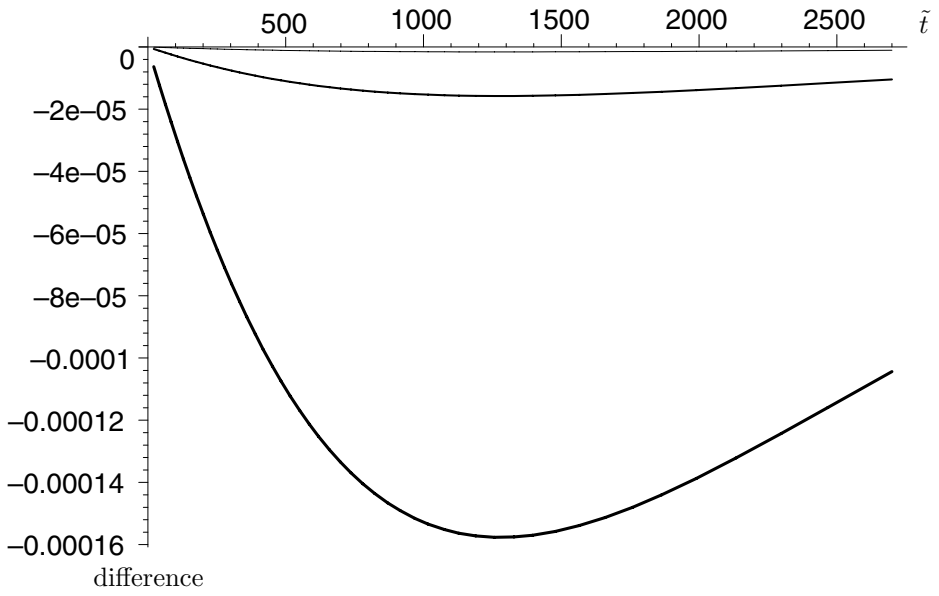


FIG. 4.2. *Absolute difference between effective rate constant solution* $\bar{B}$ *of* (4.8) *with* $\mu_r = \infty$ (*solid dextran*) *and* μ_r *finite for* (*in decreasing order of thickness*) $\mu_r = 1, 10, 100$. *Relative difference is about* 0.04%. *Association experiment.*

In Figure 4.1 (and throughout this manuscript) we graph the absolute difference because it is that difference which will be measured by the device and plotted on the sensogram. Obviously small differences will be masked by the underlying noise in any experiment. However, for comparison purposes and to extend our results to other physical systems, we also compute the relative difference. The case with $\mu_r = \infty$, graphed in Figure 1 of [5], shows that B_1 is on the order of 0.2, so the differences shown in Figure 4.1 are on the order of 1%. (Note this is the difference in $\bar{B}_1$, not in the whole solution, which would include the much greater contribution from $\bar{B}_0$.)

These results may be stated more simply in the context of an *effective rate constant* (ERC) equation, as outlined in [5], [8], and [21]. Substituting (4.5) into (3.9a), we have

$$\frac{\partial B}{\partial t} = (1 - B)\left[1 - \mathrm{Da}\frac{dB_0}{dt}h_g(x)\right] - KB + O(\mathrm{Da}^2),$$

which we may rearrange and average to obtain

$$\frac{d\bar{B}}{dt} = \frac{1 - \alpha\bar{B}}{1 + \mathrm{Da}(1 - \bar{B})\bar{h}_g} + O(\mathrm{Da}^2). \tag{4.8}$$

Equation (4.8) is an ODE for $\bar{B}$, the actual sensogram data produced by the BIAcore, and hence the solution requires no postprocessing averaging step. Equation (4.8) is in the form obtained previously [5], albeit with a different value of $\bar{h}$. This is consistent with [4], where it is shown that if B_0 is spatially uniform, the ERC approximation is robust to *any* geometry or flow.

Figure 4.2 shows the effect of μ_r on our ERC solution. The absolute error here is an order of magnitude smaller than that in Figure 4.1. This is because Figure 4.1

shows errors in $\bar{B}_1$, which have to be multiplied by Da (0.1 in our graphs) to obtain the error in the full solution as shown in Figure 4.2. The relative error here is around 0.1%, as can be seen by comparison with the full solution in Figure 1 of [8]. As before, flow penetration enhances the association process.

4.2. Dissociation experiment. A typical BIAcore experimental run begins with an association experiment run to completion. At this stage, pure buffer solution (inlet concentration zero) is injected into the device, initiating the dissociation process. This second experiment then provides additional data for rate constant estimation.

The initial condition for this phase of the experiment is the steady state of (3.9a). Since there is no flux of ligand through the exterior wall, one finds from (3.3a) that the steady state of C_{g} must be a constant. By using continuity arguments at the interface, one finds that this constant must be zero. Thus the steady state of (3.9a) is

$$B_{\mathrm{s}} = \alpha^{-1}, \tag{4.9}$$

where the subscript "s" refers to "steady state." Since (4.9) provides the initial condition for the dissociation problem, we are justified in always taking a constant initial condition for B.

With no ligand injected into the device, the equation analogous to (3.9a) becomes

$$\frac{\partial B}{\partial t} = (1-B)(-\mathrm{Da}C_{\mathrm{g}}) - KB. \tag{4.10}$$

Thus in this case $C_{\mathrm{g}} \leq 0$. The analysis proceeds in a manner analogous to the association experiment; the relevant results are given by

$$B_0(x,t) = \bar{B}_0(t) = \frac{e^{-Kt}}{\alpha}, \tag{4.11a}$$

$$\bar{B}_1 = \frac{K}{\alpha}\left(t + \frac{e^{-Kt}-1}{K\alpha}\right)\bar{h}_{\mathrm{g}}e^{-Kt}, \tag{4.11b}$$

where $\bar{h}_{\mathrm{g}}$ is given in (4.7). The same secularity problem appears with more obvious effects, since in the dissociation experiment the second term in the expansion can become larger than the first. Again, we restrict ourselves to the case where $\mathrm{Da}t = O(1)$, since constructing the multiple-scale expansion is not illuminating.

With the inlet value 1 absent from the concentration term in (4.10), the expression analogous to (4.8) is given by

$$\frac{d\bar{B}}{dt} = \frac{-K\bar{B}}{1+\mathrm{Da}(1-\bar{B})\bar{h}_{\mathrm{g}}} + O(\mathrm{Da}^2), \tag{4.12}$$

as in [5].

5. Moderate Da results. Since C_{g} depends on B, (3.9a) is nonlinear if $\mathrm{Da} = O(1)$. Thus to obtain analytic solutions we resort to short-time asymptotics by assuming a solution of the form

$$B(x,y_{\mathrm{g}},t) = B_{\mathrm{i}} + \beta(x,y_{\mathrm{g}})t + o(t), \qquad A_{\mathrm{p}}(x,y_{\mathrm{g}},t) = A_{\mathrm{p},1}(x,y_{\mathrm{g}}) + o(1), \tag{5.1a}$$

$$A_{\mathrm{h}}(x,t) = A_{\mathrm{h},1}(x) + o(1). \tag{5.1b}$$

5.1. Association experiment. Substituting (5.1) into (3.4), (3.9a), and (3.5), we have, to leading order in t,

$$C_{\mathrm{g}} = -DA_{\mathrm{p},1} + A_{\mathrm{h},1},$$

$$\beta = (1 - B_{\mathrm{i}})[1 - \mathrm{Da}(-DA_{\mathrm{p},1} + A_{\mathrm{h},1})] - KB_{\mathrm{i}}, \tag{5.2a}$$

$$\frac{\partial^2 A_{\mathrm{p},1}}{\partial y_{\mathrm{g}}^2} = \beta, \qquad \frac{\partial A_{\mathrm{p},1}}{\partial y_{\mathrm{g}}}(x,-1) = 0, \qquad A_{\mathrm{p},1}(x,0) = 0. \tag{5.2b}$$

Since (5.2a) is linear, it is most convenient to work in Laplace transform space. The relationship between $\hat{A}_{\mathrm{h},1}$ and $d\hat{A}_{\mathrm{p},1}/dy_{\mathrm{g}}(0)$ is exactly the same as in the transform of (3.7). Upon substituting that expression into the transform of (5.2) and solving, we obtain a form for $\hat{A}_{\mathrm{p},1}$ that depends explicitly on $d\hat{A}_{\mathrm{p},1}/dy_{\mathrm{g}}(0)$. Some algebraic manipulation eliminates the unknown from our solution, yielding

$$\hat{A}_{\mathrm{p},1} = -\frac{\chi}{\lambda_{\mathrm{a}}^2 r_{\mathrm{a}} s}\left[1 - \frac{\cosh\lambda_{\mathrm{a}}(y_{\mathrm{a}}+1)}{\cosh\lambda_{\mathrm{a}}}\right]\left(1 + \frac{\nu_{\mathrm{a}}^{1/3}}{s^{1/3}}\right)^{-1}, \tag{5.3}$$

$$r_{\mathrm{a}} = 1 - (1 - B_{\mathrm{i}})\mathrm{Da}\frac{v_0}{v_1}\frac{\tanh\lambda_{\mathrm{a}}}{\lambda_{\mathrm{a}}}, \qquad \lambda_{\mathrm{a}}^2 = D\mathrm{Da}(1 - B_{\mathrm{i}}), \tag{5.4a}$$

$$\nu_{\mathrm{a}} = \frac{1}{3v_1}\left\{\frac{\Gamma(2/3)}{\Gamma(1/3)}\left[\frac{1}{\mathrm{Da}(1-B_{\mathrm{i}})}\frac{\lambda_{\mathrm{a}}}{\tanh\lambda_{\mathrm{a}}} - \frac{v_0}{v_1}\right]\right\}^{-3}, \tag{5.4b}$$

where the subscript "a" denotes "association." We have written (5.4b) in a form where the correction due to v_0 can be easily seen. Recall that in deriving this form, we have already taken an asymptotic limit for small v_0. Thus, we should expect that (5.3) will hold only for those Da where the first bracketed term is much larger than the second.

To simplify the interpretation of the data, we write the average (3.8) in dimensional form with the aid of (2.4):

$$\bar{B}(\tilde{t}) \sim B_{\mathrm{i}} + S\tilde{t}, \quad \tilde{t} \to 0,$$

$$S = \frac{\tilde{k}_{\mathrm{on}} C_{\mathrm{u}}\{\mathcal{I}[\beta; x_{\max}] - \mathcal{I}[\beta; x_{\min}]\}}{x_{\max} - x_{\min}}, \tag{5.5a}$$

$$\mathcal{I}[\beta; x] \equiv \int_0^x \int_{-1}^0 \beta(\xi, y_{\mathrm{g}})\, dy_{\mathrm{g}}\, d\xi. \tag{5.5b}$$

Substituting (5.3) into the Laplace transform of (5.2b) and inverting, we have

$$\mathcal{I}[\beta; x] = \frac{\chi e^{-\nu_{\mathrm{a}} x}}{\nu_{\mathrm{a}} r_{\mathrm{a}}}\left[e^{\nu_{\mathrm{a}} x} - 1 - \left|P\left(\frac{4}{3}, -\nu_{\mathrm{a}} x\right)\right| + \left|P\left(\frac{5}{3}, -\nu_{\mathrm{a}} x\right)\right|\right]\frac{\tanh\lambda_{\mathrm{a}}}{\lambda_{\mathrm{a}}}, \tag{5.6a}$$

where P is the normalized *lower* incomplete gamma function whose definition is [1]

$$P\left(\frac{m}{3}, -\nu_{\mathrm{a}} x\right) = \frac{\gamma(m/3, -\nu_{\mathrm{a}} x)}{\Gamma(m/3)}. \tag{5.6b}$$

In the limit that $D \to 0$, $\lambda_{\mathrm{a}} \to 0$ and (5.6a) reduces to the result in the surface reaction case [7].

To estimate the rate constants from an experiment, we first run the association experiment to steady state. This will yield an estimate for α, and hence K, from (4.9). To calculate $\tilde{k}_{\mathrm{on}}$, we use the linear fit S from our short-time data in (5.5a) to

TABLE 5.1
Parameter values for Figures 5.1 *and* 5.2.

Parameter	Value	Parameter	Value
$\tilde{D}_{\mathrm{f}}$ (cm^2/s)	2.8×10^{-7}	L (cm)	2.4×10^{-1}
$\tilde{D}_{\mathrm{g}}$ (cm^2/s)	3.36×10^{-8}	R_{T} (mol/cm^2)	10^{-12}
$\tilde{k}_{\mathrm{off}}$ (s^{-1})	8.9×10^{-3}	ϕ	1

obtain $\tilde{k}_{\mathrm{on}}$. Since Da also depends on $\tilde{k}_{\mathrm{on}}$, the relationship between S and $\tilde{k}_{\mathrm{on}}$ is not a simple linear one. Using our estimates for K and $\tilde{k}_{\mathrm{on}}$ together, we may calculate $\tilde{k}_{\mathrm{off}}$.

We may asymptotically determine the behavior of S for small $\tilde{k}_{\mathrm{on}}$, which corresponds to small k. For small k, $r_{\mathrm{a}} \to 1$, $\lambda_{\mathrm{a}} \to 0$, and $\nu_{\mathrm{a}} \to 0$, so we have

$$S \sim \tilde{k}_{\mathrm{on}} C_{\mathrm{u}} \chi, \qquad \tilde{k}_{\mathrm{on}} \to 0. \tag{5.7}$$

Equation (5.7) merely shows that if there is no forward reaction ($\tilde{k}_{\mathrm{on}} = 0$), then there will be no change in the bound concentration from the initial state ($S = 0$).

At the other extreme, we cannot ascertain the behavior in the case that $k \to \infty$ due to the form of (5.4b). As k increases, so will Da, thus eventually causing the assumed ordering in (5.4b) to be violated. Essentially, because of the faster reaction, we cannot simply take the first-order convection correction; we must include additional terms in our analysis.

Since $\tanh \lambda_{\mathrm{a}} < 1$, we can still satisfy our ordering if we replace $\tanh \lambda_{\mathrm{a}}$ with 1 for simplicity. By doing so, we may construct a specific bound using the parameters in Table 5.1 (which come from [6] and [7]). The bound is calculated to be

$$k \ll 232 \mu_{\mathrm{r}}^2, \tag{5.8}$$

which is quadratic in μ_{r}—a much less restrictive bound than the linear bound in the surface reaction case [7].

5.2. Dissociation experiment. For the dissociation case, the initial condition is the steady state from the association problem, given in (4.9). In addition, the leading-order concentration is now 0, not 1. Essentially, only the parameters in the problem have changed, not the general structure. Therefore, the solution process follows as before, and our expression for $\mathcal{I}$ is

$$\mathcal{I}[\beta; x] = -\frac{K e^{-\nu_{\mathrm{d}} x}}{\alpha \nu_{\mathrm{d}} r_{\mathrm{d}}} \left[e^{\nu_{\mathrm{d}} x} - 1 - \left| P\left(\frac{4}{3}, -\nu_{\mathrm{d}} x\right) \right| + \left| P\left(\frac{5}{3}, -\nu_{\mathrm{d}} x\right) \right| \right] \frac{\tanh \lambda_{\mathrm{d}}}{\lambda_{\mathrm{d}}}, \tag{5.9}$$

$$\lambda_{\mathrm{d}}^2 = D \mathrm{Da} \left(\frac{K}{\alpha}\right), \qquad r_{\mathrm{d}} = 1 - \left(\frac{K}{\alpha}\right) \mathrm{Da} \frac{v_0}{v_1} \frac{\tanh \lambda_{\mathrm{d}}}{\lambda_{\mathrm{d}}}, \tag{5.10a}$$

$$\nu_{\mathrm{d}} = \frac{1}{3 v_1} \left\{ \frac{\Gamma(2/3)}{\Gamma(1/3)} \left[\frac{1}{\mathrm{Da}(K/\alpha)} \frac{\lambda_{\mathrm{d}}}{\tanh \lambda_{\mathrm{d}}} - \frac{v_0}{v_1} \right] \right\}^{-3}, \tag{5.10b}$$

where the subscript "d" refers to "dissociation." Similar to the previous subsection, in the limit that $D \to 0$, our results reduce to the surface reaction case in [7].

Using our new initial condition, we write our average as

$$\bar{B}(t) \sim \frac{1}{\alpha} + S\tilde{t}, \quad \tilde{t} \to 0,$$

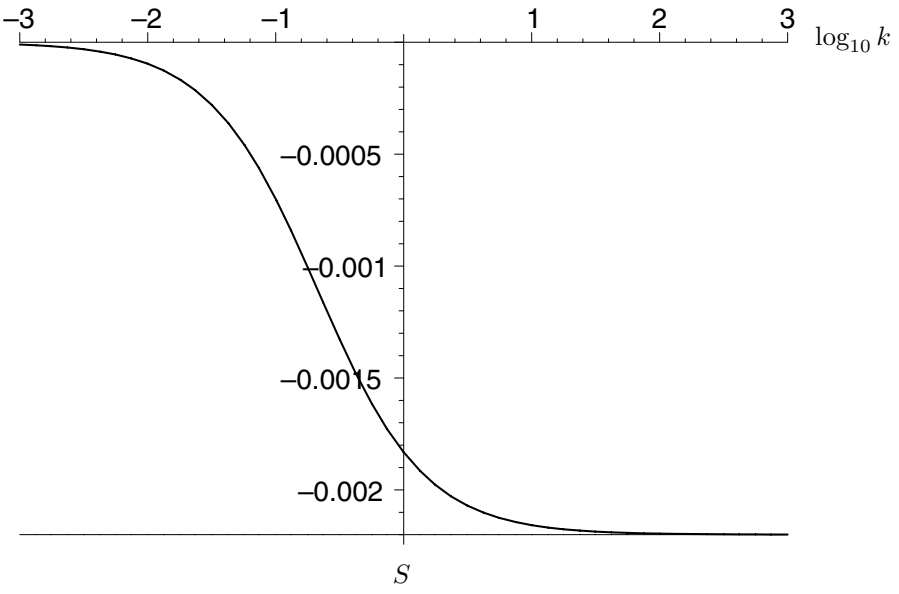


FIG. 5.1. *Thick line: S versus $\log_{10} k$, keeping $\tilde{k}_{\text{off}}$ fixed. Thin line: large-k asymptote. Dissociation experiment, $v_0 = 0$.*

where S is defined in (5.5a). Note from (5.9) that the slope is now negative, as expected for our dissociation problem.

We carefully analyze the behavior of S with respect to k, beginning with the case where $v_0 = 0$. In [5], the author kept K fixed and varied $\tilde{k}_{\text{on}}$, which necessitated (tacitly) varying $\tilde{k}_{\text{off}}$. In contrast, here we wish to keep $\tilde{k}_{\text{off}}$ fixed, which means that K will vary as $\tilde{k}_{\text{on}}$ does. This approach was taken in the study of flow penetration on the surface reaction case [7]. The value chosen for $\tilde{k}_{\text{off}}$ is listed in Table 5.1 and comes from [7].

In order to visualize the relationship between S and $\tilde{k}_{\text{on}}$, in Figure 5.1 we construct a curve using the parameters in Table 5.1 with $\mu_{\text{r}} = \infty$ (the solid dextran case). For convenience, we define the new variable

$$k = 10^{-9} \tilde{k}_{\text{on}} \frac{\text{mol} \cdot \text{s}}{\text{cm}^3}. \tag{5.11}$$

Since we take $B_{\text{i}} = 0$ for the association case, $\chi = 1$ and the solution is independent of K. It can be shown that the only qualitative difference between the graph here and with K fixed is in the asymptotes. (See [7] for a related discussion of the layer reaction case.)

For the small-$\tilde{k}_{\text{on}}$ asymptote, we first note from (5.10) that as $\tilde{k}_{\text{on}} \to 0$, $\lambda_{\text{d}} \to 0$, which causes the λ_{d} contribution to disappear, as in the previous subsection. In addition, $r_{\text{d}} \to 1$ and $\nu_{\text{d}} \to 0$. Thus we have that

$$S \sim -\tilde{k}_{\text{on}} C_{\text{u}}, \qquad \tilde{k}_{\text{on}} \to 0. \tag{5.12}$$

Note that (5.12) holds regardless of the value of v_0.

For the large-$\tilde{k}_{\text{on}}$ asymptote, we must restrict ourselves to the case with no flow:

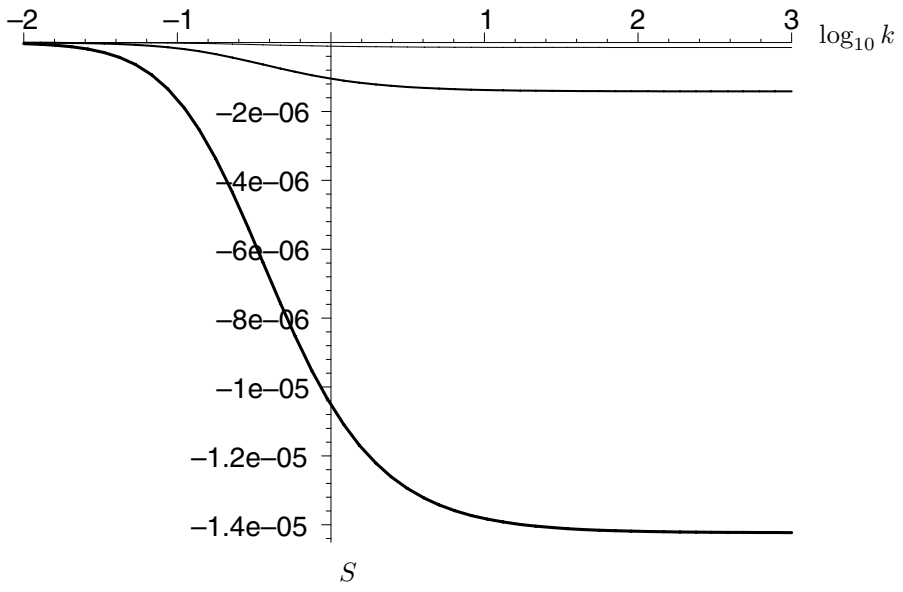


FIG. 5.2. *Absolute difference between S with $\mu_r = \infty$ (solid dextran) and μ_r finite versus $\log_{10} k$ for (in decreasing order of thickness) $\mu_r = 1$, 10, 100. Relative difference is about 1%. Dissociation experiment.*

$v_0 = 0$ and $v_1 = 1$, so $r_d = 1$. First, it is convenient to estimate λ_d in this limit:

$$\lim_{\tilde{k}_{on}\to\infty} \lambda_d \equiv \lambda_\infty = \left(\frac{\tilde{R}_T H_g}{\phi \tilde{D}_g} \frac{\tilde{k}_{off}}{C_u} \right)^{1/2}. \tag{5.13a}$$

In addition, we may use (5.10b) to calculate ν_∞ for this case:

$$\nu_\infty = \frac{1}{3\mathrm{Pe}_f} \left[\frac{\Gamma(1/3)}{\Gamma(2/3)} \frac{\tilde{R}_T H_f}{\tilde{D}_f} \frac{\tilde{k}_{off}}{C_u} \frac{\tanh \lambda_\infty}{\lambda_\infty} \right]^3, \quad v_0 = 0. \tag{5.13b}$$

Equations (5.13) illustrate a key difference between our analysis and that of [5]. In that work, K was kept fixed, so λ_d was unbounded as $\tilde{k}_{on} \to \infty$. This simplified $\mathcal{I}[\beta; x]$ greatly, leading to a relatively simple result. In our case, the asymptote has no convenient closed-form solution, but upon substituting (5.13) into (5.6a) and (5.5a), we obtain $S = -2.20 \times 10^{-3}$, which is exactly the asymptote in Figure 5.1.

Now that we have a baseline result for comparison, we next vary the viscosity ratio μ_r. Again we must preserve the assumed size ordering in (5.10b). Substituting our parameters into the above, we have

$$4.74 \times 10^{-1} \ll 16.2\mu_r, \tag{5.14}$$

which is simply a bound on μ_r that is always satisfied experimentally. Thus our expressions do not break down for large k as in the association case.

In Figure 5.2 we examine the effect of varying viscosity on the short-time asymptote S. As expected, the corrections are again small. The decrement to S increases

with decreasing μ_r, as lower μ_r means more convective transport, which enhances dissociation. By examining Figure 5.1, we see that the relative difference is about 1%.

6. Evanescent wave effects. The way in which an SPR device like the BIAcore measures binding is quite involved; here we present a brief summary. As binding occurs in the gel, the gel's index of refraction changes. A polarized light beam is aimed at the sensor surface at various angles. A sharp decrease in reflectivity is noted for a certain incidence angle, which can be related to the index of refraction and hence the binding. Since the strength of the evanescent wave (electric field) decays as it penetrates into the gel, the effect of binding on signal decays further from the surface $\tilde{y} = -H_g$ [10], [18].

By using a simple exponential decay model for the signal, we obtain the following result for the average $\bar{B}$, which replaces (3.8):

(6.1)
$$\bar{B}(t;\delta) = \frac{\delta}{(1-e^{-\delta})(x_{\max}-x_{\min})}\int_{x_{\min}}^{x_{\max}}\int_{-1}^{0} e^{-\delta(y_g+1)}B(x,y_g,t)\,dy_g\,dx, \qquad \delta = \frac{H_g}{H_w},$$

where H_w is the characteristic decay length of the wave. Here we include δ explicitly in the notation for $\bar{B}$ to indicate that we are including wave effects.

The only change to our work from previous sections is the calculation of averages. Since the leading-order solutions are independent of y_g, their averages do not change with the decaying signal strength.

6.1. Results for small Da. We begin by examining the small Da case. Using our averaging scheme in (6.1) to average h_g as given in (4.5) yields

$$\bar{h}_g(t;\delta) = D\frac{\delta^2 + 2[(\delta+1)e^{-\delta}-1]}{2\delta^2(1-e^{-\delta})} + \bar{h}. \tag{6.2}$$

It can be shown [6] that the term multiplying D is bounded between 1/3 (as in (4.7)) and 1/2; hence we expect the effect of the evanescent wave layer to be minimal. In order to plot some curves to verify this, we choose a typical decay length H_w from [25] and a typical gel width from [31]:

$$H_w = 9.5\times 10^{-6}\text{ cm}, \qquad H_g = 10^{-5}\text{ cm} \Longrightarrow \delta = 1.05. \tag{6.3}$$

As a naïve first approach, we might try to create a graph similar to Figure 4.1 to illustrate the changes of varying viscosity. However, we note from (4.4) and (6.2) that the only term involving μ_r is $\bar{h}$ (through the v_0 term). Hence the D term will always vanish when calculating

$$\bar{B}_1(\mu_r = \infty) - \bar{B}_1(\mu_r \neq \infty),$$

as in Figure 4.1. Since the D term includes the effect of decay, a graph of the above quantity will be the same whether or not decay is included.

The ERC solution does not have the unusual property described above, so to include the wave decay we substitute (6.2) into (4.8). In Figure 6.1 we plot the difference between the no-flow and viscous-flow cases of the ERC solution, including the wave decay. The graph is virtually indistinguishable from its analogue Figure 4.2, so the effect of decay is slight and the relative error is again around 0.04%.

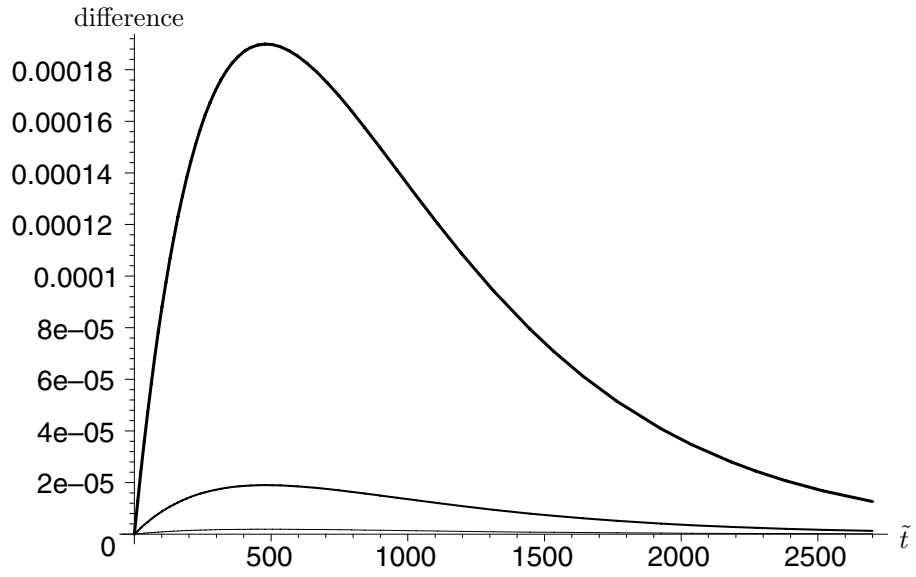


FIG. 6.1. *Difference between wave decay effective rate constant solution* $\bar{B}$ *of* (4.8) (*using* (6.1) *and* (6.2)) *with* $\mu_r = \infty$ (*solid dextran*) *and* μ_r *finite for* (*in decreasing order of thickness*) $\mu_r = 1$, 10, 100. *Relative difference is about* 0.04%. *Association experiment.*

For the dissociation experiment, the analysis is exactly analogous. The only difference is that rather than substituting (6.2) into (4.8), we substitute it into (4.12).

6.2. Results for moderate Da. The wave decay effect is more pronounced in the moderate Da case, in some cases leading to nonunique parameter estimates for the same data [6]. Proceeding in a manner analogous to (5.6a), we find that including the decay yields

$$\mathcal{I}[\beta;x] = \frac{\chi e^{-\nu_a x}}{2r_a\nu_a \cosh\lambda_a}\left[e^{\nu_a x} - 1 - \left|P\left(\frac{4}{3}, -\nu_a x\right)\right| + \left|P\left(\frac{5}{3}, -\nu_a x\right)\right|\right]$$
$$\times\left[\frac{e^{(\lambda_a-\delta)}-1}{\lambda_a-\delta} - \frac{e^{-(\lambda_a+\delta)}-1}{\lambda_a+\delta}\right]\frac{\delta}{1-e^{-\delta}}. \tag{6.4}$$

Note that the x-dependence is unchanged since the decay operates only in the y-direction.

Upon taking the limit for small $\tilde{k}_{on}$, the fact that $\lambda_a \to 0$ forces all the δ terms in the last line of (6.4) to cancel. Thus (5.7) still holds. This is because in this limit, the reaction is so slow that all transport effects are unimportant. Thus the binding will be uniform, and the wave decay cannot be discerned.

We demonstrate our results for varying μ_r in Figure 6.2. Since this is an association graph, the relevant restriction on k is given by (5.8), so the graphs end for different values of k. As before, the addition to S increases with decreasing μ_r, as lower μ_r means more convective transport, which enhances association. The actual value of S is essentially the negative of that shown in Figure 6.3, as can be seen from Figure 3 in [5]; hence the relative difference is about 1%.

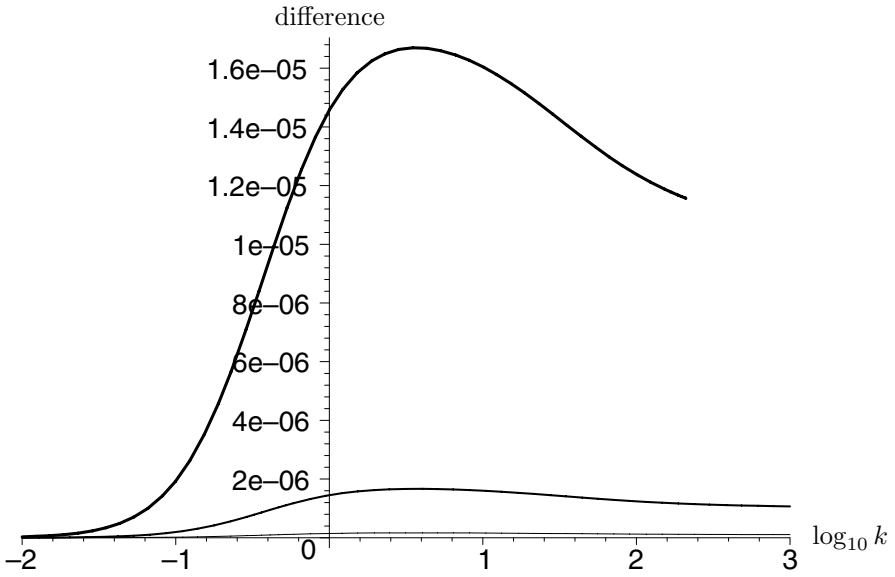


FIG. 6.2. *Absolute difference between S with $\mu_r = \infty$ (solid dextran) and μ_r finite versus $\log_{10} k$ for (in decreasing order of thickness) $\mu_r = 1$, 10, 100. Relative difference is about 1%. Association experiment, decay included.*

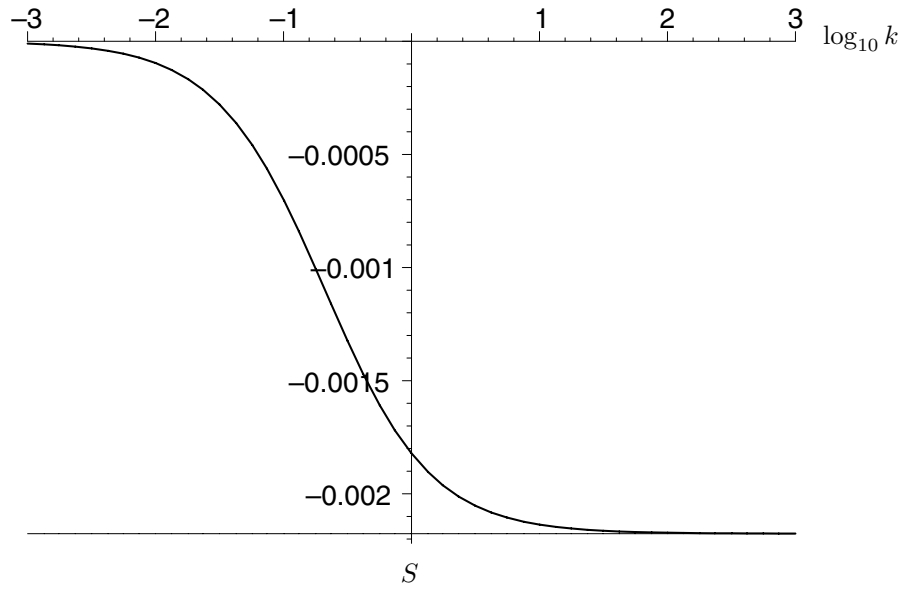


FIG. 6.3. *Thick line: S versus $\log_{10} k$, keeping $\tilde{k}_{\text{off}}$ fixed. Thin line: large-k asymptote. Dissociation experiment, $v_0 = 0$, wave decay considered.*

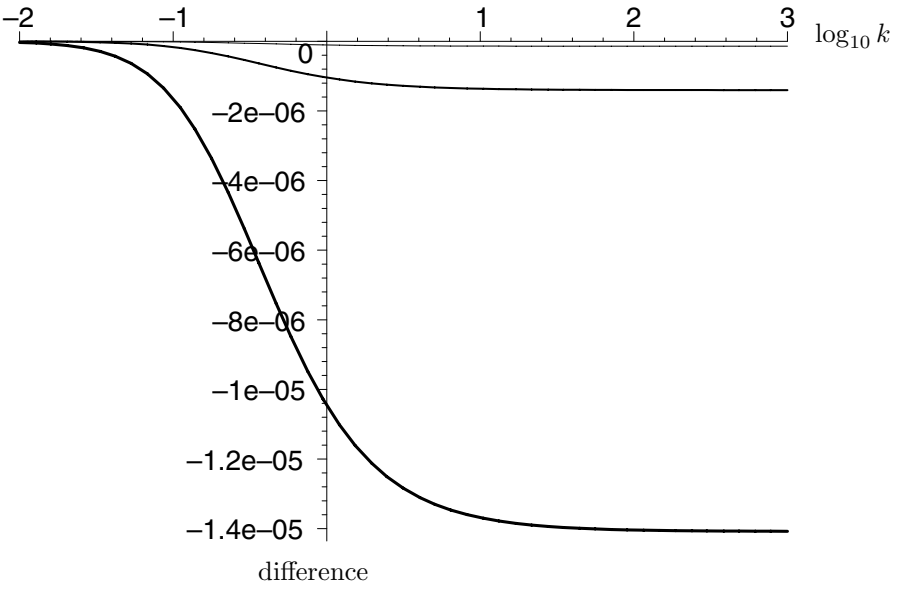


FIG. 6.4. *Absolute difference between S with $\mu_r = \infty$ (solid dextran) and μ_r finite versus $\log_{10} k$ for (in decreasing order of thickness) $\mu_r = 1$, 10, 100. Relative difference is about 1%. Dissociation experiment, decay included.*

For the dissociation case, the arguments are similar. The equation analogous to (5.9) is

$$\mathcal{I}[\beta; x] = -\frac{Ke^{-\nu_d x}}{2\alpha r_d \nu_d \cosh \lambda_d} \left[e^{\nu_d x} - 1 - \left| P\left(\frac{4}{3}, -\nu_d x\right)\right| + \left| P\left(\frac{5}{3}, -\nu_d x\right)\right|\right]$$

$$\times \left[\frac{e^{(\lambda_d - \delta)} - 1}{\lambda_d - \delta} - \frac{e^{-(\lambda_d + \delta)} - 1}{\lambda_d + \delta} \right] \frac{\delta}{1 - e^{-\delta}}. \tag{6.5}$$

As for the case without wave decay, earlier studies of the moderate Da case have kept K fixed and varied $\tilde{k}_{on}$ [6]. Thus, as in the previous subsection, in Figure 6.3 we present a demonstration graph to show what happens when we keep $\tilde{k}_{off}$ fixed instead. Again, the main difference lies in the asymptotes.

As above, when taking the limit for small $\tilde{k}_{on}$, the δ terms cancel. Thus the expression (5.12) for the small-$\tilde{k}_{on}$ asymptote still holds. For the large-$\tilde{k}_{on}$ asymptote in the no-flow case, we must use the expressions in (5.13). Substituting these parameters and our value of δ into (6.5) and (5.5a), we obtain $S = -2.18 \times 10^{-3}$, which is exactly the asymptote in Figure 6.3.

Lastly, we vary the viscosity ratio μ_r in Figure 6.4. Note that the corrections are again small and negative, as convection enhances dissociation. As in Figure 5.4, there are no restrictions on k because (5.14) is always satisfied. Therefore, our graphs go all the way to the right. Comparison with Figure 6.3 shows that the relative difference is again around 1%.

7. Conclusions. To explain BIAcore data that did not fit the traditional models, Witz [29] proposed that buffer flow from the channel penetrates into the dextran gel layer, enhancing transport. We have formulated a new model to include this effect. The key dimensionless parameter in this study is the small parameter H_r, which

measures the ratio of the widths of the gel and flow and hence characterizes the size of the velocity v_0 within the dextran gel. Its effect can be most readily seen in (3.7), where the slip condition at the flow-gel interface introduces a local depletion term that augments the integral depletion term from the no-flow case.

Since (3.9a) is a nonlinear equation, we obtained analytical results by introducing experimentally relevant simplifications. Most experiments are designed to have $\mathrm{Da} \ll 1$ to minimize transport effects, so we calculated the $O(\mathrm{Da})$ correction to the standard well-mixed case. The only effect of penetration is to introduce an additional term in $h(x)$, as defined in (4.4). We derived not only solution profiles for B, but also an ERC equation which can be used to fit sensogram data directly.

Some experiments cannot be designed such that $\mathrm{Da} \ll 1$, so we analyzed the moderate Da case by considering the short-time slope of the sensogram data. Corrections due to flow penetration appeared only in the parameter definitions in (5.4) and (5.10); the rest of the theory is the same as in the no-flow case [5]. The nature of our small-v_0 expansion dictated that we could not construct results for the case where $\mathrm{Da} \to \infty$; however, such conditions do not occur experimentally.

As in the small Da case, we examined both the association and dissociation phases of an experiment, providing (when possible) both the large- and small-$\tilde{k}_{\mathrm{on}}$ behavior of the short-time slope. In order to obtain results more consistent with experimental practice, we kept $\tilde{k}_{\mathrm{off}}$ fixed, in contrast to [5]. However, any differences between the papers were minor. In addition, since the inherent decay in the evanescent measuring wave affects only the averaging, not the transport, it was a simple matter to recast our previous results in this context.

In this manuscript we studied convection by modeling the dextran gel as a viscous fluid, though others have used more realistic polymer brush models [29]. Despite the simplicity of our model, the small thickness of the gel layer indicates that more realistic models will not produce qualitative changes in our results. We thus conclude that flow penetration effects are not likely to explain anomalous BIAcore measurements, and other effects, such as steric hindrance effects or conformational changes, should be investigated instead.

Moreover, since the size of the penetration effects are dictated by geometry, rather than properties of the gel, flow, or reactants, these insights can be extended to many similar physical systems. In particular, one may use gels and other compounds in reacting zones to reduce the size of hydrodynamic convective effects (as in [14]).

Acknowledgments. Portions of this manuscript were prepared during sabbatical stays at the Mathematical Biosciences Institute at The Ohio State University and the University of Maryland, Baltimore County.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, Appl. Math. Ser. 155, U. S. Department of Commerce, Washington, DC, 1972.

[2] S. J. Davis, E. A. Davies, A. N. Barclay, S. Daenke, D. L. Bodian, E. Y. Jones, D. I. Stuart, T. D. Butters, R. A. Dwek, and P. A. Van der Merwe, *Ligand binding by the immunoglobulin superfamily recognition molecule CD2 is glycosylation-independent*, J. Biol. Chem., 270 (1995), pp. 369–375.

[3] D. A. Edwards, *Estimating rate constants in a convection-diffusion system with a boundary reaction*, IMA J. Appl. Math., 63 (1999), pp. 89–112.

[4] D. A. Edwards, *Biochemical reactions on helical structures*, SIAM J. Appl. Math., 60 (2000), pp. 1425–1446.

[5] D. A. Edwards, *The effect of a receptor layer on the measurement of rate constants*, Bull. Math. Biol., 63 (2001), pp. 301–327.

[6] D. A. Edwards, *Refining the measurement of rate constants in the BIAcore*, J. Math. Biol., 49 (2004), pp. 272–292.
[7] D. A. Edwards, *Convection effects in the BIAcore dextran layer: surface reaction model*, Bull. Math. Biol., 68 (2006), pp. 627–634.
[8] D. A. Edwards, B. Goldstein, and D. S. Cohen, *Transport effects on surface-volume biological reactions*, J. Math. Biol., 39 (1999), pp. 533–561.
[9] D. A. Edwards and S. Swaminathan, *The effect of receptor site nonuniformity on the measurement of rate constants*, Appl. Math. Lett., 18 (2005), pp. 1101–1107.
[10] P. B. Garland, *Optical evanescent wave methods for the study of biomolecular reactions*, Quart. Rev. Biophys., 29 (1996), pp. 91–117.
[11] B. Goldstein and M. Dembo, *Approximating the effects of diffusion on reversible reactions at the cell surface: Ligand-receptor kinetics*, Biophys. J., 68 (1995), pp. 1222–1230.
[12] E. F. Grabowski, L. I. Friedman, and E. F. Leonard, *Effects of shear rate on the diffusion and adhesion of blood platelets to a foreign surface*, Ind. Eng. Chem. Fund., 11 (1972), pp. 224–232.
[13] X. Y. He, N. Li, and B. Goldstein, *Lattice Boltzmann simulation of diffusion-convection systems with surface chemical reaction*, Molec. Sim., 25 (2000), pp. 145–156.
[14] J. Jansen and B. Niemeyer, *Automated high-pressure plant for a continuous flow through a fixed bed investigation of hydrodynamic behaviour*, J. Supercrit. Fluids, 33 (2005), pp. 283–291.
[15] L. Joss, T. A. Morton, M. L. Doyle, and D. G. Myszka, *Interpreting kinetic rate constants from optical biosensor data recorded on a decaying surface*, Anal. Biochem., 261 (1998), pp. 203–210.
[16] R. Karlsson and A. Fält, *Experimental design for kinetic analysis of protein-protein interactions with surface plasmon resonance biosensors*, J. Immun. Methods, 200 (1997), pp. 121–133.
[17] R. Karlsson, A. Michaelson, and L. Mattson, *Kinetic analysis of monoclonal antibody-antigen interactions with a new biosensor based analytical system*, J. Immun. Methods, 145 (1991), pp. 229–240.
[18] B. Liedberg, I. Lundstrom, and E. Stenberg, *Principles of biosensing with an extended coupling matrix and surface-plasmon resonance*, Sens. Actuators B, 11 (1993), pp. 63–72.
[19] W. M. Long and L. V. Kalachev, *Asymptotic analysis of dissolution of a spherical bubble (case of fast reaction outside the bubble)*, Rocky Mountain J. Math., 30 (2000), pp. 293–313.
[20] S. Mann, S. L. Burkett, S. A. Davis, C. E. Fowler, N. H. Mendelson, S. D. Sims, D. Walsh, and N. T. Whilton, *Sol-gel synthesis of organized matter*, Chem. Mater., 9 (1997), pp. 2300–2310.
[21] T. Mason, A. R. Pineda, C. Wofsy, and B. Goldstein, *Effective rate models for the analysis of transport-dependent biosensor data*, Math. Biosci., 159 (1999), pp. 123–144.
[22] D. G. Myszka, X. He, M. Dembo, T. A. Morton, and B. Goldstein, *Extending the range of rate constants available from BIAcore: Interpreting mass transport influenced binding data*, Biophys. J., 75 (1998), pp. 583–594.
[23] D. J. O'Shannessy, M. Brigham-Burke, and K. Peck, *Immobilization chemistries suitable for use in the BIAcore surface plasmon resonance detector*, Anal. Biochem., 205 (1992), pp. 132–136.
[24] S. Qian, *private communication*, University of Pennsylvania, Philadelphia, PA, 2004.
[25] P. Schuck, *Kinetics of ligand binding to receptor immobilized in a polymer matrix, as detected with an evanescent wave biosensor.* I. *A computer simulation of the influence of mass transport*, Biophys. J., 70 (1996), pp. 1230–1249.
[26] A. Szabo, L. Stolz, and R. Granzow, *Surface plasmon resonance and its use in bio-molecular interaction analysis (BIA)*, Curr. Opin. Struct. Biol., 5 (1995), pp. 699–705.
[27] H. Treml, S. Woelki, and H.-H. Kohler, *Theory of capillary formation in alginate gels*, Chem. Phys., 3 (2003), pp. 341–353.
[28] L. D. Ward and D. J. Winzor, *Relative merits of optical biosensors based on flow-cell and cuvette designs*, Anal. Biochem., 285 (2000), pp. 179–193.
[29] J. Witz, *Kinetic analysis of analyte binding by optical biosensors: Hydrodynamic penetration of the analyte flow into the polymer matrix reduces the influence of mass transport*, Anal. Biochem., 270 (1999), pp. 201–206.
[30] C. Wofsy and B. Goldstein, *Effective rate models for receptors distributed in a layer above a surface: Application to cells and BIAcore*, Biophys. J., 82 (2002), pp. 1743–1755.
[31] M. L. Yarmush, D. B. Patankar, and D. M. Yarmush, *An analysis of transport resistance in the operation of BIAcore™; Implications for kinetic studies of biospecific interactions*, Molec. Immunol., 33 (1996), pp. 1203–1214.

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 1998–2026

MODEL DEVELOPMENT FOR ATOMIC FORCE MICROSCOPE STAGE MECHANISMS*

RALPH C. SMITH†, ANDREW G. HATCH†, TATHAGATA DE‡, MURTI V. SALAPAKA‡, RICARDO C. H. DEL ROSARIO§, AND JULIE K. RAYE¶

Abstract. In this paper, we develop nonlinear constitutive equations and resulting system models quantifying the nonlinear and hysteretic field-displacement relations inherent to lead zirconate titanate (PZT) devices employed in atomic force microscope stage mechanisms. We focus specifically on PZT rods utilizing d_{33} motion and PZT shells driven in d_{31} regimes, but the modeling framework is sufficiently general to accommodate a variety of drive geometries. In the first step of the model development, lattice-level energy relations are combined with stochastic homogenization techniques to construct nonlinear constitutive relations which accommodate the hysteresis inherent to ferroelectric compounds. Second, these constitutive relations are employed in classical rod and shell relations to construct system models appropriate for presently employed nanopositioner designs. The capability of the models for quantifying the frequency-dependent hysteresis inherent to the PZT stages is illustrated through comparison with experimental data.

Key words. atomic force microscope, hysteresis model, dynamics

AMS subject classifications. 74D10, 74M05

DOI. 10.1137/05063307X

1. Introduction.

Stage mechanisms employing the ferroelectric material lead zirconate titanate (PZT) have played a fundamental role in scanning tunneling microscope (STM) and atomic force microscope (AFM) design since their inception due to the high set point accuracy, large dynamic range, and relatively small temperature sensitivity exhibited by the compounds [14]. To illustrate, consider the prototypical AFM design depicted in Figure 1. To ascertain the three-dimensional (3-D) surface structure of a sample, it is moved laterally along a predetermined x-y grid by a PZT-driven stage. The response of a highly flexible microcantilever to changing atomic surface forces is monitored by a reflected laser beam measured via a photodiode, and forces corresponding to the cantilever displacement changes are determined via Hooke's law. A feedback law is used to determine voltages to a transverse PZT stage which produces displacements in the z-direction to maintain constant forces. A complete scan in this manner provides a surface image of the compounds. Additionally, PZT actuators are often used to drive the microcantilevers at resonance to achieve the tapping mode operation used to reduce damage to specimens. The reader is referred

*Received by the editors June 5, 2005; accepted for publication (in revised form) April 24, 2006; published electronically October 3, 2006.

http://www.siam.org/journals/siap/66-6/63307.html

†Department of Mathematics, Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695 (rsmith@eos.ncsu.edu, aghatch@comcast.net). The research of the first author was supported in part through NSF grant CMS-0099764 and in part by the Air Force Office of Scientific Research through the grants AFOSR-F49620-01-1-0107 and AFOSR-FA9550-04-1-0203. The research of the second author was supported by DARPA subcontract 1000-G-CF980. Both authors were supported by NSF grant CMS-0201560.

‡Electrical Engineering Department, Iowa State University, Ames, IA 50011 (tatha@iastate.edu, murti@iastate.edu). The research of these authors was supported by NSF grant CMS-0201560.

§Department of Mathematics, University of the Philippines, Diliman, Quezon City 1101 (rcdelros@math.upd.edu.ph).

¶Department of Mathematics, Virginia Commonwealth University, Richmond VA 23284 (jkraye@yahoo.com).

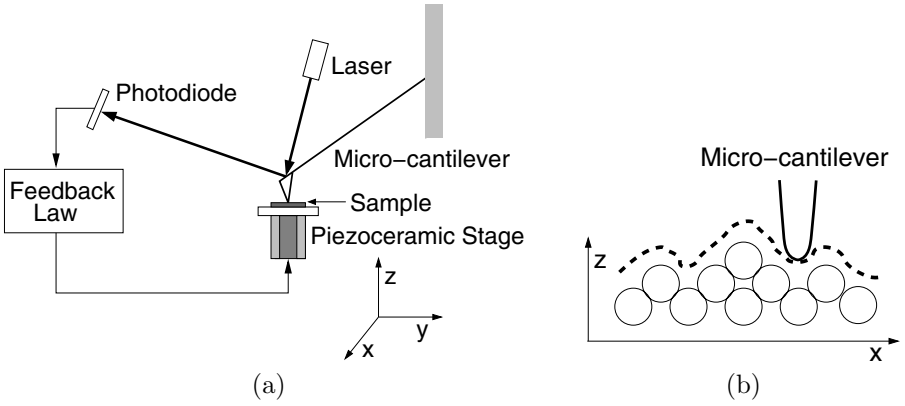


FIG. 1. (a) *Configuration of a prototypical AFM, and* (b) *surface image determined by one lateral sweep.*

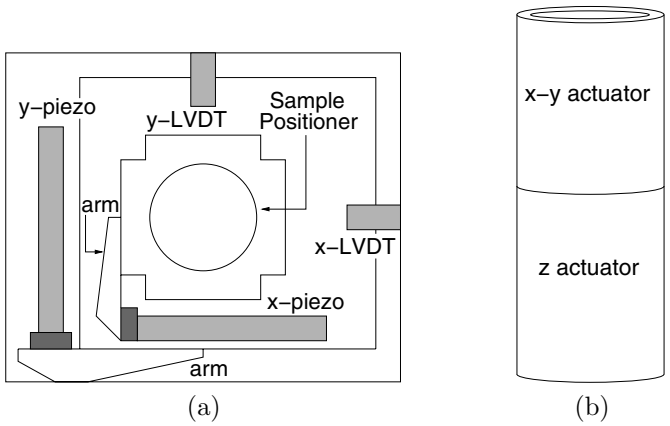


FIG. 2. *Actuator configurations employed for sample positioning in AFM:* (a) *stacked actuators employed as x- and y-stages, and* (b) *cylindrical PZT transducer.*

to [11, 13, 14] for additional details regarding AFM applications and design.

Two representative stage designs are depicted in Figure 2. The first employs stacked PZT actuators utilizing d_{33} electromechanical motion to achieve longitudinal positioning along the prespecified x-y grid. A second stage provides the transverse motion required to ascertain the sample topography. Rod models with linear and nonlinear electromechanical input relations are constructed to quantify the PZT transducer dynamics in this design. The second geometry employs a cylindrical shell—with half poled d_{33} to provide horizontal (x-y) motion and half poled d_{31} for vertical (z) motion, as depicted in Figure 2(b)—to enhance vibration isolation and reduce hysteresis and constitutive nonlinearities. Thin shell models are developed to characterize this stage design.

To illustrate issues which must be addressed by models, field-displacement data from the stacked actuator depicted in Figure 2(a) is plotted in Figures 3 and 4. The data in Figure 3 was collected at 0.1 Hz and illustrates the nested, hysteretic relation between input fields and generated displacements in a nearly quasi-static regime. The data in Figure 4 was collected at frequencies ranging from 0.279 Hz to 27.9 Hz

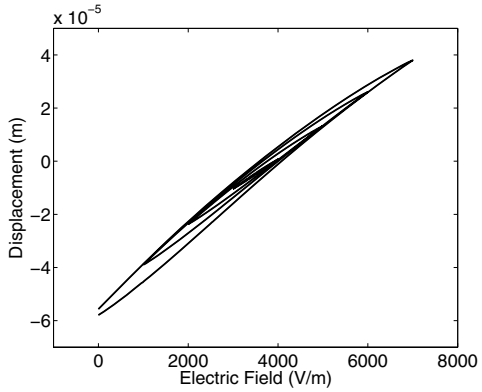


FIG. 3. *Nested minor loops in* 0.1 *Hz field-displacement data from a stacked PZT stage of the type depicted Figure* 2(a).

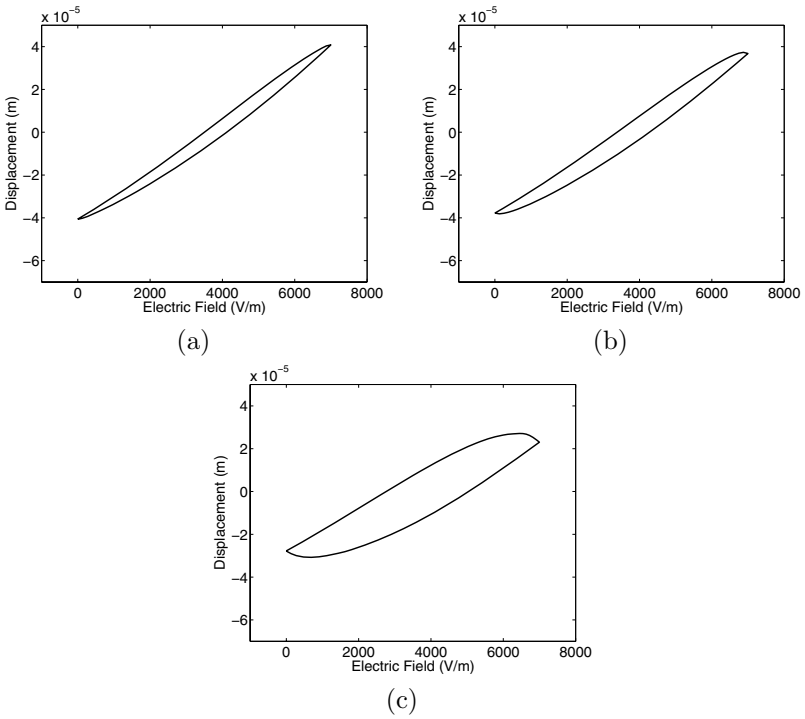


FIG. 4. *Frequency-dependent field-displacement behavior of a stacked PZT stage of the type depicted in Figure* 2(a)*: sample rates of* (a) 0.279 *Hz,* (b) 5.58 *Hz, and* (c) 27.9 *Hz.*

to illustrate the frequency-dependence of the hysteresis as well as certain dynamic effects.

At low frequencies, the inherent hysteresis can be accommodated through proportional-integral-derivative (PID) or robust control designs [7, 8, 22, 29]. However, at the higher frequencies required for applications including real-time monitoring of biological processes (e.g., protein unfolding), comprehensive product diagnostics, and single electron spin detection [28, 40], increasing noise-to-data ratios and diminishing

high-pass characteristics of control filters preclude a sole reliance on feedback laws to eliminate hysteresis.

Alternatively, it is illustrated in [20, 21] that use of charge- or current-controlled amplifiers can essentially eliminate hysteresis. However, this mode of operation can be prohibitively expensive when compared with the more commonly employed voltage-controlled amplifiers, and current control is ineffective if maintaining DC offsets, as is the case when the x-stage of an AFM is held in a fixed position while a sweep is performed with the y-stage.

The need to significantly increase scanning speeds with general amplifiers motivates the development of models and model-based control designs which accommodate the frequency-dependent hysteresis inherent to the PZT actuators employed in the AFM stages. As detailed in [30], there exist a number of general approaches and frameworks for quantifying the constitutive nonlinearities and hysteresis in the general class of ferroelectric materials which encompass PZT. These include phenomenological macroscopic models [24], Preisach models [12, 27], domain wall models [33, 34], micromechanical models [6, 18, 19], mesoscopic energy relations [5, 17], and homogenized energy models [32, 39]. Within the context of AFM design, Croft, Shed, and Devasia [7] have employed a combination of a viscoelastic creep model and nonlinear Preisach representation to compensate for hysteresis and creep in an AFM stage, whereas a domain wall model was employed in [35] for the characterization of hysteresis in certain stage constructs. Primary requirements for nonlinear hysteresis models for the PZT actuators in an AFM are (i) flexibility with regard to frequency-dependent hysteresis effects (the frameworks of [7, 35] are limited in this regard), (ii) exact or approximate invertibility for linear control design, and (iii) sufficient efficiency for real-time implementation at the speeds required for present and future applications.

In this paper, we develop AFM transducer models, based on a homogenized energy framework for characterizing hysteresis and constitutive nonlinearities in ferroelectric materials, which meet these criteria. In section 2, we summarize the framework developed in [15, 31, 37, 38, 39] for quantifying hysteresis in the field-polarization relation and develop constitutive equations which characterize the elastic and electromechanical behavior of the PZT material. These constitutive relations are employed in section 3 to construct rod and shell models for the stages depicted in Figure 2, and the well-posedness of the models is established in section 4. Numerical approximation techniques for the transducer models are summarized in section 5, and the capability of the framework to quantify the biased and frequency-dependent hysteresis behavior of the transducers is illustrated in section 6 through a comparison with the experimental data plotted in Figures 2 and 3.

To place this framework in perspective, we briefly summarize the manner in which it compares and contrasts with previous models. As illustrated in [30, 36], the homogenized energy framework provides an energy basis for certain extended Preisach models. However, it also differs in five fundamental ways, and these are detailed in Remark 4 in section 2.4, following the model development. The domain wall model employed in the constitutive relations of [35] is efficient for characterizing hysteresis when inputs are known a priori. As detailed in [30], however, it does not guarantee the closure of biased minor loops in quasi-static regimes, nor does it provide the capability for including frequency-dependent effects due to thermal activation or creep. Hence the present framework, which automatically incorporates these mechanisms, provides significantly more flexibility for the range of operating regimes required for present and future atomic force microscopy applications. Certain preliminary aspects

of the model are presented in [16], which includes the rod model with initial model validation as a prelude for open loop control design. The present work significantly extends that modeling framework through the unified consideration of rod and shell geometries with extensive experimental validation and the presentation of analysis to establish model well-posedness.

The capability of the framework for characterizing frequency-dependent effects, and hence achieving criterion (i), is illustrated in section 6. With regard to criteria (ii) and (iii), the construction and experimental implementation of model inverses to linearize the nonlinear dynamics is demonstrated in [16]. Hence the models provide a framework for characterizing the hysteresis and nonlinear dynamics inherent to PZT-based nanopositioners in a manner which promotes stage and control design.

2. Constitutive relations. In this section, we summarize the development of constitutive relations which quantify the nonlinear and hysteretic map between input fields E and stresses σ and the polarization P and strains ε generated in ferroelectric materials. These relations are developed in three steps. In the first, Helmholtz and Gibbs energy relations are constructed at the lattice level to quantify the *local* dependence of P and ε on E and σ for regimes in which relaxation due to thermal processes is either negligible or significant. In the second step of the development, material nonhomogeneities, polycrystallinity, and variable field effects are incorporated through the assumption that certain material properties are manifestations of underlying distributions rather than constants. Stochastic homogenization in this manner yields macroscopic models which quantify the bulk hysteretic E-P behavior measured in ferroelectric materials. Finally, necessary conditions associated with minimization of the Gibbs energy are invoked to obtain 1-D and 2-D constitutive relations quantifying the elastic and electromechanical behavior of the transducer materials.

2.1. Helmholtz and Gibbs energy relations. As detailed in [39], an appropriate Helmholtz energy relation is

$$\psi(P,\varepsilon) = \psi_P(P) + \frac{1}{2}Y\varepsilon^2 - a_1\varepsilon P - a_2\varepsilon P^2, \tag{1}$$

where the component

$$\psi_P(P) = \begin{cases} \frac{1}{2}\eta(P+P_R)^2, & P \le -P_I, \\ \frac{1}{2}\eta(P-P_R)^2, & P \ge P_I, \\ \frac{1}{2}\eta(P_I-P_R)\left(\frac{P^2}{P_I}-P_R\right), & |P| < P_I, \end{cases}$$

quantifies the internal energy due to dipole processes. As shown in Figure 5, P_I is the positive inflection point that delineates the transition between stable and unstable regions, P_0 denotes the unstable equilibrium, and P_R is the value of P at which the positive local minimum of ψ occurs. The parameter η is the reciprocal of the slope of the E-P relation after switching occurs. The second term on the right-hand side of (1) quantifies the elastic energy, whereas the third and fourth terms quantify electromechanical coupling effects. Here Y denotes the Young's modulus, and a_1 and a_2 are electromechanical coupling coefficients.

The Gibbs energy relation

$$G(E,\sigma,P,\varepsilon) = \psi_P(P) + \frac{1}{2}Y\varepsilon^2 - a_1\varepsilon P - a_2\varepsilon P^2 - EP - \sigma\varepsilon \tag{2}$$

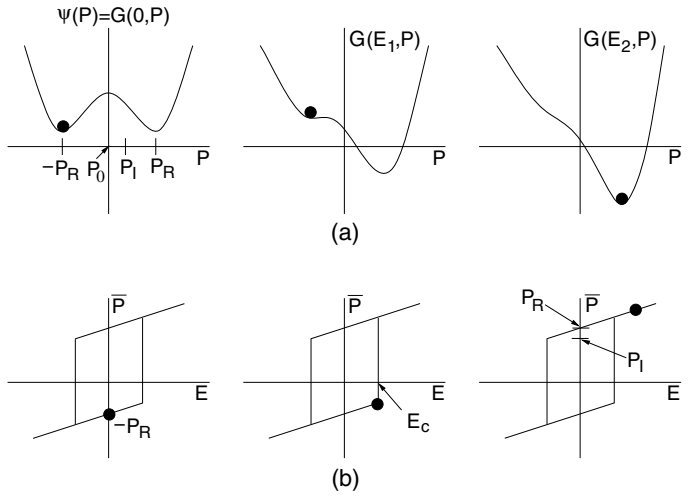


FIG. 5. (a) *Helmholtz energy ψ and Gibbs energy G for $\sigma = 0$ and increasing fields E.* (b) *Switch in the local polarization $\overline{P}$ that occurs as E is increased beyond the local coercive field E_c given by* (4) *in the absence of thermal activation.*

incorporates the elastic work $\sigma\varepsilon$ and electromechanical work EP. This provides the functional that is minimized or balanced with the relative thermal energy to provide local E-P relations and global electromechanical constitutive equations. The reader is referred to [30, 32] for details regarding the manner in which the Gibbs energy incorporates the dependent variables ε and P in terms of the independent variables σ and E.

2.2. Polarization kernel—Negligible thermal activation. For operating regimes in which relaxation or creep due to thermal processes is negligible, the local E-P relation is determined from the equilibrium conditions

$$\frac{\partial G}{\partial P} = 0, \qquad \frac{\partial^2 G}{\partial P^2} > 0.$$

For the piecewise quadratic functional (2), this yields a polarization kernel of the form

$$\overline{P}(E) = \frac{E}{\eta - 2a_2\varepsilon} + \delta\frac{P_R\eta + \delta a_1\varepsilon}{\eta - 2a_2\varepsilon}, \tag{3}$$

where $\delta = 1$ for positively oriented dipoles and $\delta = -1$ for those having negative orientation. To specify δ, and hence $\overline{P}$, in terms of the initial dipole configurations and previous switches, we let $\delta_0 = \pm 1$ designate the initial dipole orientation and let

$$E_c = \eta(P_R - P_I) \tag{4}$$

define the local coercive field at which the negative well ceases to exist and hence a dipole switch occurs. The local polarization is then given by

$$[\overline{P}(E; E_c, \delta_0)](t) = \begin{cases} \frac{E}{\eta - 2a_2\varepsilon} + \delta_0\frac{P_R\eta + \delta_0 a_1\varepsilon}{\eta - 2a_2\varepsilon}, & \tau = \emptyset, \\ \frac{E(t) - P_R\eta + a_1\varepsilon}{\eta - 2a_2\varepsilon}, & \tau \neq \emptyset,\ E(\max\tau) = -E_c, \\ \frac{E(t) + P_R\eta + a_1\varepsilon}{\eta - 2a_2\varepsilon}, & \tau \neq \emptyset,\ E(\max\tau) = E_c. \end{cases} \tag{5}$$

Here $\emptyset$ denotes the empty set, and the set of transition times is designated by

$$\tau = \{t \in (0, t_f] \,|\, E(t) = -E_c \text{ or } E(t) = E_c\},$$

where t_f denotes the final time under consideration.

Remark 1. For the drive levels employed for nanopositioning, the stress effects on the polarization are typically negligible, which motivates taking $\varepsilon = 0$ in (3) and (5). Hence the relations

$$\overline{P}(E) = \frac{1}{\eta}E + P_R\delta$$

or

$$[\overline{P}(E; E_c, \delta_0)](t) = \begin{cases} \frac{E(t)}{\eta} + P_R\delta_0, & \tau = \emptyset, \\ \frac{E(t)}{\eta} - P_R, & \tau \neq \emptyset,\ E(\max\tau) = -E_c, \\ \frac{E(t)}{\eta} + P_R, & \tau \neq \emptyset,\ E(\max\tau) = E_c, \end{cases} \tag{6}$$

are usually employed when characterizing AFM stages.

2.3. Polarization kernel—Thermal activation. If thermal relaxation or creep is significant, the Gibbs energy G and relative thermal energy kT/V are balanced through the Boltzmann relation

$$\mu(G) = Ce^{-GV/kT}. \tag{7}$$

Here k is Boltzmann's constant, V denotes a reference volume chosen to ensure physical relaxation behavior, and C is chosen to ensure integration to unity for the complete set of admissible inputs. As detailed in [30, 39], this yields the local polarization relation

$$\overline{P} = x_+ \langle P_+ \rangle + x_- \langle P_- \rangle . \tag{8}$$

The fractions x_+ and x_- of positively and negatively oriented dipoles are quantified by the differential equations

$$\dot{x}_+ = -p_{+-}x_+ + p_{-+}x_-,$$
$$\dot{x}_- = -p_{-+}x_- + p_{+-}x_+,$$

which can be simplified to

$$\dot{x}_+ = -p_{+-}x_+ + p_{-+}(1 - x_+)$$

through the identity

$$\dot{x}_+ + \dot{x}_- = 1.$$

The expected polarizations due to positively and negatively oriented dipoles are

$$\langle P_+ \rangle = \frac{\int_{P_I}^{\infty} Pe^{-G(E,P)V/kT}dP}{\int_{P_I}^{\infty} e^{-G(E,P)V/kT}dP}, \qquad \langle P_- \rangle = \frac{\int_{-\infty}^{-P_I} Pe^{-G(E,P)V/kT}dP}{\int_{-\infty}^{-P_I} e^{-G(E,P)V/kT}dP}, \tag{9}$$

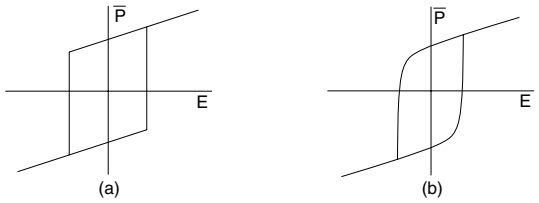


FIG. 6. *Hysteron provided by* (a) *the relation* (6) *with negligible thermal relaxation, and* (b) *the relation* (8)*, which incorporates relaxation mechanisms.*

where the denominator results from the evaluation of C in (7). The likelihoods of switching from positive to negative, and conversely, are given by

$$p_{+-} = \frac{1}{\mathcal{T}(T)} \frac{\int_{P_I-\epsilon}^{P_I} e^{-G(E,P)V/kT} dP}{\int_{P_I-\epsilon}^{\infty} e^{-G(E,P)V/kT} dP}, \quad p_{-+} = \frac{1}{\mathcal{T}(T)} \frac{\int_{-P_I}^{-P_I+\epsilon} e^{-G(E,P)V/kT} dP}{\int_{-\infty}^{-P_I+\epsilon} e^{-G(E,P)V/kT} dP}, \tag{10}$$

where ϵ is taken to be a small positive constant. The relaxation time $\mathcal{T}$ is the reciprocal of the frequency at which dipoles attempt to switch. It is proven in [30, 39] that $\overline{P}$ given by (8) converges to the local polarization (6) in the limit $kT/V \to 0$ of negligible thermal activation.

Remark 2. When constructing the expected polarization relations (9) and likelihoods (10), we use the notation $G(E, P)$ to indicate that we take $\varepsilon = \sigma = 0$ in (2) in accordance with the assumption that stress effects on the polarization are negligible at the drive levels employed in AFM stages. This approximation is employed only when defining the stress-independent polarization, and the full expression (2) is employed when constructing elastic constitutive relations in section 2.5.

2.4. Macroscopic polarization model. The local polarization relations (6) and (8) exhibit the behavior depicted in Figure 6 and provide reasonable characterization of the E-P behavior of certain single crystal compounds. However, to incorporate the effects of material and stress nonhomogeneities, polycrystallinity, and variable effective fields $E_e = E + E_I$, we assume that the interaction field E_I and local coercive field E_c given by (4) are manifestations of underlying distributions rather than constants. If we designate the associated densities by ν_1 and ν_2, the macroscopic field-polarization behavior is quantified by the relation

$$[P(E)](t) = \int_0^{\infty} \int_{-\infty}^{\infty} \nu_1(E_c)\nu_2(E_I)[\overline{P}(E + E_I; E_c, \xi)](t)\, dE_I\, dE_c, \tag{11}$$

where the kernel $\overline{P}$ is given by (6) or (8).

As detailed in [30, 32], the densities ν_1 and ν_2 are assumed to satisfy the physical criteria

$$\begin{aligned} &\text{(i)} && \nu_1(x) \ \text{ defined for } \ x > 0, \\ &\text{(ii)} && \nu_2(-x) = \nu_2(x), \\ &\text{(iii)} && |\nu_1(x)| \le c_1 e^{-a_1 x}, \\ & && |\nu_2(x)| \le c_2 e^{-a_2|x|} \end{aligned} \tag{12}$$

for positive c_1, a_1, c_2, a_2. The restricted domain in (i) reflects the fact that the coercive field E_c is positive, whereas the symmetry enforced in the interaction field through (ii) yields the symmetry observed in low-field Rayleigh loops. Hypothesis (iii) incorporates the physical observation that the coercive and interaction fields decay as a function of distance, and guarantees that integration against the piecewise linear kernel yields finite polarization values.

2.4.1. Model implementation. Approximation of (11) through Gaussian quadrature techniques yields the approximate relation

$$[P(E)](t) = \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \nu_1(E_{c_i}) \nu_2(E_{I_j}) [\overline{P}(E_{I_j} + E; E_{c_i}, \xi_j)](t) v_i w_j, \tag{13}$$

where E_{I_j}, E_{c_i} denote the abscissas associated with respective quadrature formulae and v_i, w_j are the respective weights—e.g., see [30]. Highly efficient algorithms for implementing the approximate polarization model (13) with the kernel (6) for the case of negligible thermal activation can be found in [30, 39]. Algorithms for implementing the model with the thermally active kernel (8) are presented in [4]. MATLAB code for both cases can be accessed at the website http://www.siam.org/books/fr32 associated with [30].

2.4.2. Density estimation. Techniques for identifying the densities ν_1 and ν_2 are illustrated in [30, 32]. For certain applications, reasonable accuracy is provided by a priori functions satisfying the physical criteria (12) and having a small number of parameters to be estimated through least squares fits to data—e.g., variances and means in normal and lognormal relations. For more general applications requiring high accuracy for a wide range of operating conditions, the $N_i + N_j$ discretized density values $\nu_1(E_{c_i})$ and $\nu_2(E_{I_j})$ can be estimated through the least squares techniques detailed in [30, 32].

Remark 3. From the perspective of both numerical and experimental implementation and the establishment of the well-posedness of resulting transducer models, it is important to quantify the regularity between input fields and the polization predicted by (11). In the appendix, it is established that P given by (11) is continuous with respect to E.

Remark 4. The formulation of the model as a superposition of kernels bears some resemblance to Preisach models, and it is illustrated in [30, 36] that the framework provides an energy basis for certain extended Preisach formulations. However, the energy framework differs from the classical Preisach model, characterized by the properties of deletion and congruency, in five aspects which prove crucial for actuator characterization and model-based control design. (i) For certain density choices, parameters can be correlated with attributes of the data to facilitate model construction and updating. (ii) The incorporation of relative thermal energy provides the thermal activation mechanisms required to characterize relaxation and creep. (iii) Stress and temperature-dependencies (e.g., see [1, 26]) are incorporated into the kernel rather than weights, as is the case for Preisach models, which eliminates the necessity of vector-valued lookup tables. (iv) Derivation of the kernels using Ising theory yields hysterons which accommodate measured noncongruencies and avoids the input or output-dependent densities associated with Preisach models. (v) The framework automatically incorporates low-field reversible behavior without the extensions required by Preisach theory.

2.5. Constitutive relations. To obtain elastic constitutive relations, the equilibrium condition

$$\frac{\partial G}{\partial \varepsilon} = 0$$

is invoked to obtain

$$\sigma = Y\varepsilon - a_1 P - a_2 P^2,$$

which reduces to Hooke's law when $P = 0$. To incorporate internal damping, we posit that in the absence of electromechanical effects, stress is proportional to a linear combination of strain and strain rate (the Kelvin–Voigt damping hypothesis). Finally, we note that the PZT stage mechanisms are poled and hence operate about the remanence polarization $P = P_R$ rather than the depoled state $P = 0$. (The remanence polarization is that which remains when the applied field is reduced to zero following positive saturation.) When combined with the polarization model (11), this yields the 1-D constitutive relations

$$\begin{aligned} &\sigma = Y\varepsilon + C\dot{\varepsilon} - a_1(P - P_R) - a_2(P - P_R)^2, \\ &[P(E)](t) = \int_0^\infty \int_{-\infty}^\infty \nu_1(E_c)\nu_2(E_I)[\overline{P}(E + E_I; E_c, \xi)](t)\, dE_I\, dE_c, \end{aligned} \tag{14}$$

where C is the Kelvin–Voigt damping coefficient. These relations are employed when constructing rod models to characterize the hysteretic dynamics shown in Figures 3 and 4 for the stacked actuators employed in the stage construction depicted in Figure 2(a).

The constitutive behavior of the PZT shell depicted in Figure 2(b) differs from that of the rod in two fundamental aspects: (i) the longitudinal actuation is due to d_{31} rather than d_{33} electromechanical coupling mechanisms and (ii) longitudinal and circumferential stresses and strains are coupled due to the curvature. To designate the coupled material behavior, we let ε_x, σ_x and $\varepsilon_\theta, \sigma_\theta$ denote the normal strains and stresses in the longitudinal and circumferential directions, respectively, and we denote shear strains and stresses by $\varepsilon_{x\theta}$ and $\sigma_{x\theta}$. Finally, we let ν denote the Poisson ratio for the material. The resulting 2-D constitutive relations

$$\begin{aligned} \sigma_x &= \frac{Y}{1-\nu^2}(\varepsilon_x + \nu\varepsilon_\theta) + \frac{C}{1-\nu^2}(\dot{\varepsilon}_x + \nu\dot{\varepsilon}_\theta) - \frac{1}{1-\nu}\left[a_1(P - P_R) + a_2(P - P_R)^2\right], \\ \sigma_\theta &= \frac{Y}{1-\nu^2}(\varepsilon_\theta + \nu\varepsilon_x) + \frac{C}{1-\nu^2}(\dot{\varepsilon}_\theta + \nu\dot{\varepsilon}_x) - \frac{1}{1-\nu}\left[a_1(P - P_R) + a_2(P - P_R)^2\right], \\ \sigma_{x\theta} &= \frac{Y}{2(1+\nu)}\varepsilon_{x\theta} + \frac{C}{2(1+\nu)}\dot{\varepsilon}_{x\theta}, \\ [P(E)](t) &= \int_0^\infty \int_{-\infty}^\infty \nu_1(E_c)\nu_2(E_I)[\overline{P}(E + E_I; E_c, \xi)](t)\, dE_I\, dE_c \end{aligned} \tag{15}$$

are employed when constructing transducer models for cylindrical nanopositioning stages.

3. Transducer models for stacked and cylindrical AFM stages. We now employ the 1-D constitutive relation (14) and 2-D relation (15) to construct models for the stacked and cylindrical AFM stages depicted in Figure 2. For the stacked actuator, we consider two frameworks: (i) a distributed PDE model, which quantifies displacements along the rod length as a function of the input field, and (ii) a lumped model, which exploits the assumption of uniform stresses and fields along the rod length to motivate an ODE quantifying displacements only at the rod end. A comparison between characterization capabilities provided by the two frameworks is provided in section 6. For the cylindrical shell design, we summarize a Donnell–Mushtari model which quantifies vertical motion provided by the z-component of the stage depicted in Figure 2(b).

3.1. Rod model for the stacked actuator.

3.1.1. Distributed rod model. We consider first the development of a distributed rod model which quantifies the displacement $u(t,x)$ along the rod length. In accordance with present stage design, one end of the rod is assumed fixed, while the other encounters resistance due to the connecting mechanisms, as depicted in Figure 7. We assume that this latter contribution can be modeled as a damped elastic system with mass m_ℓ, stiffness k_ℓ, and damping coefficient c_ℓ. The density, cross-sectional area, and length of the rod are denoted by ρ, A, and ℓ, and, in accordance with (14), the Young's modulus and Kelvin–Voigt damping parameter are denoted by Y and C.

Force balancing yields the relation

$$\rho A \frac{\partial^2 u}{\partial t^2} = \frac{\partial \mathcal{N}}{\partial x}, \tag{16}$$

where the resultant $N = \int_A \sigma dA$ is given by

$$N = YA\frac{\partial u}{\partial x} + CA\frac{\partial^2 u}{\partial x \partial t} - a_1[P(E) - P_R] - a_2[P(E) - P_R]^2,$$

once the linear relation $\varepsilon = \frac{\partial u}{\partial x}$ is employed for the strains in (14). The nonlinear and hysteretic map between input fields E and the polarization P is specified by (11). The fixed-end condition yields $u(t,0) = 0$, and balancing forces at $x = \ell$ yields the energy dissipating end condition

$$N(t,\ell) = -k_\ell u(t,\ell) - c_\ell \frac{\partial u}{\partial t}(t,\ell) - m_\ell \frac{\partial^2 u}{\partial t^2}(t,\ell).$$

Finally, initial conditions are taken to be $u(0,x) = u_0(x)$ and $\frac{\partial u}{\partial t}(0,x) = u_1(x)$. This provides a strong formulation of the stacked actuator model.

To define a weak or variational form of the model which is appropriate for well-posedness analysis, approximation, or control design, states $z = (u(\cdot), u(\ell))$ are con-

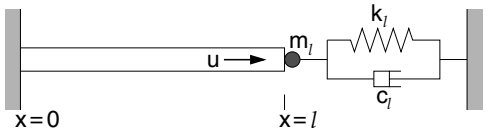


FIG. 7. *Rod of length ℓ and cross-sectional area A with a fixed end at $x = 0$ and energy dissipating boundary conditions at $x = \ell$.*

sidered in the state space $X = L^2(0,\ell) \times \mathbb{R}$ with the inner product

$$\langle \Phi_1, \Phi_2 \rangle_X = \int_0^\ell \rho A \phi_1 \phi_2 dx + m_\ell \varphi_1 \varphi_2, \tag{17}$$

where $\Phi_1 = (\phi_1, \varphi_1), \Phi_2 = (\phi_2, \varphi_2)$ with $\varphi_1 = \phi_1(\ell)$, $\varphi_2 = \phi_1(\ell)$. The space of test functions is taken to be

$$V = \left\{ \Phi = (\phi, \varphi) \in X \,|\, \phi \in H^1(0,\ell), \phi(0) = 0, \phi(\ell) = \varphi \right\}$$

with the inner product

$$\langle \Phi_1, \Phi_2 \rangle_V = \int_0^\ell Y A \phi_1' \phi_2' dx + k_\ell \varphi_1 \varphi_2. \tag{18}$$

Multiplication by $\phi \in H_0^1(0,\ell) = \{\phi \in H^1(0,\ell) \,|\, \phi(0) = 0\}$ and integration by parts in space yields the weak model formulation

$$\begin{aligned} &\int_0^\ell \rho A \frac{\partial^2 u}{\partial t^2} \phi dx + \int_0^\ell \left[Y A \frac{\partial u}{\partial x} + C A \frac{\partial^2 u}{\partial x \partial t} \right] \frac{d\phi}{dx} dx \\ &\qquad = \int_0^\ell f \phi dx + A \left[a_1 (P - P_R) + a_2 (P - P_R)^2 \right] \int_0^\ell \frac{d\phi}{dx} dx \\ &\qquad\quad - \left[k_\ell u(t,\ell) + c_\ell \frac{\partial u}{\partial t}(t,\ell) + m_\ell \frac{\partial^2 u}{\partial t^2}(t,\ell) \right] \phi(\ell), \end{aligned} \tag{19}$$

which must be satisfied for all $\phi \in V$.

3.1.2. Lumped rod model. The assumption that fields and stresses are uniform along the rod length motivates the conclusion that strains (relative displacements) also exhibit negligible x-dependence. Since the position of the sample is dictated by the position of the rod tip at $x = \ell$, this motivates the development of a lumped model which quantifies $u_\ell(t) = u(t,\ell)$.

From the assumption of uniform strains along the rod length, we take

$$\varepsilon(t) = \frac{u_\ell(t)}{\ell}$$

in (14). Balancing the forces σA for the rod with those of the restoring mechanism yields the lumped model

$$\begin{aligned} \rho A \ell \frac{d^2 u_\ell}{dt^2}(t) + \frac{CA}{\ell} \frac{du_\ell}{dt}(t) + \frac{YA}{\ell} u_\ell(t) &= -m_\ell \frac{d^2 u_\ell}{dt^2}(t) - c_\ell \frac{du_\ell}{dt}(t) - k u_\ell(t) \\ &\quad + A a_1 [P(E(t)) - P_R] + A a_2 [P(E(t)) - P_R]^2 \end{aligned}$$

or, equivalently,

$$m \frac{d^2 u_\ell}{dt^2}(t) + c \frac{du_\ell}{dt}(t) + k u_\ell(t) = \widetilde{a}_1 [P(E(t)) - P_R] + \widetilde{a}_2 [P(E(t)) - P_R]^2, \tag{20}$$

where

$$m = \rho A \ell + m_\ell, \quad c = \frac{CA}{\ell} + c_\ell, \quad k = \frac{YA}{\ell} + k_\ell, \quad \widetilde{a}_1 = A a_1, \quad \widetilde{a}_2 = A a_2, \tag{21}$$

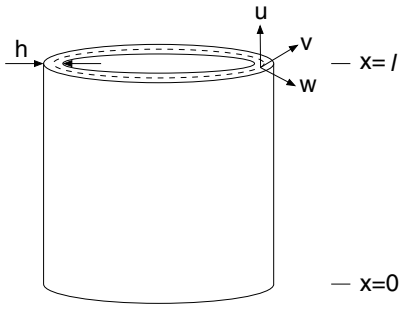


FIG. 8. *Orientation of the shell geometry used when quantifying the longitudinal, circumferential, and transverse displacements u, v, and w.*

and the initial conditions are $u_\ell(0) = u_0$ and $\frac{du_\ell}{dt}(0) = u_1$. The polarization P is specified by the model (11) or discretized model (13).

The model can also be written as the first-order system

$$\begin{aligned} \dot{\vec{u}}_\ell(t) &= A\vec{u}_\ell(t) + \vec{\mathcal{P}}(E(t)), \\ \vec{u}_\ell(0) &= \vec{u}_0, \end{aligned} \tag{22}$$

where $\vec{u}_\ell(t) = [u_\ell(t), \dot{u}_\ell(t)]^T$, $\vec{u}_\ell(0) = [u_0, u_1]^T$ and

$$A = \begin{bmatrix} 0 & 1 \\ -k/m & -c/m \end{bmatrix},$$

$$\vec{\mathcal{P}}(E(t)) = \frac{1}{m}\left[\widetilde{a}_1(P(E(t)) - P_R) + \widetilde{a}_2(P(E(t)) - P_R)^2\right] \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

3.2. Cylindrical shell model. To quantify the dynamics of the cylindrical stage depicted in Figure 2(b), we construct a linear shell model with nonlinear inputs quantified by the 2-D constitutive relation (15). We focus on the actuator employed for transverse displacements since real-time control of this component is required to maintain constant forces between the sample and microcantilever. The mass of the shell employed for horizontal translation is combined with the mass of the sample to provide an inertial force acting on the free end of the vertical actuator.

For modeling purposes, we assume that the shell has length ℓ, thickness h, and radius R. The axial direction is specified along the x-axis, and the longitudinal, circumferential, and transverse displacements are respectively denoted by u, v, and w, as depicted in Figure 8. The density is designated by ρ, and the region occupied by the reference or middle surface of the shell is specified by $\Gamma_0 = [0, \ell] \times [0, 2\pi]$. In accordance with the constitutive relations (15), Y, C, and ν denote the Young's modulus, Kelvin–Voigt damping coefficient, and Poisson ratio for the material. We point out that $\varepsilon_x, \varepsilon_\theta$, and $\varepsilon_{x\theta}$ in (15) denote strains at points throughout the shell thickness, whereas 2-D shell models are formulated in terms of strains e_x, e_θ, and $e_{x\theta}$ in the reference surface of the shell. The relationship between the two is established through the assumption that displacements are linear through the shell thickness, which comprises one of the fundamental tenets of linear shell theory [3, 30].

We consider the case in which the bottom edge of the shell ($x = 0$) is clamped and the opposite end ($x = \ell$) is acted upon only by the inertial force associated with the combined mass m of the x-y actuator and the sample.

As detailed in [3, 30], force and moment balancing yield the Donnell–Mushtari shell equations

$$
\begin{aligned}
&R\rho h\frac{\partial^2 u}{\partial t^2} - R\frac{\partial N_x}{\partial x} - \frac{\partial N_{x\theta}}{\partial \theta} = 0,\\
&R\rho h\frac{\partial^2 v}{\partial t^2} - \frac{\partial N_\theta}{\partial \theta} - R\frac{\partial N_{x\theta}}{\partial x} = 0,\\
&R\rho h\frac{\partial^2 w}{\partial t^2} - R\frac{\partial^2 M_x}{\partial x^2} - \frac{1}{R}\frac{\partial^2 M_\theta}{\partial \theta^2} - 2\frac{M_{x\theta}}{\partial x\partial \theta} + N_\theta = 0,
\end{aligned}
\tag{23}
$$

where the force and moment resultants are

$$
\begin{aligned}
N_x &= \frac{Yh}{1-\nu^2}(e_x + \nu e_\theta) + \frac{Ch}{1-\nu^2}(\dot{e}_x + \nu \dot{e}_\theta) - \frac{h}{1-\nu}[a_1(P-P_R) + a_2(P-P_R)^2],\\
N_\theta &= \frac{Yh}{1-\nu^2}(e_\theta + \nu e_x) + \frac{Ch}{1-\nu^2}(\dot{e}_\theta + \nu \dot{e}_x) - \frac{h}{1-\nu}[a_1(P-P_R) + a_2(P-P_R)^2],\\
N_{x\theta} &= \frac{Yh}{2(1+\nu)}e_{x\theta} + \frac{Ch}{2(1+\nu)}\dot{e}_{x\theta}
\end{aligned}
\tag{24}
$$

and

$$
\begin{aligned}
M_x &= \frac{Yh^3}{12(1-\nu^2)}(\kappa_x + \nu\kappa_\theta) + \frac{Ch^3}{12(1-\nu^2)}(\dot{\kappa}_x + \nu\dot{\kappa}_\theta),\\
M_\theta &= \frac{Yh^3}{12(1-\nu^2)}(\kappa_\theta + \nu\kappa_x) + \frac{Ch^3}{12(1-\nu^2)}(\dot{\kappa}_\theta + \nu\dot{\kappa}_x),\\
M_{x\theta} &= \frac{Yh^3}{24(1+\nu)}\kappa_{x\theta} + \frac{Ch^3}{24(1+\nu)}\dot{\kappa}_{x\theta}.
\end{aligned}
\tag{25}
$$

The midsurface strains and changes in curvature are

$$
\begin{aligned}
&e_x = \frac{\partial u}{\partial x}, \quad e_\theta = \frac{1}{R}\frac{\partial v}{\partial \theta} + \frac{w}{R}, \quad e_{x\theta} = \frac{\partial v}{\partial x} + \frac{1}{R}\frac{\partial u}{\partial \theta},\\
&\kappa_x = -\frac{\partial^2 w}{\partial x^2}, \quad \kappa_\theta = -\frac{1}{R^2}\frac{\partial^2 w}{\partial \theta^2}, \quad \kappa_{x\theta} = -\frac{2}{R}\frac{\partial^2 w}{\partial x\partial \theta}.
\end{aligned}
\tag{26}
$$

The boundary conditions for the fixed-end at $x = 0$ are taken to be

$$u = v = w = \frac{\partial w}{\partial x} = 0,$$

whereas the conditions

$$
\begin{aligned}
&N_x = -m\frac{\partial^2 u}{\partial t^2}, \qquad && N_{x\theta} + \frac{M_{x\theta}}{R} = 0,\\
&Q_x + \frac{1}{R}\frac{\partial M_{x\theta}}{\partial \theta} = 0, \qquad && M_x = 0,
\end{aligned}
$$

are employed at $x = \ell$. The first resultant condition incorporates the inertial force due to the mass m of the PZT actuator employed for x-y translation along with the mass of the sample.

To reduce smoothness requirements for approximation and eliminate the Dirac behavior of external inputs at $x = \ell$, we also consider a weak formulation of the model. The state is taken to be $z = (u(\cdot,\cdot), v(\cdot,\cdot), w(\cdot,\cdot), u(\ell,\cdot))$ in the state space

$$X = L^2(\Omega) \times L^2(\Omega) \times L^2(\Omega) \times L^2(0, 2\pi),$$

where

$$\Omega = [0, \ell] \times [0, 2\pi]$$

denotes the shell region. The space of test functions is specified as

$$V = \left\{ \Phi = (\phi_1, \phi_2, \phi_3, \eta) \in X \,|\, \phi_1 \in H_0^1(\Omega), \phi_2 \in H_0^1(\Omega), \phi_3 \in H_0^2(\Omega) \right\},$$

where $\eta(\theta) = \phi_1(\ell, \theta)$ and

$$\begin{aligned} H_0^1(\Omega) &= \left\{ \phi \in H^1(\Omega) \,|\, \phi(0,\theta) = 0 \right\}, \\ H_0^2(\Omega) &= \left\{ \phi \in H^2(\Omega) \,|\, \phi(0,\theta) = \phi'(0,\theta) = 0 \right\}. \end{aligned} \tag{27}$$

Through either variation principles—e.g., see [3]—or integration by parts, one obtains the weak formulation of the thin shell model,

$$\begin{aligned} &\int_\Omega \left\{ R\rho h \frac{\partial^2 u}{\partial t^2}\phi_1 + RN_x \frac{\partial \phi_1}{\partial x} + N_{x\theta}\frac{\partial \phi_1}{\partial \theta} \right\} d\omega = 0, \\ &\int_\Omega \left\{ R\rho h \frac{\partial^2 v}{\partial t^2}\phi_2 + N_\theta \frac{\partial \phi_2}{\partial \theta} + RN_{x\theta}\frac{\partial \phi_2}{\partial x} \right\} d\omega = 0, \\ &\int_\Omega \left\{ R\rho h \frac{\partial^2 w}{\partial t^2}\phi_3 - RM_x \frac{\partial^2 \phi_3}{\partial x^2} - 2M_{x\theta}\frac{\partial^2 \phi_3}{\partial x \partial \theta} - \frac{1}{R} M_\theta \frac{\partial^2 \phi_3}{\partial \theta^2} + N_\theta \phi_3 \right\} d\omega = 0, \end{aligned} \tag{28}$$

which must be satisfied for all $\Phi \in V$. The resultants are given by (24) and (25) with midsurface strains and changes in curvature designated in (26).

Remark 5. It is noted that the d_{31} poling, used to generate vertical motion in the stage, produces no polarization contributions to the moments. However, transverse displacements w in the shell model are generated by the N_θ resultant in the w relation, and hence all three components of the displacement are coupled.

3.3. Frequency-dependent dynamics. One of the requirements of the nanopositioner models is the capability to characterize the frequency-dependent behavior shown in Figure 2. This behavior is due to a combination of dielectric losses, thermal relaxation processes, and elastic and damping properties, and it is incorporated in the framework in two places. The dielectric losses and relaxation behavior are incorporated through the balance of the Gibbs and relative thermal energies via the Boltzmann relation (7) and subsequent average polarization relations (9) and likelihood expressions (10). Hence this component of the polarization model incorporates the property that dipole dynamics can lag behind field dynamics as frequencies are increased. Dynamics associated with inertial, elastic, and internal damping properties of the actuators are incorporated through the force balances (16) and (23) and resultant definitions. In combination, this provides the framework with significant flexibility regarding a range of dynamics operating regimes.

4. Model well-posedness.

4.1. Rod model. To provide a framework which facilitates the establishment of criteria that guarantee the existence of a unique solution to the distributed rod model with nonlinear inputs, we consider a Hilbert space formulation of the weak model formulation (19) with the state and test function spaces

$$X = L^2(0,\ell) \times \mathbb{R},$$
$$V = \left\{\Phi = (\phi,\varphi) \in X \,|\, \phi \in H^1(0,\ell), \phi(0) = 0, \phi(\ell) = \varphi\right\}$$

and inner products

$$\begin{aligned} \langle \Phi_1, \Phi_2 \rangle_X &= \int_0^\ell \rho A \phi_1 \phi_2 dx + m_\ell \varphi_1 \varphi_2, \\ \langle \Phi_1, \Phi_2 \rangle_V &= \int_0^\ell Y A \phi_1' \phi_2' dx + k_\ell \varphi_1 \varphi_2, \end{aligned} \tag{29}$$

where $\Phi_1 = (\phi_1, \varphi_1), \Phi_2 = (\phi_2, \varphi_2)$ with $\varphi_1 = \phi_1(\ell)$, $\varphi_2 = \phi_1(\ell)$.

It is observed that V is densely and continuously embedded in X with $|\Phi|_X \leq c|\Phi|_V$; this is expressed by $V \hookrightarrow X$. Moreover, when one defines conjugate dual spaces X^* and V^*—e.g., V^* denotes the linear space of all conjugate linear continuous functionals on V—two observations are important: (i) X^* can be identified with X through the Riesz map and (ii) $X^* \hookrightarrow V^*$. Hence the two spaces comprise what is termed a Gelfand triple, $V \hookrightarrow X \cong X^* \hookrightarrow V^*$ with pivot space X and duality pairing (duality product) $\langle \cdot, \cdot \rangle_{V^*,V}$. The latter is defined as the extension by continuity of the inner product $\langle \cdot, \cdot \rangle_X$ from $V \times X$ to $V^* \times X$. Hence elements $v^* \in V^*$ have the representation $v^*(v) = \langle v^*, v \rangle_{V^*,V}$.

We now define the stiffness and damping sesquilinear forms $\sigma_i : V \times V \to \mathbb{C}$, $i = 1, 2$, by

$$\begin{aligned} \sigma_1(\Phi_1, \Phi_2) &= \langle \Phi_1, \Phi_2 \rangle_V, \\ \sigma_2(\Phi_1, \Phi_2) &= \int_0^\ell C A \phi_1' \phi_2' dx + c_\ell \varphi_1 \varphi_2. \end{aligned} \tag{30}$$

It can be directly verified that the stiffness form satisfies

$$\begin{aligned} &\text{(H1)} \quad |\sigma_1(\Phi_1, \Phi_2)| \leq c_1 |\Phi_1|_V |\Phi_2|_V \text{ for some } c_1 \in \mathbb{R} \qquad \text{(bounded)}, \\ &\text{(H2)} \quad \operatorname{Re} \sigma_1(\Phi_1, \Phi_1) \geq c_2 |\Phi_1|_V^2 \text{ for some } c_2 > 0 \qquad (V\text{-elliptic}), \\ &\text{(H3)} \quad \sigma_1(\Phi_1, \Phi_2) = \overline{\sigma_1(\Phi_2, \Phi_1)} \qquad \text{(symmetric)}, \end{aligned}$$

for all $\psi, \phi \in V$. Moreover, the damping term σ_2 satisfies

$$\begin{aligned} &\text{(H4)} \quad |\sigma_2(\Phi_1, \Phi_2)| \leq c_3 |\Phi_1|_V |\Phi_2|_V \text{ for some } c_3 \in \mathbb{R} \qquad \text{(bounded)}, \\ &\text{(H5)} \quad \operatorname{Re} \sigma_2(\Phi_1, \Phi_1) \geq c_4 |\Phi_1|_V^2 \text{ for some } c_4 > 0 \qquad (V\text{-elliptic}). \end{aligned} \tag{31}$$

The input space is taken to be the Hilbert space $U = \mathbb{R}$, and the input operator $B : U \to V^*$ is defined by

$$\langle [B(E)](t), \Phi \rangle_{V^*,V} = \left[a_1(P(E(t)) - P_R) + a_2(P(E(t)) - P_R)^2\right] \int_0^\ell \phi' \, dx \tag{32}$$

for $\Phi = (\phi, \varphi)$ with $\varphi = \phi(\ell)$. It is observed that B can be expressed as

$$[B(E)](t) = [b(E)](t) \cdot g, \qquad g \in V^*, \tag{33}$$

where

$$\begin{aligned} &[b(E)](t) = (P(E(t)) - P_R) + a_2(P(E(t)) - P_R)^2, \\ &g(\Phi) = \int_0^\ell \phi' \, dx. \end{aligned} \tag{34}$$

The model (19) can then be written in the abstract weak formulation

$$\begin{aligned} &\langle \ddot{u}(t), \Phi \rangle_{V^*,V} + \sigma_2(\dot{u}(t), \Phi) + \sigma_1(u(t), \Phi) = \langle [B(E)](t), \Phi \rangle_{V^*,V}, \\ &u(0) = u_0, \quad \dot{u}(0) = u_1, \end{aligned} \tag{35}$$

for all $\Phi \in V$.

Alternatively, one can define the operators $A_i \in \mathcal{L}(V, V^*), i = 1, 2$, by

$$\langle A_i \Phi_1, \Phi_2 \rangle_{V^*,V} = \sigma_i(\Phi_1, \Phi_2) \tag{36}$$

and formulate the model in operator form as

$$\begin{aligned} &\ddot{u}(t) + A_2 \dot{u}(t) + A_1 u(t) = [B(E)](t), \\ &u(0) = u_0, \quad \dot{u}(0) = u_1, \end{aligned} \tag{37}$$

in the dual space V^*. This formulation illustrates the analogy between the infinite-dimensional, strongly damped elastic model and the familiar finite-dimensional relations (22).

4.1.1. Model well-posedness. As a prelude to establishing the well-posedness of the beam model with hysteretic E-P relations, we provide a lemma which quantifies the smoothness of the input operator.

LEMMA 1. *Consider field inputs $E \in C[0,T]$. The input operator B defined by* (32) *then satisfies*

$$B(E) \in L^2(0, T; V^*). \tag{38}$$

Proof. In the appendix, we establish that for continuous input fields E the polarization satisfies $P \in C[0,T]$, which implies that b defined by (34) satisfies $b(\cdot) : C[0,T] \to C[0,T]$. Hence the norm

$$\|[B(E)](t)\|_{V^*} = \sup_{v \in V} \frac{|[b(E)](t) \cdot g(v)|}{\|v\|_V}$$

exists for each $t \in [0,T]$. Since $\|[B(E)](t)\|_{V^*} = |[b(E)](t)| \cdot \|g\|_{V^*}$, it follows that

$$\|B(E)\|^2_{L^2(0,T;V^*)} \leq \max_{t \in [0,T]} \left\{ |[b(E)](t)|^2 \right\} \cdot T \cdot \|g\|^2_{V^*},$$

which implies that

$$B(E) \in L^2(0, T; V^*). \qquad \square$$

The well-posedness of the model is established by the following theorem, whose proof follows directly from Theorem 4.1 of [3] or Theorem 2.1 and Remark 2.1 of [2].

THEOREM 2. *Let σ_1 and σ_2 be given by* (30)*, and consider continuous field inputs $E \in C[0,T]$. There then exists a unique solution w to* (35)*, or equivalently* (37)*, which satisfies*

$$u \in C(0,T;V),$$
$$\dot{u} \in C(0,T;X).$$

4.2. Shell model. Similar well-posedness results can be obtained for the shell model (28) through consideration of an analogous Hilbert space formulation of the model. Details regarding the construction of appropriate inner product spaces, sesquilinear forms, and operators can be found in [30, 35].

5. Numerical approximation techniques. To implement the distributed models for either the rectangular stacked actuator or the cylindrical actuator, it is necessary to develop appropriate approximation techniques to discretize the modeling PDE. To accomplish this, we consider general Galerkin methods in which basis functions are comprised of spline or spline-Fourier tensor products. The resulting methods can accommodate a variety of boundary conditions, are sufficiently accurate to resolve fine-scale dynamics, and can be employed for constructing reduced-order proper orthogonal decomposition approximates for real-time implementation.

5.1. Stacked actuator model. To approximate the weak form of the stacked actuator model (19), we employ a finite element discretization and a finite difference discretization in time. The semidiscrete system resulting from the finite element approximation is appropriate for finite-dimensional continuous time control design, whereas the fully discrete system is amenable to simulations and control implementation.

To obtain a semidiscrete system, we consider a uniform partition of $[0,\ell]$ with points $x_j = jh$, $j = 0,1,\dots,N$, with step size $h = \ell/N$, where N denotes the number of subintervals. The spatial basis $\{\phi_j\}_{j=1}^N$ is then comprised of linear splines

$$\phi_j(x) = \frac{1}{h}\begin{cases} (x - x_{j-1}), & x_{j-1} \le x < x_j, \\ (x_{j+1} - x), & x_j \le x \le x_{j+1}, \\ 0, \quad \text{otherwise}, \end{cases} \qquad i = 1,\dots,N-1,$$

$$\phi_N(x) = \frac{1}{h}\begin{cases} (x - x_{N-1}), & x_{N-1} \le x \le x_N, \\ 0, \quad \text{otherwise} \end{cases}$$

(see [25] for details regarding the convergence analysis for the method). The solution $u(t,x)$ to (19) is subsequently approximated by the expansion

$$u^N(t,x) = \sum_{j=1}^{N} u_j(t)\phi_j(x)\,.$$

Through construction, the approximate solution satisfies the essential boundary condition $u^N(t,0) = 0$ and can attain arbitrary displacements at $x = \ell$.

The projection of the problem (19) onto the finite-dimensional subspace V^N yields the semidiscrete system

$$\begin{aligned} &\dot{\mathbf{z}}(t) = \mathbb{A}\,\mathbf{z}(t) + A\left[a_1(P(t)-P_R) + a_2(P(t)-P_R)^2\right]\mathbf{B}, \\ &\mathbf{z}(0) = \mathbf{z}_0, \end{aligned} \tag{39}$$

where $\mathbf{z}(t) = [u_1(t), \ldots, u_N(t), \dot{u}_1(t), \ldots, \dot{u}_N(t)]^T$ and

$$\mathbb{A} = \begin{bmatrix} 0 & \mathbb{I} \\ -\mathbb{M}^{-1}\mathbb{K} & -\mathbb{M}^{-1}\mathbb{Q} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 \\ \mathbb{M}^{-1}\mathbf{b} \end{bmatrix}. \tag{40}$$

The mass, stiffness, and damping matrices have the components

$$[\mathbb{M}]_{ij} = \begin{cases} \displaystyle\int_0^\ell \rho A \phi_i \phi_j \, dx, & i \neq N \text{ or } j \neq N, \\ \displaystyle\int_0^\ell \rho A \phi_i \phi_j \, dx + m_\ell, & i = N \text{ and } j = N, \end{cases}$$

$$[\mathbb{K}]_{ij} = \begin{cases} \displaystyle\int_0^\ell Y A \phi_i' \phi_j' \, dx, & i \neq N \text{ or } j \neq N, \\ \displaystyle\int_0^\ell Y A \phi_i' \phi_j' \, dx + k_\ell, & i = N \text{ and } j = N, \end{cases}$$

and

$$[\mathbb{Q}]_{ij} = \begin{cases} \displaystyle\int_0^\ell c A \phi_i' \phi_j' \, dx, & i \neq N \text{ or } j \neq N, \\ \displaystyle\int_0^\ell c A \phi_i' \phi_j' \, dx + c_\ell, & i = N \text{ and } j = N\,, \end{cases}$$

and the force vector is defined by

$$[\mathbf{b}\,]_i = \int_0^\ell \phi_i' dx.$$

The system (39) can be employed for finite-dimensional control design. For subsequent implementation, we consider a temporal discretization of (39) using a modified trapezoid rule. For temporal stepsizes Δt, this yields the difference equation

$$\mathbf{z}_{k+1} = \mathbb{W}\mathbf{z}_k + \frac{1}{2}\left[a_1\widetilde{P}(t_k) + a_1\widetilde{P}(t_{k+1}) + a_2\widetilde{P}^2(t_k) + a_2\widetilde{P}^2(t_{k+1})\right]\mathbb{V}\mathbf{B}, \tag{41}$$

where $\widetilde{P} = P - P_R$, $t_j = j\Delta t$, $\mathbf{z}_j$ approximates $\mathbf{z}(t_j)$, and

$$\mathbb{W} = \left(\mathbb{I} - \frac{\Delta t}{2}\mathbb{A}\right)^{-1}\left(\mathbb{I} + \frac{\Delta t}{2}\mathbb{A}\right), \qquad \mathbb{V} = \Delta t\left(\mathbb{I} - \frac{\Delta t}{2}\mathbb{A}\right)^{-1}.$$

This yields an A-stable method requiring moderate storage and providing moderate accuracy.

5.2. Cylindrical actuator model. Due to the inherent coupling between longitudinal, circumferential, and transverse displacements in combination with the 2-D support of the middle surface, the numerical approximation of the model for the cylindrical actuator is significantly more complicated than the approximation of the stacked actuator model. Among the issues which must be addressed when constructing finite element or general Galerkin methods for the shell is the choice of elements which avoid shear and membrane locking and the maintenance of boundary conditions. We summarize here a spline-based Galerkin method developed in [9] for thin shells and direct the reader to that source for details regarding the construction of constituent matrices and convergence properties of the method. Details regarding the use of this approximation method for LQR (linear quadratic regulator) control of shells utilizing PZT actuators can be found in [10].

The bases for the u, v, and w displacements are respectively taken to be

$$\Phi_{u_k}(\theta, x) = e^{im\theta}\phi_{u_n}(x), \quad \Phi_{v_k}(\theta, x) = e^{im\theta}\phi_{v_n}(x), \quad \Phi_{w_k}(\theta, x) = e^{im\theta}\phi_{w_n}(x),$$

where ϕ_{u_n}, ϕ_{v_n}, and ϕ_{w_n} are cubic B-splines modified to satisfy the boundary conditions (e.g., see p. 79 of [25]). The approximating subspaces are

$$V_u^N = \operatorname{span}\{\Phi_{u_k}\}_{k=1}^{N_u}, \quad V_v^N = \operatorname{span}\{\Phi_{v_k}\}_{k=1}^{N_v}, \quad V_w^N = \operatorname{span}\{\Phi_{w_k}\}_{k=1}^{N_w},$$

and the approximate displacements are represented by the expansions

$$\begin{aligned} u^N(t,\theta,x) &= \sum_{k=1}^{N_u} u_k(t)\Phi_{u_k}(\theta,x), \\ v^N(t,\theta,x) &= \sum_{k=1}^{N_v} v_k(t)\Phi_{v_k}(\theta,x), \\ w^N(t,\theta,x) &= \sum_{k=1}^{N_w} w_k(t)\Phi_{w_k}(\theta,x). \end{aligned} \tag{42}$$

The restriction of the problem (28) to the approximating subspaces and construction of the forcing vectors subsequently yields the matrix system

$$\begin{aligned} &\dot{\mathbf{z}}^N(t) = \mathbb{A}\,\mathbf{z}(t) + \left[a_1(P(t) - P_R) + a_2(P(t) - P_R)^2\right]\mathbf{B}, \\ &\mathbf{z}(0) = \mathbf{z}_0, \end{aligned} \tag{43}$$

where $\mathbf{z} = [\boldsymbol{\vartheta}(t), \dot{\boldsymbol{\vartheta}}(t)]^T$, with $\boldsymbol{\vartheta}(t) = [\mathbf{u}(t), \mathbf{v}(t), \mathbf{w}(t)]^T$, and

$$\mathbb{A} = \begin{bmatrix} 0 & \mathbb{I} \\ -\mathbb{M}^{-1}\mathbb{K} & -\mathbb{M}^{-1}\mathbb{Q} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 \\ \mathbb{M}^{-1}\mathbf{b} \end{bmatrix}.$$

The reader is referred to [9, 30] for details concerning the construction of the mass, stiffness, and damping matrices $\mathbb{M}, \mathbb{K}$, and $\mathbb{Q}$.

6. Model validation.

6.1. Characterization of the stacked actuator. We consider the capability of the modeling framework for characterizing the dynamics of the stacked actuator depicted in Figure 2(a). The PZT actuator had a length of $\ell = 2 \times 10^{-2}$ m and a

TABLE 1
Parameters employed in the distributed (PDE) model (19) *and lumped (ODE) model* (20) *for the stacked actuator.*

Distributed model					
Parameter	ρ	Y	C	m_ℓ	
value	7600	7×10^{10}	5×10^6	4.015	
Parameter	k_ℓ	c_ℓ	a_1	a_2	
value	8.49×10^{-5}	440	1.54×10^{11}	0	
Lumped model					
Parameter	m	k	c	$\widetilde{a}_1$	$\widetilde{a}_2$
value	4.21	8.75×10^7	1.52×10^5	8.75×10^7	0

square cross-sectional face of width $w = 5 \times 10^{-3}$ m, so that the cross-sectional area is $A = 2.5 \times 10^{-5}$ m^2. As illustrated in Figure 7, one end of the actuator was considered fixed, whereas the other encountered elastic, damping, and inertial effects due to the attached components of the stage mechanism.

To validate and illustrate properties of the models, we consider three regimes: (i) end displacements quantified by the lumped model (20) with the thermally inactive kernel (6) employed in the polarization model (14), (ii) displacements characterized by the lumped model with the thermally active polarization kernel (8), and (iii) end displacements quantified by the discretization (41) of the distributed model (19). It is illustrated that whereas the latter choice incorporates the distributed rod nature of the device, the fact that fields and stresses are uniform along the rod length implies that relative displacements are also uniform. A comparison of the ODE and PDE model predictions at the rod tip ($x = \ell$) illustrates that, as a result, the ODE provides a highly accurate characterization with significantly less computation cost. Hence the ODE model is advantageous for real-time experimental implementation.

The construction of the models requires the estimation of elastic, damping, and electromechanical parameters in addition to identification of the densities ν_1 and ν_2. The densities were estimated through least squares fits to the data using the techniques detailed in [30, 32]. The manufacturer specifications $\rho = 7600$ kg/m^3 and $Y = 7 \times 10^{10}$ N/m^2 were employed for the density and Young's modulus, and remaining parameters were estimated through a least squares fit to the data. The resulting values are summarized in Table 1. The relation between the rod and spring parameters is provided by (21).

6.1.1. Lumped model—No thermal activation in polarization relation. We consider first the characterization of the biased minor loop data shown in Figure 3 and frequency-dependent data from Figure 4 using the lumped model (20) with the thermally inactive kernel (6) employed in the polarization model (14). It should be noted that the stage was disassembled between the quasi-static biased minor loop experiments and the frequency-dependent experiments, which necessitated the re-identification of densities for the two cases.

In the first set of experiments, displacement data measured with an LVDT (linear variable differential transformer) was collected at a sample rate of 0.1 Hz and four input field levels to generate a set of biased and nested transducer responses ranging from nearly linear to hysteretic and nonlinear, as shown in Figures 3 and 9. The densities ν_1 and ν_2 and parameters summarized in Table 1 were obtained through a least squares fit to the full data set comprised of four loops. The resulting model accurately quantifies both the nest behavior and the hysteresis measured at increasing

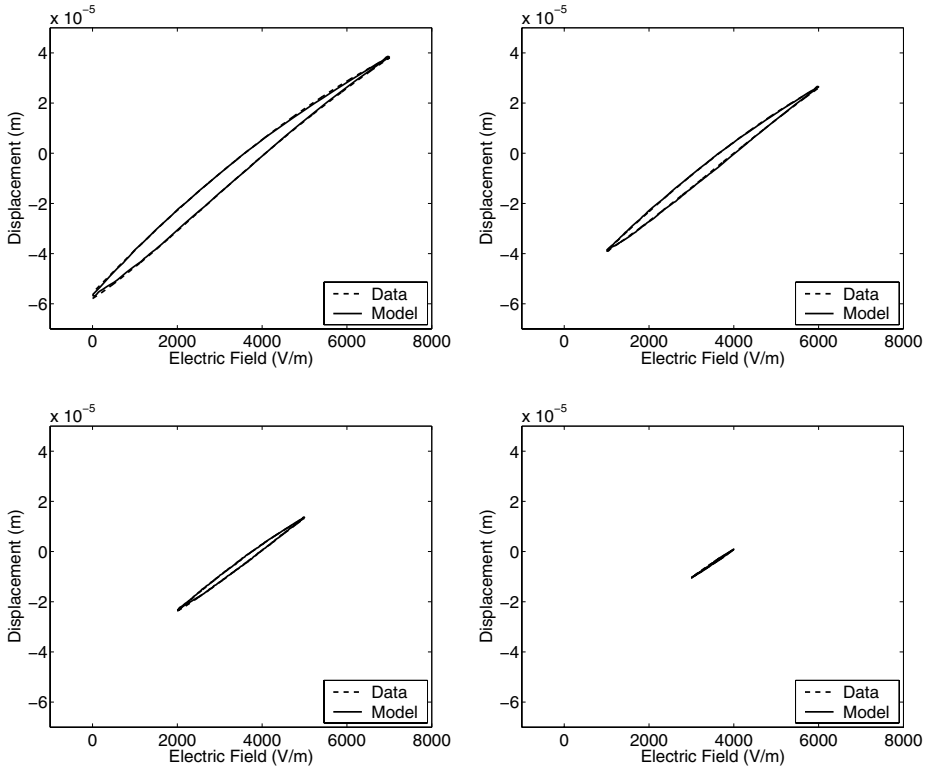


FIG. 9. *Characterization of AFM field-displacement behavior at* 0.1 *Hz using the ODE model* (20) *with the thermally inactive kernel* (6).

input levels.

In a second set of experiments, data was collected at frequencies ranging from 0.279 Hz to 27.9 Hz, yielding the behavior shown in Figure 4. These experiments took longer (approximately 30 minutes), which led to slight heating and accompanying changes in the material constitutive behavior. To accommodate these operating conditions, the data from four frequencies was used to reidentify parameters in the polarization model, thus yielding the fits shown in Figure 10. It is observed that the model characterizes the augmented hysteresis arising at higher frequencies but slightly overpredicts the increase in displacement following field reversal that is due primarily to inertial effects.

6.1.2. Lumped model—Thermal activation in polarization relation. We next employ the thermally active kernel (8) in the polarization model to incorporate relaxation effects. Parameters in the polarization model were again identified through a least squares fit to the four frequency data sets, thus yielding the model fit shown in Figure 11. It is observed that use of this more general kernel provides additional accuracy at higher frequencies. Whereas this improves characterization capabilities, the added accuracy comes at the cost of decreased efficiency, and the criteria of accuracy versus efficiency must be balanced when employing the model for real-time control design, as discussed in [16]. We note that use of the thermally active kernel (8) is required when characterizing the creep measured when one actuator is held fixed while a sweep is performed with the other.

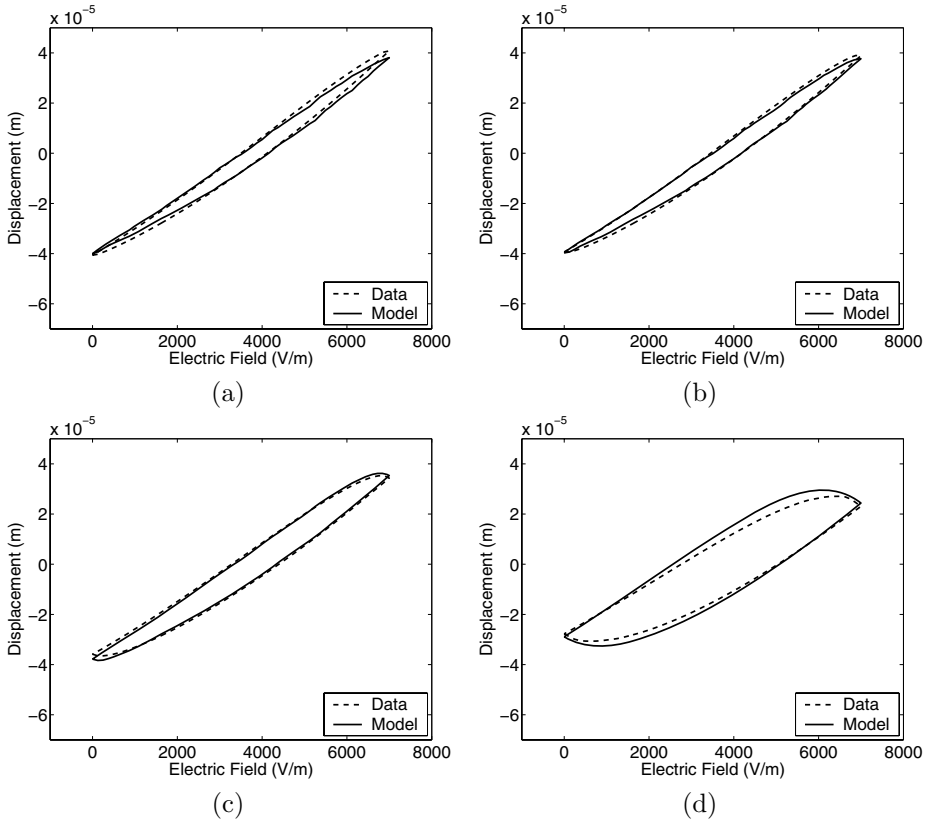


FIG. 10. *Characterization of AFM field-displacement behavior using the ODE model* (20) *with the* thermally inactive *kernel* (6)*, with sample rates of* (a) 0.279 *Hz,* (b) 1.12 *Hz,* (c) 5.58 *Hz, and* (d) 27.9 *Hz.*

6.1.3. Lumped model versus distributed model. It has been observed that whereas quantification of the physics of the stacked actuator leads to the rod model (19), the fact that stresses and fields are uniform along the rod length implies that relative displacements will also be uniform. This motivates consideration of the lumped model (20), which yielded the fits shown in Figures 10 and 11.

To illustrate the validity of this assumption, the difference between the displacement $u(t, \ell)$, given by the discretization (41) of (19) with $N = 16$ basis functions, and the displacement $u_\ell(t)$, resulting from (20), is plotted in Figure 12. We emphasize that when constructing the PDE model we employed the parameter values summarized in Table 1, which are *consistent* with the spring parameters due to the relation (21). The maximal difference of 5×10^{-10} is five orders of magnitude less than the micron-level displacements being characterized, thus verifying the validity of the ODE model in this regime. The accuracy of the ODE model has important ramifications for control design since the discretized ODE model is significantly more efficient to implement than the discretized PDE model.

6.2. Characterization of the shell actuator. We discuss here the performance of the cylindrical shell model detailed in section 3.2, when discretized using the Galerkin techniques summarized in section 5.2, for characterizing the longitudinal

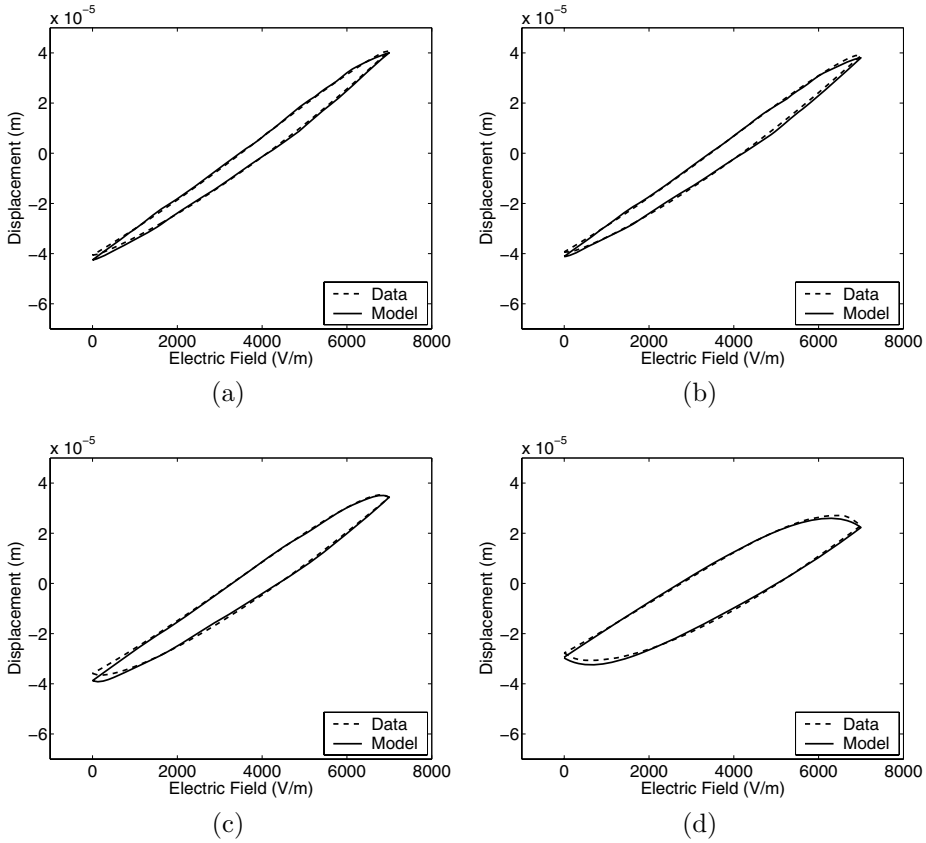


FIG. 11. *Characterization of AFM field-displacement behavior using the ODE model* (20) *with the* thermally active *kernel* (8)*, with sample rates of* (a) 0.279 *Hz,* (b) 1.12 *Hz,* (c) 5.58 *Hz, and* (d) 27.9 *Hz.*

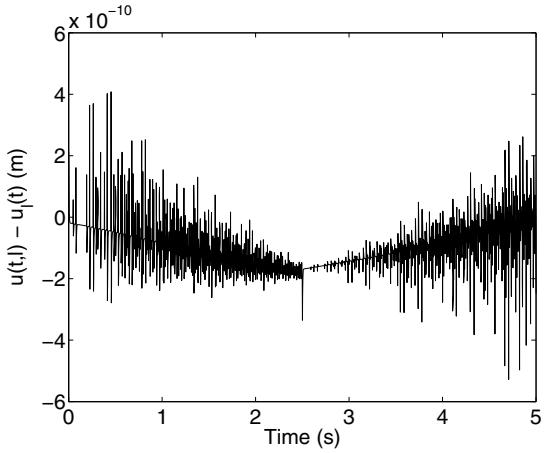


FIG. 12. *Difference between the displacement* $u(t, \ell)$ *given by the distributed model* (19) *and* $u_\ell(t)$ *given by the lumped model* (20).

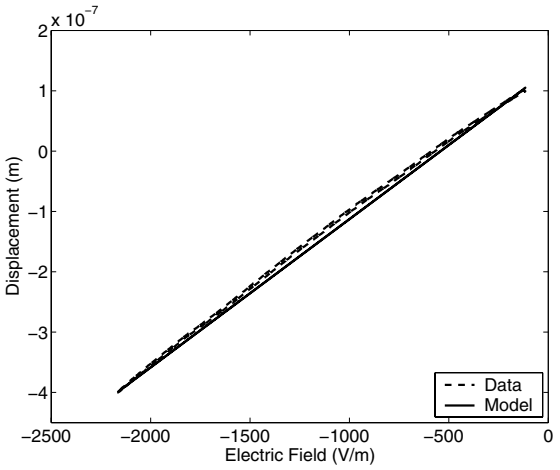


FIG. 13. *Characterization of the relation between the field and longitudinal displacements for the cylindrical actuator depicted in Figures* 2(b) *and* 7.

displacements of the cylindrical PZT shell transducer depicted in Figures 2(b) and 7. Whereas the cylindrical PZT elements employed in this design are more complex than the rod elements used in the stage design depicted in Figure 2(a), the overall transducer is simpler and has the advantage of enhanced vibration isolation and diminished hysteresis.

The experimental cylindrical transducer had a length of $\ell = 0.0396$ m, radius of $R = 0.0056$ m, and thickness of $h = 0.0015$ m. The manufacturer specifications $\rho = 7600$ kg/m^3 and $Y = 7.1 \times 10^10$ N/m^2 were employed for the density and Young's modulus, and remaining model parameters were estimated through a least squares fit to the data.

The longitudinal displacement u^N provided by (42) is compared in Figure 13 with experimental data collected under quasi-static operating conditions. We note that due to the inherent coupling between the longitudinal, circumferential, and transverse displacements, u, v, and w in the model (23), the approximate displacements (42) are also coupled and all are obtained through solution of (43)—we plot only u^N since it corresponds to measured data. For this operating regime, the dynamics were resolved with eight cubic B-splines in x and seven Fourier elements in θ, so $N_u = N_v = N_w = 56$.

The nearly linear behavior of the data reflects the low drive levels under consideration. The accurate fit provided by the model illustrates the property that the hysteretic E-P model (11) yields approximately linear behavior in low drive regimes. The fidelity of the model further illustrates the accuracy and flexibility of the modeling framework.

7. Concluding remarks. This characterization framework quantifies both the approximately linear and hysteretic properties of the PZT device employed in AFM positioning mechanisms. In the first step of the development, constitutive relations are constructed through a combination of energy analysis at the lattice level and stochastic homogenization techniques based on the assumption that certain parameters are manifestations of underlying distributions. These relations quantify the frequency-dependent hysteresis exhibited by the materials for general drive regimes

while reducing to approximately linear behavior at low drive regimes. In the second step of the development, these constitutive relations are used to construct lumped and distributed rod and shell models for the various PZT transducer geometries. The accuracy of the models is illustrated through comparison with experimental data from AFM stages.

An important property of the framework is the fact that resulting models can be approximately inverted with nearly the same efficiency as the forward models [16]. This provides a framework with the capability for providing inverse compensators for linear control design [22, 23]. The implementation of feedback control designs for high-speed scanning, using these model-based compensators, is under present investigation.

For either linear control designs employing model inverses or model-based nonlinear control designs, it is crucial that discretized models be implementable in real-time. For the stacked actuator model, discretization limits are sufficiently small (e.g., $2N = 32$ for the first-order system) to permit efficient implementation with present hardware. Furthermore, it was demonstrated that lumped models provide sufficient accuracy for the considered architecture. The behavior of the shell transducer is significantly more complex due to the coupling between longitudinal, circumferential and transverse displacements, and the construction of reduced-order models based on proper orthogonal decomposition (POD) techniques is under investigation.

Appendix. Continuity of the polarization model. We establish here the continuity of the homogenized energy model (11),

$$[P(E)](t) = \int_0^\infty \int_{-\infty}^\infty [\overline{P}(E + E_I; E_c, \xi)](t)\nu_1(E_c)\nu_2(E_I)\, dE_I\, dE_c, \tag{44}$$

as a function of both field and time in the case of negligible thermal activation. The densities ν_1 and ν_2 satisfy the conditions (12), and the kernel $\overline{P}$ has the form

$$\overline{P}(E) = \frac{E}{\eta} + P_R\delta(E; E_c, E_I)$$

specified in (6).

We first note that there are at most three values at which δ can change sign: $-E_c$, E_c, and $-E_c \le E_T \le E_c$. The third is determined by the initial dipole distribution ξ, as depicted in Figure 14(a), and is typically chosen so that $E_T = 0$ when $E + E_I = 0$.

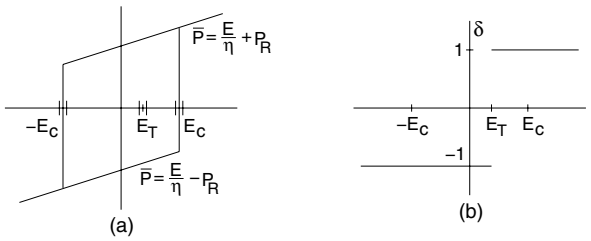


FIG. 14. (a) *Points* $E + E_I = -E_c, E_c$, *and* E_T *at which* $\delta = \pm 1$ *changes sign, and* (b) *behavior of* δ *associated with the initial dipole distribution at* $E + E_I = E_T$.

We also note that the decay conditions (12) dictate that ν_1 and ν_2 satisfy the relations

$$|\nu_2(E_I)| \leq c_2,$$

$$\int_{-\infty}^{\infty} \nu_2(E_I) dE_I \leq b_2,$$

$$\int_{0}^{\infty} \nu_1(E_c) dE_c \leq b_1,$$

where b_1, b_2, and c_2 are finite constants.

To establish the continuity of P with respect to E, we consider the behavior at field values E_0 and E_1, where without loss of generality we take $E_0 < E_1$. When integrating with respect to E_I, we decompose the interval $(-\infty, \infty)$ into seven regions delineated by the points $-E_c, E_c, E_T$, as shown in Figure 14(a). For this decomposition, we note that

$$\left|\overline{P}(E_1 + E_I; E_c, \xi) - \overline{P}(E_0 + E_I; E_c, \xi)\right| = \begin{cases} \frac{1}{\eta}(E_1 - E_0), & \text{region excludes} \\ & -E_c, E_c, E_T, \\ \frac{1}{\eta}(E_1 - E_0) + 2P_R, & \text{region includes} \\ & -E_c, E_c, E_T\,. \end{cases}$$

To consolidate notation, we define the integrals

$$I(a,b) = \int_a^b \frac{1}{\eta}(E_1 - E_0)\nu_2(E_I)\, dE_I,$$

$$I_{P_R}(a,b) = \int_a^b \left[\frac{1}{\eta}(E_1 - E_0) + 2P_R\right] \nu_2(E_I)\, dE_I.$$

It subsequently follows that

$$\begin{aligned} |P(E_1) - P(E_0)| \;\leq\; & \int_0^\infty \{|I(-\infty, -E_c - E_1)| + |I_{P_R}(-E_c - E_1, -E_c - E_0)| \\ & + |I(-E_c - E_0, E_T - E_1)| + |I_{P_R}(E_T - E_1, E_T - E_0)| \\ & + |I(E_T - E_0, E_c - E_1)| + |I_{P_R}(E_c - E_1, E_c - E_0)| \\ & + |I(E_c - E_0, \infty)|\}\, \nu_1(E_c)\, dE_c \\ \leq\; & (E_1 - E_0) \int_0^\infty \left\{\frac{4c_2}{\eta} + 3b_2\left[\frac{1}{\eta}(E_1 - E_0) + 2P_R\right]\right\} \nu_1(E_c)\, dE_c \\ \leq\; & (E_1 - E_0) b_1 \left(\frac{4c_2}{\eta} + 3b_2\left[\frac{1}{\eta}(E_1 - E_0) + 2P_R\right]\right). \end{aligned}$$

For $\varepsilon > 0$, take

$$\delta = \min\left\{\frac{\varepsilon}{b_1\left(\frac{4c_2}{\eta} + 3b_2\left[\frac{1}{\eta}(E_1 - E_0) + 2P_R\right]\right)}, 1\right\}.$$

Under the assumption that E is continuous in time and $E_0 = E(t_0), E_1 = E(t_1)$, for every $\delta > 0$ there exists $\widetilde{\delta} > 0$ such that if $|t_1 - t_0| < \widetilde{\delta}$, we are guaranteed that $|E_1 - E_0| < \delta$. It follows that if $|t_1 - t_0| < \widetilde{\delta}$, the polarization values satisfy the bound

$$|[P(E)](t_1) - [P(E)](t_0)| \leq \varepsilon,$$

thus establishing the continuity of the hysteresis model. This holds for all major and minor loops. As illustrated in Figure 6, the behavior of the model that incorporates thermal activation is smoother than the thermally inactive case considered here. For brevity, we omit the proof of this second case.

REFERENCES

[1] B.L. Ball, R.C. Smith, S-J. Kim, and S. Seelecke, *A stress-dependent hysteresis model for ferroelectric materials,* J. Intelligent Material Systems Structures, to appear.

[2] H.T. Banks, K. Ito, and Y. Wang, *Well-posedness for damped second order systems with unbounded input operators,* Differential Integral Equations, 8 (1995), pp. 587–606.

[3] H.T. Banks, R.C. Smith, and Y. Wang, *Smart Material Structures: Modeling, Estimation and Control*, Masson/John Wiley, Paris/Chichester, 1996.

[4] T.R. Braun and R.C. Smith, *Efficient implementation algorithms for homogenized energy models,* Continuum Mech. Thermodynamics, 18 (2006), pp. 137–155.

[5] W. Chen and C.S. Lynch, *A model for simulating polarization switching and AF-F phase changes in ferroelectric ceramics,* J. Intelligent Material Systems Structures, 9 (1998), pp. 427–431.

[6] W. Chen and C.S. Lynch, *A micro-electro-mechanical model for polarization switching of ferroelectric materials,* Acta Mater., 46 (1998), pp. 5303–5311.

[7] D. Croft, G. Shed, and S. Devasia, *Creep, hysteresis, and vibration compensation for piezoactuators: Atomic force microscopy application,* J. Dynamic Systems, Measurement, and Control, 23 (2001), pp. 35–43.

[8] A. Daniele, S. Salapaka, M.V. Salapaka, and M. Dahleh, *Piezoelectric scanners for atomic force microscopes: Design of lateral sensors, identification and control,* in Proceedings of the America Control Conference, San Diego, CA, 1999, IEEE Press, Piscataway, NJ, pp. 253–257.

[9] R.C.H. del Rosario and R.C. Smith, *Spline approximation of thin shell dynamics,* Internat. J. Numer. Methods Engrg., 40 (1997), pp. 2807–2840.

[10] R.C.H. del Rosario and R.C. Smith, *LQR control of thin shell dynamics: Formulation and numerical implementation,* J. Intelligent Material Systems Structures, 9 (1998), pp. 301–320.

[11] R. García and R. Pérez, *Dynamic atomic force microscopy methods,* Surface Sci. Reports, 47 (2002), pp. 197–301.

[12] P. Ge and M. Jouaneh, *Modeling hysteresis in piezoceramic actuators,* Precision Engineering, 17 (1995), pp. 211–221.

[13] R.J. Giessibl, *Advances in atomic force microscopy,* Rev. Modern Phys., 75 (2003), pp. 949–983.

[14] P.K. Hansma, V.B. Elings, O. Marti, and C.E. Bracker, *Scanning tunneling microscopy and atomic force microscopy: Application to biology and technology,* Science, 242 (1988), pp. 209–242.

[15] A.G. Hatch, *Model Development and Control Design for Atomic Force Microscopy*, Ph.D. Dissertation, Department of Mathematics, North Carolina State University, Raleigh, NC, 2004.

[16] A.G. Hatch, R.C. Smith, T. De, and M.V. Salapaka, *Construction and experimental implementation of a model-based inverse filter to attenuate hysteresis in ferroelectric transducers,* IEEE Trans. Control Systems Technol., to appear.

[17] L. Huang and H.F. Tiersten, *An analytic description of slow hysteresis in polarized ferroelectric ceramic actuators,* J. Intelligent Material Systems Structures, 9 (1998), pp. 417–426.

[18] C.M. Landis, *Non-linear constitutive modeling of ferroelectrics,* Current Opinion in Solid State and Materials Sci., 8 (2004), pp. 59–69.

[19] W. Lu, D.-N. Fang, and K-C. Hwang, *Nonlinear electric-mechanical behavior and micromechanics modelling of ferroelectric domain evolution,* Acta Mater., 47 (1999), pp. 2913–2926.

[20] J.A. Main, E. Garcia, and D.V. Newton, *Precision position control of piezoelectric actuators using charge feedback,* J. Guidance, Control, and Dynamics, 18 (1995), pp. 1068–73.
[21] J.A. Main, D. Newton, L. Massengil, and E. Garcia, *Efficient power amplifiers for piezoelectric applications,* Smart Materials and Structures, 5 (1996), pp. 766–775.
[22] J.M. Nealis and R.C. Smith, *$\mathcal{H}_\infty$ control design for a magnetostrictive transducer,* in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, IEEE Press, Piscataway, NJ, pp. 1801–1806.
[23] J.M. Nealis and R.C. Smith, *Model-based robust control design for magnetostrictive transducers operating in hysteretic and nonlinear regimes,* IEEE Trans. Control Systems Technol., to appear.
[24] M.B. Ozer and T.J. Royston, *Modeling the effect of piezoceramic hysteresis in structural vibration control,* Smart Structures and Materials 2001, Proc. SPIE 4326, SPIE, Bellingham, WA, 2001, pp. 89–100.
[25] P.M. Prenter, *Splines and Variational Methods*, Wiley, New York, 1975.
[26] J.K. Raye and R.C. Smith, *A temperature-dependent hysteresis model for relaxor ferroelectric compounds,* Proc. SPIE, Smart Structures and Materials 2004, 5383, SPIE, Bellingham, WA, 2004, pp. 1–10.
[27] G. Robert, D. Damjanovic, and N. Setter, *Preisach modeling of piezoelectric nonlinearity in ferroelectric ceramics,* J. Appl. Phys., 89 (2001), pp. 5067–5074.
[28] D. Rugar, O. Züger, S.T. Hoen, C.S. Yannoni, H.-M. Vieth and R.D. Kendrick, *Force detection of nuclear magnetic resonance,* Science, 264 (1994), pp. 1560–1563.
[29] S. Salapaka, A. Sebastian, J.P. Cleveland, and M.V. Salapaka, *High bandwidth nanopositioner: A robust control approach,* Rev. Scientific Instruments, 73 (2002), pp. 3232–3241.
[30] R.C. Smith, *Smart Material Systems: Model Development*, Frontiers in Appl. Math. 32, SIAM, Philadelphia, PA, 2005.
[31] R.C. Smith and A. Hatch, *Parameter estimation techniques for nonlinear hysteresis models,* in Smart Structures and Materials 2004, Proc. SPIE 5383, SPIE, Bellingham, WA, 2004, pp. 155–163.
[32] R.C. Smith, A. Hatch, B. Mukherjee, and S. Liu, *A homogenized energy model for hysteresis in ferroelectric materials: General density formulation,* J. Intelligent Material Systems Structures, 16 (2005), pp. 713–732.
[33] R.C. Smith and C.L. Hom, *Domain wall theory for ferroelectric hysteresis,* J. Intelligent Material Systems Structures, 10 (1999), pp. 195–213.
[34] R.C. Smith and Z. Ounaies, *A domain wall model for hysteresis in piezoelectric materials,* J. Intelligent Material Systems Structures, 11 (2000), pp. 62–79.
[35] R.C. Smith and M. Salapaka, *Model Development for the Positioning Mechanisms in an Atomic Force Microscope,* Internat. Ser. Numer. Math. 143, Birkhäuser, Basel, Switzerland, 2002, pp. 249–269.
[36] R.C. Smith and S. Seelecke, *An energy formulation for Preisach models,* Proc. SPIE, Smart Structures and Materials 2002, 4693, SPIE, Bellingham, WA, 2002, pp. 173–182.
[37] R.C. Smith, S. Seelecke, M.J. Dapino and Z. Ounaies, *A unified model for hysteresis in ferroic materials,* in Smart Structures and Materials 2003, Proc. SPIE, 5049, SPIE, Bellingham, WA, 2003, pp. 88–99.
[38] R.C. Smith, S. Seelecke, M.J. Dapino, and Z. Ounaies, *A unified framework for modeling hysteresis in ferroic materials,* J. Mech. Phys. Solids, 54, (2005), pp. 46–85.
[39] R.C. Smith, S. Seelecke, Z. Ounaies, and J. Smith, *A free energy model for hysteresis in ferroelectric materials,* J. Intelligent Material Systems Structures, 14 (2003), pp. 719–739.
[40] S.A. Wolf, D.D. Awschalom, R.A. Buhrman, J.M. Daughton, S. von Molnár, M.L. Chtchelkanova, and D.M. Teger, *Spintronics: A spin-based electronics vision for the future*, Science, 294 (2001), pp. 1488–1495.

SIAM J. Appl. Math.
Vol. 66, No. 6, pp. 2027–2048

OSCILLATIONS IN A MATURATION MODEL OF BLOOD CELL PRODUCTION*

IVANA DROBNJAK†, A. C. FOWLER†, AND MICHAEL C. MACKEY‡

Abstract. We present a mathematical model of blood cell production which describes both the development of cells through the cell cycle, and the maturation of these cells as they differentiate to form the various mature blood cell types. The model differs from earlier similar ones by considering primitive stem cells as a separate population from the differentiating cells, and this formulation removes an apparent inconsistency in these earlier models. Three different controls are included in the model: proliferative control of stem cells, proliferative control of differentiating cells, and peripheral control of stem cell committal rate. It is shown that an increase in sensitivity of these controls can cause oscillations to occur through their interaction with time delays associated with proliferation and differentiation, respectively. We show that the characters of these oscillations are quite distinct and suggest that the model may explain an apparent superposition of fast and slow oscillations which can occur in cyclical neutropenia.

Key words. maturation, mathematical model, blood cell production, chronic myelogenous leukemia, delay equation, cyclical neutropenia

AMS subject classifications. 00A69, 00A71

DOI. 10.1137/050648055

1. Introduction.

A number of hematological diseases are characterized by oscillations in the circulating density of various types of blood cells. These include chronic myelogenous leukemia (CML), cyclical neutropenia (CN), polycythemia vera (PV) and aplastic anemia (AA). Examples of blood cell counts for CML and CN are shown in Figures 1 and 2.

A review of the clinical data and a discussion of possible mechanisms for the oscillations are given by Haurie, Dale, and Mackey [10]. These mechanisms focus on the role of negative feedback control on proliferation and differentiation of blood cells within the bone marrow, together with time delays due to cell cycling and maturation. There are consequently a number of different ways in which oscillations can occur, and one object of mathematical modelling of blood cell development is to understand which of these effects may be responsible for these oscillations.

Blood cells are produced through a process of differentiation from primitive stem cells in the bone marrow. These pluripotential stem cells begin to develop along one of several different cell lineages, forming blast cells which eventually develop through a number of different stages to form the various kinds of blood cells. The most numerous are the red blood cells, or erythrocytes, whose normal density in the blood is about 5×10^6 cells $\mu\mathrm{l}^{-1}$. Their primary function is in transporting oxygen to the tissues. Platelets are formed by the fragmentation of megakaryocytes, which develop in the bone marrow. Platelets are present at levels of 5×10^5 cells $\mu\mathrm{l}^{-1}$, and their function is in blood clotting. Finally, there are a number of different white blood cells, the most

*Received by the editors December 20, 2005; accepted for publication (in revised form) May 17, 2006; published electronically October 4, 2006. This work was supported by the Natural Sciences and Engineering Research Council (NSERC grant OGP-0036920, Canada) and MITACS.
http://www.siam.org/journals/siap/66-6/64805.html
†Mathematical Institute, Oxford University, 24–29 St Giles', Oxford OX1 3LB, UK (ivana.drobnjak@linacre.oxford.ac.uk, fowler@maths.ox.ac.uk).
‡Centre for Nonlinear Dynamics, McGill University, Montreal H3G 1Y6, QC, Canada (michael.mackey@mcgill.ca).

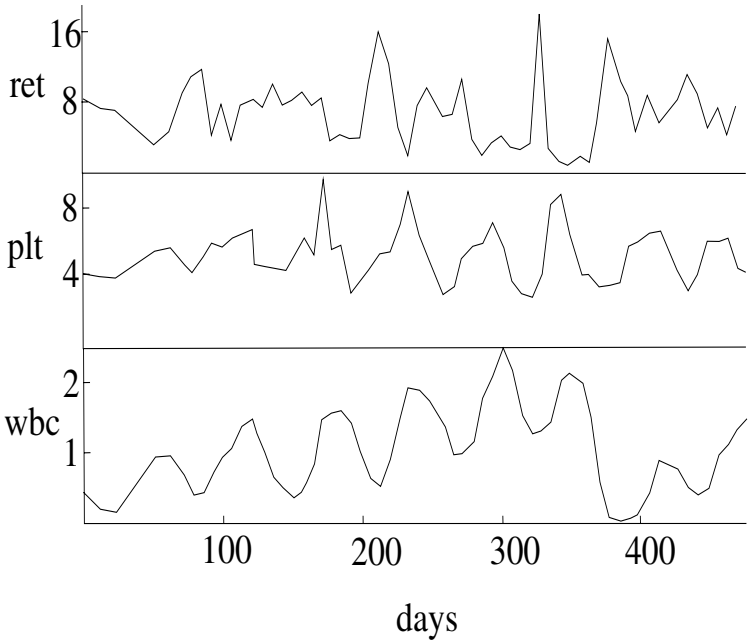


FIG. 1. *Oscillations in white blood cell, platelet, and reticulocyte numbers in a patient with chronic myelogenous leukemia. The units are white blood cells,* 10^5 *cells* μl^{-1}*; platelets,* 10^5 *cells* μl^{-1}*; reticulocytes,* 10^4 *cells* μl^{-1}*. This research was originally published in* Blood*. G. Chikkappa et al. Periodic oscillation of blood leukocytes, platelets, and reticulocytes in a patient with chronic myelocytic leukemia. Blood.* 1976;47:1023–1030. ©*the American Society of Hematology.*

common of which are neutrophils (5000 cells μl^{-1}) and lymphocytes (2000 cells μl^{-1}), which form constituent parts of the immune system. The normal levels of these cells are controlled by a number of mechanisms, and an excess or deficit of the various cell types defines certain kinds of disease; for example, anemia refers to a low red blood cell count, below 4×10^6 cells μl^{-1}.

There are a number of features in Figures 1 and 2 which are of note. In CML, there are regular oscillations in white blood cell counts with a long period ranging from 40 to 80 days (see Fortin and Mackey [7]). The other cell lines (platelets and reticulocytes, i.e., erythrocyte precursors) also oscillate in a similar fashion (Figure 1 does not show this well; see Fortin and Mackey [7] for other examples).

A similar observation is true of cyclical neutropenia. Oscillation periods are on the order of 20 days, during which there is a marked collapse of the neutrophil count to vanishingly low levels (see Dale and Hammond [4] and Guerry et al. [9]). Other cell types oscillate, but only the neutrophils appear to oscillate fairly regularly: Oscillations in other cell types (e.g., red blood cells, platelets, reticulocytes, and lymphocytes) are marked by irregularity and high frequency noise (see Guerry et al. [9]). This latter feature is well illustrated in Figure 2.

The purpose of the present paper is to throw some light on these observations by the study of a model of blood cell proliferation and differentiation. This model is similar to those of previous authors, particularly that of Mackey and Rudnicki [13], and describes the stem cell and developing (blast) cell populations as functions of time, age (time through the proliferative cell cycle), and maturation (stage in the differentiation process). Fokas, Keller, and Clarkson [6] describe a model with discrete generations in the development of blast cells, while Mackey and Rudnicki

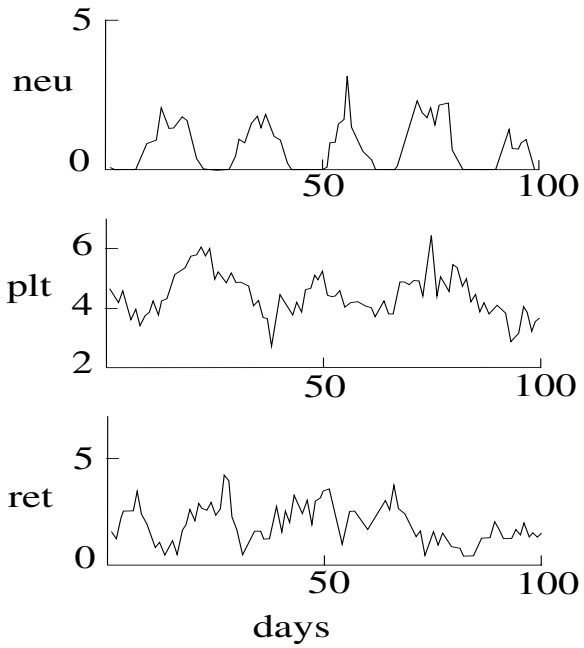


FIG. 2. *Oscillations in neutrophil, platelet, and reticulocyte numbers in a patient with cyclical neutropenia. The units are neutrophils,* 10^3 *cells* μl^{-1}*; platelets,* 10^5 *cells* μl^{-1}*; and reticulocytes,* 10^4 *cells* μl^{-1}*. This research was originally published in* Blood. *C. Haurie, D. C. Dale, and M. C. Mackey. Cyclical neutropenia and other periodic hematological disorders: A review of mechanisms and mathematical models. Blood.* 1998;92:2629–2640. ©*the American Society of Hematology.*

[13] develop a corresponding continuous model (i.e., the developmental stage is a continuous variable).

In this paper we use a continuous model to describe the development of a single cell lineage following the committal of stem cells. Three separate controls are implemented in the model, namely the proliferative control of stem cells, the proliferative control of developing blast cells, and the peripheral control of stem cell committal by circulating blood cell density. We show that variation of parameters in all three control systems can cause oscillations, and that the characters of these oscillations are very different. This allows us some potential insight into the mechanisms that may be operative in some of these dynamic blood diseases.

2. A model of maturation of blood cell production. We consider all cell lineages to consist of populations of two types, proliferative and resting phase (cf. Figure 3). These are denoted p and n, respectively, and are functions of age a (time since the inception of the proliferative cell cycle) and maturation m (degree of maturation, measured in maturation units (mat), which could be, for example, cell generation number). Also, p and n are functions of time t. Thus, we have $p = p(t, m, a)$ and $n = n(t, m, a)$. The dimensions of p and n are cells age^{-1} mat^{-1}.

In the cell population, there will be a finite number which are primitive and have not begun differentiation. These *cannot* be characterized in terms of p and n at $m = 0$, since the latter are cell densities with respect to a *and* m. The primitive stem cells are characterized by cell densities $p_0(t, a)$ and $n_0(t, a)$, such that $p_0\, da$ and $n_0\, da$ are the numbers of primitive stem cells (with $m = 0$) of age in $(a, a + da)$.

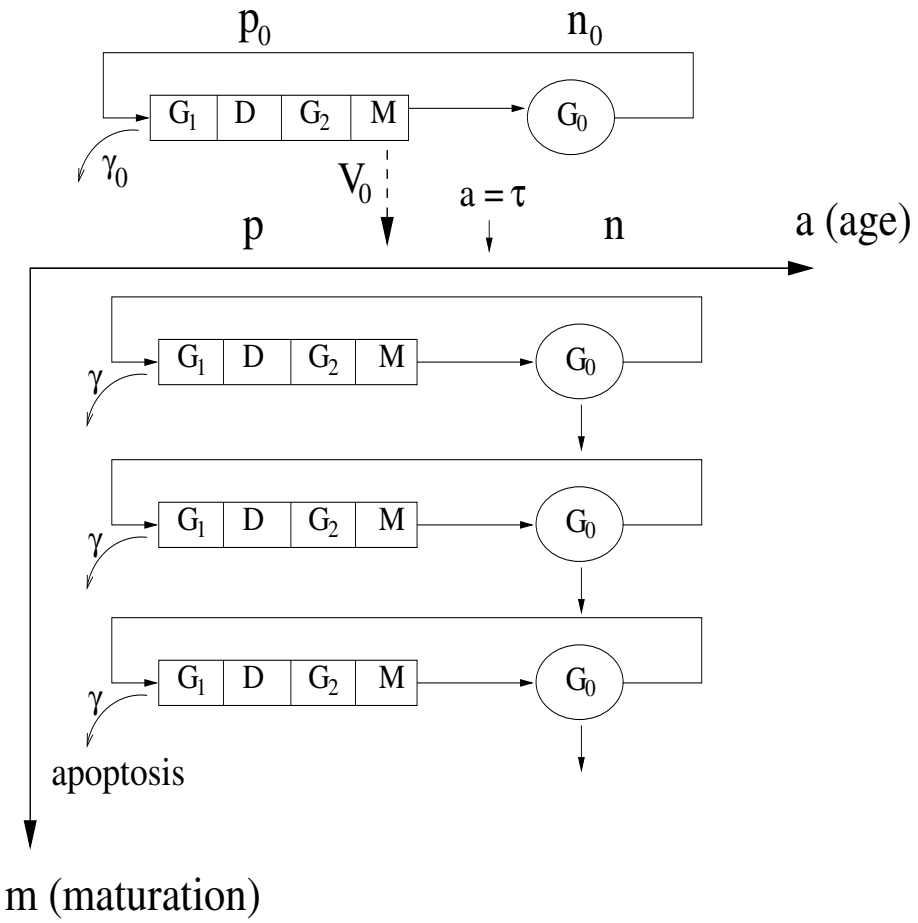


FIG. 3. *Schematic evolution of cells. Each cell ages as it goes through its cell cycle, before dividing and entering a resting (G_0) phase; at the same time, the cells mature. The time-like variables a (age) and m (maturation) are independent.*

The basic model is similar to that described by Mackey and Rudnicki [13]. It has been analyzed in various versions by Rey and Mackey [16], Crabb, Mackey, and Rey [3], and Mackey and Rudnicki [14]. However, a particular feature of these previous models was the assumption of zero maturation rate at maturation state zero. This leads to some odd behavior, which we believe arises because the model does not properly identify the role played by primitive stem cells. In our formulation of the model, we do not make this assumption.

Specifically, Mackey and Rudnicki [13], [14] assumed a model in which the cell density depended on stage of maturation m, age through the cell cycle a, and time t. As a consequence, the number of cells of zero age in the cell cycle is zero, as is the number of cells of zero maturation. As is common in age-structured population models, the migration of cells away from age zero through the cell cycle is balanced by a renewal equation, which here takes the form of recruitment from the resting cell population. However, there is in general no recruitment to the stem cell population at zero maturation, and as a consequence, the Mackey–Rudnicki model generally leads to a disappearance of the cells as they migrate down the maturation stage without

replenishment. In their model, they were able to avoid this unattractive option by allowing a rate of maturation $V(m)$ which tended to zero as $m \to 0$. The effect of this assumption is to make the time of maturation infinite, and in addition it causes strange irregular behavior to occur (see Dyson, Villella-Bresson, and Webb [5]). We consider that an implication of the lack of recruitment of stem cells is that they must be represented in the model as a compartment of finite number and therefore cannot be represented in terms of a density dependent on both m and a. In fact, consistent with previous models of stem cells (see Mackey [11], [12]), we suppose that the stem cell population can be represented as a density dependent on a and t only. This assumption removes the Mackey–Rudnicki awkwardness and does not require an artificial and unnatural choice (with its artificial consequences) of maturation rate.

The evolution of the system is illustrated schematically in Figure 3. We suppose that cell mortality occurs at a rate γ (for proliferating cells only), and that cell maturation occurs continuously at a rate V (for both proliferative and resting phases). We suppose that both γ and V may depend on maturation stage m, but not on t. Conservation of proliferative cells then implies

$$\frac{\partial p}{\partial t} + \frac{\partial p}{\partial a} + \frac{\partial (Vp)}{\partial m} = -\gamma p, \tag{2.1}$$

where the units of V are mat d^{-1} (maturation units per day). We suppose (2.1) applies during a cycle of length τ (which might depend on m), thus for $0 < a < \tau$. For $a > \tau$, the cells in the resting phase satisfy the equation

$$\frac{\partial n}{\partial t} + \frac{\partial n}{\partial a} + \frac{\partial (Vn)}{\partial m} = -Rn, \tag{2.2}$$

which differs from (2.1) by the rate of recruitment R back to the proliferative phase, where resting cell mortality is taken to be zero. Equation (2.2) applies for $a > \tau$.

At the end of the cell cycle, $a = \tau$, we apply a boundary condition describing the conversion of p to n. Mackey and Rudnicki [13] allow a very general condition, on the basis that cells at maturation M can divide to form cells at maturation $g(M) \le M$. Specifically, this implies $2p[t, M, \tau(M)]\, dM = n[t, g(M), \tau\{g(M)\}]\, dg$. If we write $m = g(M)$, $M = h(m)$ (thus $h = g^{-1}$), then this becomes

$$n[t, m, \tau(m)] = 2p[t, h(m), \tau\{h(m)\}]h'(m), \tag{2.3}$$

where $g(m) \le m$ implies $h(m) \ge m$.

A boundary condition for p at $a = 0$ follows from the recruitment condition (the renewal equation)

$$p(t, m, 0) = RN(t, m), \tag{2.4}$$

where we introduce the total resting cell density

$$N = \int_{\tau}^{\infty} n\, da. \tag{2.5}$$

Now we integrate (2.2) from $a = \tau$ to $a = \infty$, taking $n \to 0$ as $a \to \infty$ (which is necessary if there is a finite number of cells). We suppose that $V = V(m)$ is independent of a and t, and that $R = R(t, m)$ is independent of a. Then

$$\frac{\partial N}{\partial t} + \frac{\partial (VN)}{\partial m} = -RN + 2p[t, h(m), \tau\{h(m)\}]h'(m), \tag{2.6}$$

adopting (2.3).

We need to solve (2.1) for p. We use the method of characteristics and begin by applying the recruitment condition (2.4). Specifically, we apply the parametric conditions

$$t = s, \quad m = \mu, \quad a = 0, \quad p = R(s,\mu)N(s,\mu), \tag{2.7}$$

valid for $s, \mu > 0$; then the characteristic solution is

$$\begin{aligned} &a = t - s, \quad \int_{\mu}^{m} \frac{d\rho}{V(\rho)} = t - s, \\ &p = R(s,\mu)N(s,\mu) \exp\left[-\int_{s}^{t} [\gamma + V'(m)]\, dt \right]. \end{aligned} \tag{2.8}$$

Define a function $\nu(m,a)$ by

$$\int_{\nu}^{m} \frac{d\rho}{V(\rho)} = a. \tag{2.9}$$

Then $a = t - s$, $\mu = \nu(m,a)$. Also $dt = dm/V(m)$ on a characteristic; thus for $t > a$ (and also $\nu > 0$),

$$p(t,m,a) = R[t-a,\nu(m,a)]N[t-a,\nu(m,a)] \exp\left[-\int_{\nu(m,a)}^{m} \{\gamma + V'(\rho)\} \frac{d\rho}{V(\rho)} \right]. \tag{2.10}$$

Simplifying and putting $a = \tau$, we have

$$p(t,m,\tau) = R[t-\tau,\nu(m,\tau)]N[t-\tau,\nu(m,\tau)] \exp\left[-\int_{\nu(m,\tau)}^{m} \frac{\gamma\, d\rho}{V(\rho)} \right] \frac{V[\nu(m,\tau)]}{V(m)} \tag{2.11}$$

for $t > \tau$ and $\nu > 0$. Finally, (2.6) becomes

$$\begin{aligned} \frac{\partial N}{\partial t} + \frac{\partial}{\partial m}(VN) = &- RN \\ &+ 2h'(m)R[t-\tau,\nu\{h(m),\tau\}]N[t-\tau,\nu\{h(m),\tau\}] \\ &\times \exp\left[-\int_{\nu\{h(m),\tau\}}^{h(m)} \frac{\gamma\, d\rho}{V(\rho)} \right] \frac{V[\nu\{h(m),\tau\}]}{V[h(m)]}. \end{aligned} \tag{2.12}$$

Note that

$$\int_{\nu(m,\tau)}^{m} \frac{d\rho}{V(\rho)} \equiv \tau. \tag{2.13}$$

It is convenient to define a modified maturation variable ξ by

$$\xi = \int_{0}^{m} \frac{d\rho}{V(\rho)}; \tag{2.14}$$

ξ has units of time, and indeed it is equal to the elapsed time during maturation. Note that $\nu > 0$ if $\xi > \tau$. The lower limit can be chosen for convenience and allows

us to fix ξ at some reference point; here we choose this to be the initial maturation stage (note that this cannot be done if $V(0) = 0$). Define also

$$\eta(\xi) = \int_0^{h(m)} \frac{d\rho}{V(\rho)} \tag{2.15}$$

(note $\eta \geq \xi$ since $h \geq m$). Now if

$$F(m) \equiv f(\xi), \tag{2.16}$$

then we find

$$\begin{aligned} F[h(m)] &= f(\eta), \\ F[\nu\{h(m), \tau\}] &= f(\eta - \tau). \end{aligned} \tag{2.17}$$

We change the variable from m to ξ and define

$$\begin{aligned} v(\xi) &\equiv V(m), \\ M &\equiv NV \end{aligned} \tag{2.18}$$

(note that $Md\xi = Ndm$, so that M is cell density in terms of the variable ξ; the units of M are cells d^{-1}). After a little manipulation, we find

$$\frac{\partial M}{\partial t} + \frac{\partial M}{\partial \xi} = -RM + Q, \tag{2.19}$$

where

$$Q = 2\eta'(\xi) R[t - \tau, \eta - \tau] M[t - \tau, \eta - \tau] \exp\left[- \int_{\eta - \tau}^{\eta} \gamma \, d\xi \right], \tag{2.20}$$

where we write γ, R, and M as dependent on ξ rather than m. This equation applies if $t > \tau$ and $\eta > \tau$.

In order to find the form of the source term for $t < \tau$ or $\eta < \tau$, we must solve (2.1) for p using the initial data from $m = 0$ and $t = 0$. If, specifically, we have an initial condition

$$p = p_I(m, a) \quad \text{at} \quad t = 0, \tag{2.21}$$

then after some algebra we find that

$$Q = 2\eta'(\xi) p_I[\eta - t, \tau - t] v(\eta - t) \exp\left[- \int_{\eta - t}^{\eta} \gamma \, d\xi \right], \quad t < \tau, \quad \eta > t. \tag{2.22}$$

The definition of Q in $t > \eta$ and $\eta < \tau$ requires consideration of the stem cell evolution, and we now turn to this. Conservation laws for the stem cell densities p_0 and n_0 are

$$\begin{aligned} \frac{\partial p_0}{\partial t} + \frac{\partial p_0}{\partial a} &= -(\gamma_0 + V_0) p_0, \\ \frac{\partial n_0}{\partial t} + \frac{\partial n_0}{\partial a} &= -(V_0 + R_0) n_0, \end{aligned} \tag{2.23}$$

where V_0 is the rate of loss of stem cells to maturation, R_0 is the stem cell recruitment rate from the resting phase, and γ_0 is the mortality rate of stem cells in the proliferative

phase. We allow R_0, V_0, and γ_0 to depend on t, but we assume they are independent of a. Note that $V_0 \neq 0$, indeed $V_0 \neq V(0)$, as the units of V_0 and V are not even the same: V has units of mat d^{-1}, while V_0 has units of d^{-1}. Note also that p_0 and n_0 have units of cells age^{-1} (unlike p and n).

The primitive loss to maturation must balance the source for p and n at $m = 0$; thus

$$V_0 p_0 = (Vp)|_{m=0}, \quad V_0 n_0 = (Vn)|_{m=0}, \tag{2.24}$$

and the units are consistent.

Analogously to (2.4) and (2.3), we have

$$\begin{aligned} p_0(t,0) &= R_0(t) N_0(t), \\ n_0(t,\tau) &= 2p_0(t,\tau), \end{aligned} \tag{2.25}$$

where

$$N_0 = \int_\tau^\infty n_0 \, da. \tag{2.26}$$

Integration over a now yields

$$\frac{dN_0}{dt} = -V_0 N_0 - R_0 N_0 + 2p_0|_{a=\tau}, \tag{2.27}$$

and

$$(NV)|_{m=0} = N_0 V_0. \tag{2.28}$$

In order to find p_0 we must solve

$$\frac{\partial p_0}{\partial t} + \frac{\partial p_0}{\partial a} = -(\gamma_0 + V_0) p_0 \tag{2.29}$$

with parametric initial conditions

$$p_0 = p_{00}(\alpha), \quad a = \alpha > 0, \quad t = 0,$$

$$p_0 = R_0(s) N_0(s), \quad a = 0, \quad t = s > 0. \tag{2.30}$$

For $t > a$, the solution is

$$p_0 = R_0(t-a) N_0(t-a) \exp\left[-\int_{t-a}^t [\gamma_0(t') + V_0(t')] \, dt' \right], \tag{2.31}$$

whereas for $t < a$,

$$p_0 = p_{00}(a-t) \exp\left[-\int_0^t [\gamma_0(t') + V_0(t')] \, dt' \right]. \tag{2.32}$$

Putting $a = \tau$, we find

$$\frac{dN_0}{dt} = -(R_0 + V_0) N_0 + 2R_0(t-\tau) N_0(t-\tau) \exp\left[-\int_{t-\tau}^t [\gamma_0(t') + V_0(t')] \, dt' \right], \ t > \tau, \tag{2.33}$$

which prescribes the control system for N_0, analogously to that of Mackey [11]. For $t < \tau$, the equation for N_0 involves the initial condition for p_0, and we can equivalently simply prescribe initial data for N_0 there.

Finally, we complete the definition of Q in (2.19) by solving (2.1) using the initial data on $m = 0$:

$$m = 0, \quad a = \alpha > 0, \quad t = s > 0, \quad V(0)p = V_0(s)p_0(s, \alpha). \tag{2.34}$$

We find

$$p(t, \xi, a) = \frac{V_0(t-\xi)p_0(t-\xi, a-\xi)}{v(\xi)} \exp\left[-\int_0^\xi \gamma \, d\xi\right], \quad \xi < a, \quad \xi < t, \tag{2.35}$$

from which it follows that

$$Q = 2\eta'(\xi)V_0(t-\eta)p_0[t-\eta, \tau-\eta] \exp\left[-\int_0^\eta \gamma \, d\xi\right], \quad t > \eta, \quad \eta < \tau. \tag{2.36}$$

Along with (2.20) and (2.22), this completes the definition of Q for $t > 0$, $\eta > 0$ (and thus $\xi > 0$). In summary,

$$Q = \begin{cases} 2\eta'(\xi)R[t-\tau, \eta-\tau]M[t-\tau, \eta-\tau] \exp\left[-\displaystyle\int_{\eta-\tau}^\eta \gamma \, d\xi\right], & t > \tau, \quad \eta > \tau, \\ 2\eta'(\xi)p_I[\eta-t, \tau-t]v(\eta-t) \exp\left[-\displaystyle\int_{\eta-t}^\eta \gamma \, d\xi\right], & t < \tau, \quad \eta > t, \\ 2\eta'(\xi)V_0(t-\eta)p_0[t-\eta, \tau-\eta] \exp\left[-\displaystyle\int_0^\eta \gamma \, d\xi\right], & t > \eta, \quad \eta < \tau. \end{cases} \tag{2.37}$$

The two equations (2.33) and (2.19) are coupled through (2.28), which provides the requisite boundary condition for M at $\xi = 0$:

$$M = V_0 N_0 \quad \text{at} \quad \xi = 0. \tag{2.38}$$

We see that by an appropriate consideration of the primitive stem cells, we derive a coherent model which does not require $V(0) = 0$.

Many authors (for example, see Rey and Mackey [16] and Dyson, Villella-Bresson, and Webb [5]) study the differential equation (2.12) for N, under the assumption that V does tend to zero as $m \to 0$, for example,

$$V = rm. \tag{2.39}$$

The reasoning behind this is that, if primitive stem cells mature at a finite rate, then such cells will be immediately lost to $m > 0$, which makes no physiological sense, because the cell population then inexorably disappears. Only by choosing $V(0) = 0$ can we allow primitive stem cells to endure. In the present version of the model, it is also possible to allow $V(0) = 0$; for example, (2.39) would then imply

$$M \to 0 \quad \text{as} \quad \xi \to -\infty. \tag{2.40}$$

The sensitivity of the solution to this condition has led to the idea that the system may have unstable and even chaotic solutions (e.g., Crabb, Mackey, and Rey [3]), because of the degeneracy of the equation at $m = 0$. Our considerations here suggest that the requirement that $V(0) = 0$ is inaccurate, because it does not properly address the biological question of how the primitive stem cells should be described.

3. Dimensionless model. How we solve the model depends on the complexity of our assumptions about γ, R, η, and R_0. In the remainder of this paper we will assume $g(m) = m$ (thus $\eta = \xi$) and that γ and γ_0 are constant. The equation for the maturing cells, (2.19), is

$$\frac{\partial M}{\partial t} + \frac{\partial M}{\partial \xi} = -RM + Q, \tag{3.1}$$

and is a hyperbolic delay partial differential equation. Figure 4 shows the regions where the different definitions of Q apply. In regions II and segment (a) of region III, that is, for $t < \tau$ and all $\eta = \xi > 0$, Q depends on the initial data, either p_{I} (in II) or p_{00} (in III(a)). Thus we may equivalently simply choose instead to prescribe M in $0 < t < \tau$, and this we do. In fact, since ξ is finite, the part of the solution which depends on this initial data will wash out of the system in a finite time. It is therefore apparently of little concern.

We therefore confine ourselves to consideration of the definition of Q in regions I and III(b); with the assumptions we have made, we find that for $t > \tau$,

$$Q = \begin{cases} 2e^{-\gamma\tau} R[t-\tau, \xi-\tau] M[t-\tau, \xi-\tau], & \xi > \tau, \\ 2e^{-\gamma_0(\tau-\xi)} e^{-\gamma\xi} \exp\left[-\displaystyle\int_{t-\tau}^{t-\xi} V_0(t')\, dt' \right] V_0(t-\xi) R_0(t-\tau) N_0(t-\tau), & \xi < \tau. \end{cases} \tag{3.2}$$

We thus have to solve (3.1) with (3.2) in $t > \tau$, with the boundary condition (2.38) on $\xi = 0$, and prescription of an initial function for M in $0 < t < \tau$.

A principal issue of focus is how the recruitment rates R and R_0 and the committal rate V_0 depend on M, N_0, and ξ. There is very little to constrain our choice. In what follows, we assume $R_0 = R_0(N_0)$ (stem cell proliferation is controlled by stem cell density). We follow Mackey and Rudnicki [13] in supposing that R depends on the total differentiating cell population $\bar{M}$, where

$$\bar{M}(t) = \int_0^{\xi_F} M\, d\xi, \tag{3.3}$$

with ξ_F being the time of final maturation,

$$\xi_F = \int_0^{m_F} \frac{d\rho}{V(\rho)}, \tag{3.4}$$

and $m = m_F$ at full maturity. We suppose that the rate of committal V_0 should depend on the peripheral blood cell count, B, and thus $V_0 = V_0(B)$. A simple model for B is

$$\frac{dB}{dt} = M|_{\xi_F} - \gamma_B B, \tag{3.5}$$

where γ_B is the specific decay rate of the peripheral blood cells, and the source term $M|_{\xi_F}$ is the delivery rate to the blood from the maturation phase cells. Peripheral control models of similar type have been studied by Bernard, Bélair, and Mackey [1]. Assumptions of this type are liable to be important in the evolution of diseases such as cyclical neutropenia, which is thought to be due to an instability in the peripheral

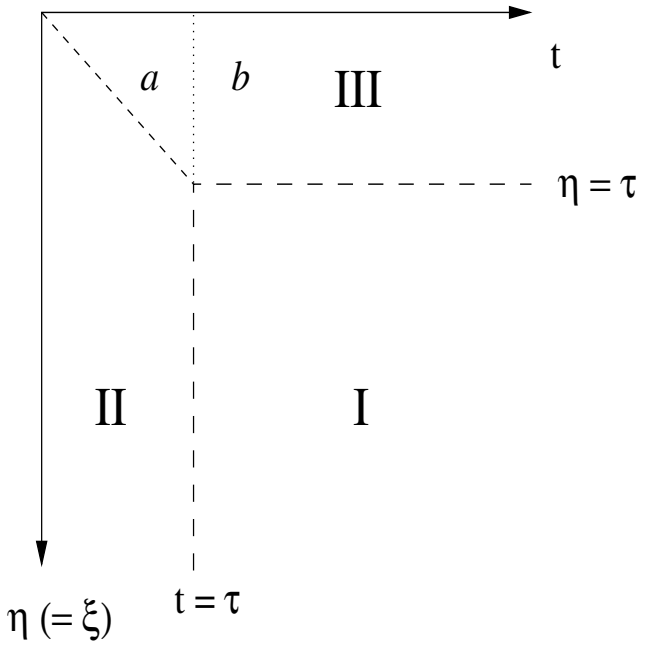


FIG. 4. *Regions of different definitions of* Q *in* (2.37). *Regions* I, II, *and* III *correspond to the first, second, and third definitions of* Q *and their locations of validity in the* (t, η) *plane. The vertical dotted line in region* III *(where* Q *is defined in terms of* p_0*) separates the region (a), where* (2.32) *applies (to the left), from (b), where* (2.31) *applies (to the right).*

control of stem cell committal. In addition, it is likely that other controls affect rate of apoptosis, maturation rate, cell cycle time, and so on.

The equation for N_0 (2.33) now takes the form

$$\dot{N}_0 = -[V_0 + R_0(N_0)]N_0 + 2e^{-\gamma_0\tau}R_0(N_{0\tau})N_{0\tau}\exp\left[-\int_{t-\tau}^{t} V_0[B(t')]\,dt'\right], \tag{3.6}$$

where $N_{0\tau} = N_0(t-\tau)$. This is precisely the model of Mackey [11] if V_0 is constant, and it has been studied by Fowler and Mackey [8] in the limit

$$V_0\tau \ll 1, \tag{3.7}$$

when it is shown that relaxation oscillations will occur for a further parameter $\mu_0 = [2e^{-(\gamma_0+V_0)\tau} - 1]/V_0\tau$ within a certain $O(1)$ range. (Note that in the notation of Fowler and Mackey's model, $\gamma = \gamma_0 + V_0$, $\delta = V_0$.) When such oscillations occur, they will propagate through the maturing cells; however, we show in this paper that the resultant amplitude of oscillations of mature blood cells is small unless amplification also occurs during maturation.

We can write (3.2) in abbreviated form as

$$Q = \begin{cases} 2e^{-\gamma\tau}R(\bar{M}_\tau)M_{\tau,\tau}, & \xi > \tau, \\ 2e^{-\gamma_0(\tau-\xi)}e^{-\gamma\xi}\exp\left[-\displaystyle\int_{t-\tau}^{t-\xi} V_0[B(t')]\,dt'\right] V_0[B(t-\xi)]R_0(N_{0\tau})N_{0\tau}, & \xi < \tau, \end{cases} \tag{3.8}$$

where $\bar{M}_\tau = \bar{M}(t-\tau)$, and $M_{\tau,\tau} = M(t-\tau, \xi-\tau)$.

We nondimensionalize the model by following the analysis of Fowler and Mackey [8], which motivates a choice of scales for the variables as follows:

$$t, \xi \sim \tau, \quad M \sim M^*, \quad N_0 = N_0^* S, \quad Q \sim \frac{M^*}{\tau},$$

$$R_0 = R_0^* h_0, \quad R = R^* h, \quad V_0 = V_0^* v_0, \quad B \sim \frac{M^*}{\gamma_B}, \tag{3.9}$$

where $R_0^*,\ R^*,\ V_0^*,\ N_0^*,\ M^*$ are determined by the control functions (so that they are $O(1)$ functions of $O(1)$ variables). For example, Mackey [11] chooses for R_0 the Hill function

$$R_0 = \frac{R_0^*}{1 + (N_0/\theta)^n}. \tag{3.10}$$

In this case, we would choose $N_0^* = \theta$, and h_0 is the Hill function

$$h_0(S) = \frac{1}{1+S^n}. \tag{3.11}$$

The dimensionless stem cell equation is

$$\dot{S} = b_0 \left[(1+\lambda_0) h_0(S_1) S_1 \exp\left(\varepsilon_0 \left\{ 1 - \int_{t-1}^{t} v_0[B(t')]\, dt' \right\} \right) - h_0 S \right] - \varepsilon_0 v_0 S, \tag{3.12}$$

where

$$\varepsilon_0 = V_0^* \tau, \quad \lambda_0 = 2e^{-(\gamma_0 + V_0^*)\tau} - 1, \quad b_0 = R_0^* \tau. \tag{3.13}$$

The dimensionless form of (3.8) is

$$\frac{\partial M}{\partial t} + \frac{\partial M}{\partial \xi} = -b h(\bar{M}) M + Q, \tag{3.14}$$

where

$$Q = \begin{cases} b(1+\lambda) & h(\bar{M}_1) M_{1,1}, \quad \xi > 1, \\ \nu b_0 (1+\lambda_0) & e^{-\alpha\xi} \exp\left[\varepsilon_0 \left\{ 1 - \displaystyle\int_{t-1}^{t-\xi} v_0[B(t')]\, dt' \right\} \right] v_0[B(t-\xi)] h_0(S_1) S_1 \quad \xi < 1, \end{cases} \tag{3.15}$$

in which

$$\nu = \frac{N_0^* V_0^*}{M^*}, \quad b = R^* \tau, \quad \lambda = 2e^{-\gamma\tau} - 1, \quad \alpha = (\gamma - \gamma_0)\tau. \tag{3.16}$$

This is analogous to the scaling used by Fowler and Mackey [8]. The boundary condition for M is

$$M = \nu v_0 S \quad \text{at} \quad \xi = 0, \tag{3.17}$$

and if

$$M = M_f \quad \text{at} \quad \xi = \xi_f, \tag{3.18}$$

then

$$\delta \dot{B} = M_f - B, \tag{3.19}$$

where

$$\delta = \frac{1}{\gamma_B \tau}, \quad \xi_f = \frac{\xi_F}{\tau}. \tag{3.20}$$

This completes the statement of the dimensionless form of the model.

Parameter values. Equation (3.12) is exactly that studied by Fowler and Mackey [8] (if $v_0 \equiv 1$). However, their model can also be interpreted as a lumped, or compartmentalized, version of (3.14) for the maturing cells. One way of enabling this is if we make the special assumption that the maturation rate $V \to 0$ as both $m \to 0$ and $m \to m_F$, as also assumed by Mackey and Rudnicki [13]. In this case the range of ξ is $(-\infty, \infty)$, and we have $M \to 0$ at both limits. Then integration of (3.14) over ξ again leads to an equation of the form of (3.12). In the present paper we assume V is finite at $m = m_F$, i.e., the mature blood cells are delivered to the bloodstream at a finite rate, and this is then the essential difference between the models with and without maturation.

In estimating the parameters, we follow Fowler and Mackey [8] in choosing $\tau \sim 2$ d, and we suppose that proliferative control is effected at typical rates $R^* \sim R_0^* \sim 2$ d^{-1}. We suppose that apoptosis rates are of order $\gamma \sim \gamma_0 \sim 0.2$ d^{-1} and that committal rates are of order $V_0^* \sim 0.05$ d^{-1}, and from these we find

$$b \sim b_0 \sim 4, \quad \lambda \sim \lambda_0 \sim 0.3, \quad \varepsilon_0 \sim 0.1. \tag{3.21}$$

The parameter α is not independent of the others, as

$$\alpha = \varepsilon_0 + \ln\left(\frac{1+\lambda_0}{1+\lambda}\right), \tag{3.22}$$

and plausibly $\alpha \approx \varepsilon_0$.

The remaining parameters are δ, ν, and ξ_f. For δ, we assume a half-life (γ_B^{-1}) of 7 hours, appropriate for neutrophils (but, for example, certainly not for erythrocytes); then

$$\delta \sim 0.15. \tag{3.23}$$

We can get some sense of the size of the remaining parameters ν and ξ_f by considering the nature of stem cells. These are difficult to isolate; indeed it is not yet clear whether genuine stem cells have ever really been isolated. The reason for this is that there are few of them, and maturing cells will typically undergo about (or at least) 20 divisions before emerging as mature blood cells. A typical numerical estimate for the total number of blast cells is 10^{12} per kg body weight, while for stem cells, a corresponding estimate is 10^6 (see Bernard, Bélair, and Mackey [1], Mackey [12]). If this is the case, then it successively implies that the parameter ν in (3.16) is very small ($\approx 10^{-6}$), and therefore also that the maturation time is long. Typical estimates of $\xi_F \approx 10$–20 days are consistent with values of $\xi_f \approx 5$–10, and in fact the small parameter ξ_f^{-1} then plays the role corresponding to that of the small parameter ε in Fowler and Mackey's [8] analysis.

Steady state. To elaborate this discussion, we now describe the steady state. For simplicity, we ignore the distinct definition of Q in $\xi < 1$ and extend the definition in $\xi > 1$ back to $\xi = 0$. The steady solution of (3.14) and (3.15) is, with $v_0 = S = 1$,

$$M = \nu e^{s\xi}, \tag{3.24}$$

where s is the unique positive solution of the pair

$$s = bh(\bar{M}) \left[(1+\lambda)e^{-s} - 1\right],$$

$$\bar{M} = \frac{\nu}{s}\left(e^{s\xi_f} - 1\right). \tag{3.25}$$

(Uniqueness follows from the fact that $\bar{M}$ is monotonically increasing with s, hence $h(\bar{M})$ is monotonically decreasing with s, and hence $bh(\bar{M})\left[(1+\lambda)e^{-s} - 1\right]$ is monotonically decreasing with s, while evidently s is increasing.) We can see that $s < \ln(1+\lambda)$, and s will be close to this value if b is large. Note also that by choosing λ_0 and ε_0 to have certain specific values which depend on λ, b, ν, and ξ_f, this solution consistently extends back to $\xi = 0$, even allowing for the distinct definition of Q in $\xi < 1$.

Numerical solutions do confirm the exponential variation of M with ξ. In general, it is found that M decreases for $0 < \xi < 1$, before subsequently increasing.

4. Periodic solutions. We are interested in finding whether periodic solutions can occur. There are three different controllers in the model, and thus three different ways in which oscillations can occur: These are described below. We utilize a reference set of parameters based on the estimates in Fowler and Mackey [8], and these are given in Table 1. They are those suggested by independent estimate, except that we take $\nu = 10^{-2}$ rather than 10^{-6}. This is partly for numerical expediency, as smaller values of ν require larger ξ_f and thus longer computation times, and also because the value of ν is not well constrained.

TABLE 1
The reference parameter values, based on estimates in Fowler and Mackey [8].

Symbol	Typical value
n	3
ε_0	0.1
λ_0	0.3
b_0	4
ν	10^{-2}
b	4
λ	0.3
ξ_f	5
δ	0.2
v^*	1
v'	1

We choose the Hill function controller (3.11) for both h and h_0; thus

$$h(\bar{M}) = \frac{1}{1+\bar{M}^n}, \quad h_0(S) = \frac{1}{1+S^n}, \tag{4.1}$$

and we take the peripheral controller v_0 to have the form

$$v_0 = [v^* - v'B]_+, \tag{4.2}$$

with default values of the amplitude and slope parameters to be $v^* = v' = 1$. The choice of a threshold in (4.2) is motivated by the observation that neutrophil populations can dwindle to zero in cyclical neutropenia, which would appear to require zero production for sufficiently high blood cell counts.

With this choice of the controller functions and using the default parameters, the steady state is stable. Instabilities arising from parameter variations are described below.

Numerical method. We have to solve the two ordinary differential equations (3.12) and (3.19), and the partial differential equation (3.14). The solution is complicated by the presence of the integral in (3.12). We define

$$U = S \exp\left[\varepsilon_0 \int^t v_0[B(t')]\, dt'\right], \tag{4.3}$$

and then S and U satisfy the pair of equations

$$\dot{S} = S\left[\frac{\dot{U}}{U} - \varepsilon_0 v_0\right],$$

$$\dot{U} = b_0\left[(1+\lambda)e^{\alpha} h_0(S_1)U_1 - h_0(S)U\right]. \tag{4.4}$$

On the assumption that S remains bounded, U grows exponentially as $U \sim \exp(\overline{v}_0 t)$, where $\overline{v}_0$ is the mean of v_0. This is likely to cause difficulty in numerical solutions, and this can be reduced by using the algebraically growing function $W = \ln U$, whence

$$\dot{S} = S\left[\dot{W} - \varepsilon_0 v_0\right],$$

$$\dot{W} = b_0\left[(1+\lambda)h_0(S_1)e^{\alpha + W_1 - W} - h_0(S)\right]. \tag{4.5}$$

In our numerical solutions, we solve (4.5) using the second order accurate improved Euler method, and we similarly solve (3.14) along the characteristics $\xi - t = \eta$, on which the function Q takes the form

$$Q = \begin{cases} b(1+\lambda)h(\bar{M}_1)M_1\,, & \xi > 1, \\ \\ b_0(1+\lambda)e^{\alpha(1-\xi)}M_0(\eta)h_0(S_1)\exp\left[W_1 - W(\eta)\right], & \xi < 1, \end{cases} \tag{4.6}$$

where $M_0(\eta) = M(\eta, 0)$ and $M_1 = M_{1,1}$, i.e., M delayed by one along the characteristic.

Accurate solutions are obtained with a time step $\Delta t = \Delta\xi = 0.05$, and these are checked against values $\Delta t = \Delta\xi = 0.02$ (which are used to give the figures).

Stem cell oscillations. Oscillations in the primitive stem cell population will occur for a finite range of the parameter λ_0/ε_0, as described by Fowler and Mackey [8], when $v_0 = 1$. For the default values of $b_0 = 4$, $n = 3$, the approximate range of instability is $0.5\varepsilon_0 \lesssim \lambda_0 \lesssim 1.5\varepsilon_0$, and this is modified in an obvious way when the peripheral controller alters the value of v_0. Figure 5 shows the oscillations which occur in the stem cell population when λ_0 is reduced to 0.05. It is an interesting fact that these oscillations are hardly manifested in the blood cell population. The apparent reason for this is that the small value of ν means that oscillations in M_0, and therefore also in M_f, are small because small perturbations propagate stably down the maturation axis. The blood cell population is therefore stable, and $B \approx M_f$.

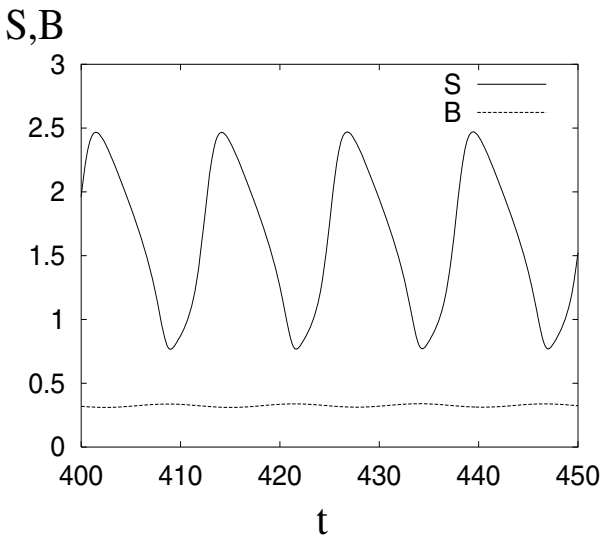


FIG. 5. *Default parameter values, except that* $\lambda_0 = 0.05$. *Stem cell oscillations are induced, without any significant effect on blood cells.*

Proliferation-controlled oscillations. We use the term proliferation-controlled oscillations to refer to oscillations induced by destabilization of the proliferative feedback control function $h(\bar{M})$. If we compare the stem cell model (3.12) (with $v_0 = 1$)

$$\dot{S} = b_0 \left[(1+\lambda_0)h_0(S_1)S_1 - h_0(S)S\right] - \varepsilon_0 S \tag{4.7}$$

with the blast cell model (along the characteristics)

$$\dot{M} = b\left[(1+\lambda)h(\bar{M}_1)M_{1,1} - h(\bar{M})M\right], \tag{4.8}$$

it is not difficult to sense that modification of the parameters b or λ may cause the blast maturation to proceed unstably.

This is what we find if λ is increased to 0.6, and the consequent oscillations are shown in Figure 6. The steady exponential proliferation of blast cells is unstable, which causes oscillations to occur in the maturation profile, and these oscillations propagate along the characteristics, as shown in Figure 7.

The oscillations have period equal to the cell cycling time, equal to one in our scaled model. A partial understanding of these oscillations is afforded by the observation that if h is constant and M is periodic with period $2\pi/\omega$, then (4.8) admits a solution

$$M = \sum_{p,q} c_{pq} e^{\sigma_q \xi + ip\omega(t-\xi)}, \tag{4.9}$$

provided that σ_q satisfies

$$\sigma = -A - Ge^{-\sigma}, \tag{4.10}$$

where

$$A = bh, \quad G = -bh(1+\lambda). \tag{4.11}$$

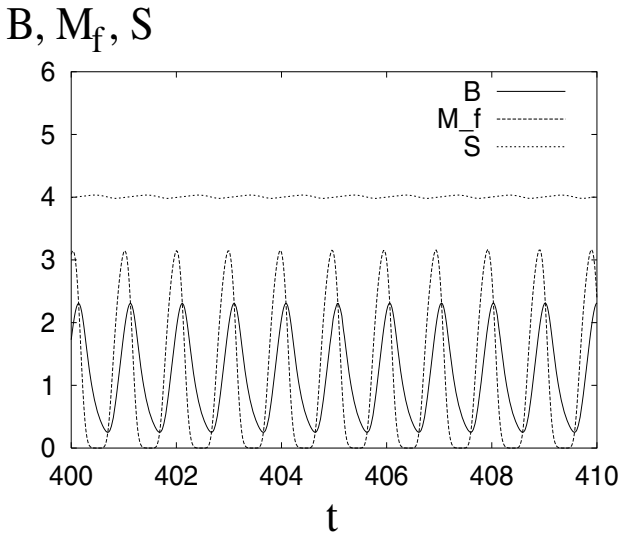


FIG. 6. *Proliferatively controlled oscillations due to increased proliferation. Default parameter values are used, except that* $\lambda = 0.6$.

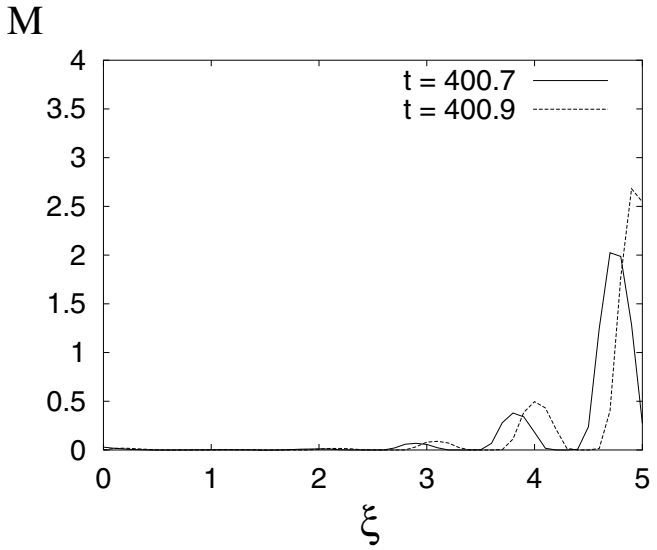


FIG. 7. *Two snapshots of the maturation profile for the numerical solution in Figure* 6. *An exponentially growing travelling wave propagates down the maturation axis.*

Since $\lambda > 0$, we have $G + A < 0$, and it is straightforward to show that there is always a single positive root, which can be labelled with $q = 0$. The others are complex (conjugates), and are labelled with increasing frequency as $q = \pm 1, \pm 2$, etc. Consideration of these complex roots then shows that for small $|G|$, $\mathrm{Re}\,\sigma_q < 0$, so that the effect of the oscillations dies away as the cells mature; this is what happens in Figure 5. However, for larger $|G|$, $\mathrm{Re}\,\sigma_q > 0$, and the oscillations grow in amplitude as the cells mature. This causes $\bar{M}$ to fluctuate, and thus also h, presumably entraining the period of the oscillations to that of the delay. This description is consistent with

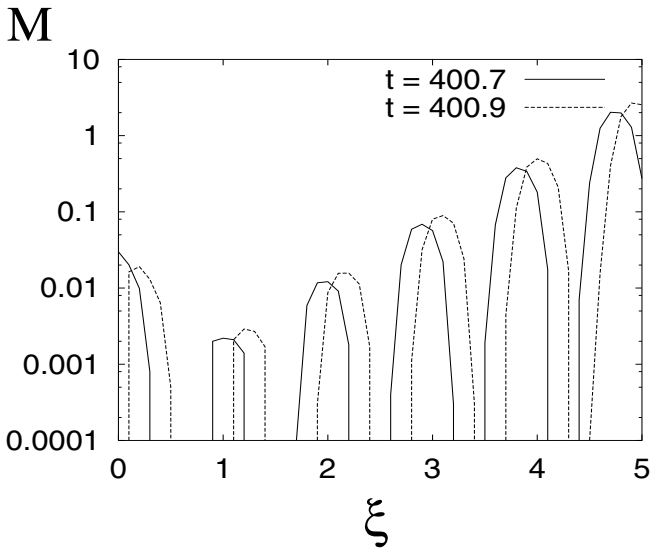


FIG. 8. *The same graph as in Figure* 7*, except that a logarithmic scale is used. The exponential increase with* ξ *is clearly seen.*

what is seen in Figure 7 (see also Figure 8). An approximate criterion for growth of periodic perturbations with ξ is when $G < -[A^2 + \pi^2]^{1/2}$, i.e.,

$$\lambda \gtrsim \left[1 + \left(\frac{\pi}{bh}\right)^2\right]^{1/2} - 1. \tag{4.12}$$

Differentiation-controlled oscillations. The final kind of oscillation that we see is induced by the peripheral control of stem cell committal through the function $v_0(B)$. These are essentially delay induced oscillations, where now the delay involved is the maturation time. Because we suppose maturation time is large, these are long period oscillations. They can be caused by increasing the sensitivity of the peripheral controller, as shown in Figure 9.

To understand the origin of these oscillations, let us suppose that $\xi_f \gg 1$, or $\xi_F \gg \tau$, meaning that the maturation time is significantly longer than the cell cycle time, or equivalently, that there are a large number of generations in the cell lineage. Let us define

$$\varepsilon = \frac{1}{\xi_m}, \tag{4.13}$$

and the slow time and maturation scales

$$T = \varepsilon t, \quad X = \varepsilon \xi. \tag{4.14}$$

We also define μ via

$$\lambda = \varepsilon \mu, \tag{4.15}$$

and suppose that $\mu = O(1)$. Essentially we are revisiting the relaxation oscillation analysis of Fowler and Mackey [8]. The partial differential equation for M takes the

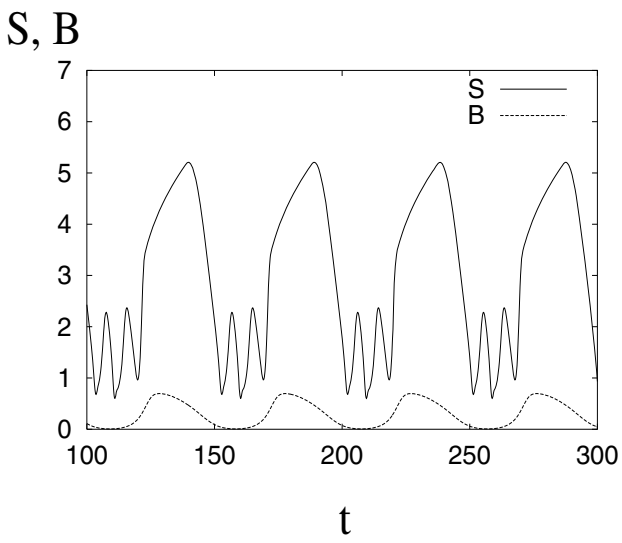


FIG. 9. *Differentiation-controlled oscillations due to enhancement of the peripheral controller function. Default parameters are used, except that* $v^* = v' = 2$.

form

$$\frac{\partial M}{\partial T} + \frac{\partial M}{\partial X} = \frac{b\left[h_{\varepsilon,\varepsilon} M_{\varepsilon,\varepsilon} - hM\right]}{\varepsilon} + \mu b h_{\varepsilon,\varepsilon} M_{\varepsilon,\varepsilon}, \tag{4.16}$$

and expanding in a Taylor series, we have

$$\frac{\partial[(1+bh)M]}{\partial T} + \frac{\partial[(1+bh)M]}{\partial X} \approx \mu b h M, \tag{4.17}$$

with the boundary condition (taking $S = 1$)

$$M = \nu v_0(B) \quad \text{at} \quad X = 0. \tag{4.18}$$

If we suppose h is constant (it is not, but it is not the dependence of h on $\bar{M}$ which causes the oscillations), then the solution of this is

$$M = \nu v_0[B(T-X)] \exp\left[\frac{\mu b h X}{1+bh}\right], \tag{4.19}$$

and the cell efflux at $X = 1$ ($\xi = \xi_f$) is

$$M(1) = \nu a v_0[B(T-1)], \tag{4.20}$$

where the amplification factor a is

$$a = \exp\left[\frac{\mu b h}{1+bh}\right]. \tag{4.21}$$

Therefore the blood cell conservation law (3.19) becomes the delay recruitment model

$$\varepsilon\delta \frac{dB}{dT} = \nu a v_0(B_1) - B. \tag{4.22}$$

This is a standard delay recruitment equation with a unique steady state. Oscillations will occur as a consequence of instability if there are solutions σ of (4.10), i.e.,

$$\sigma = -A - Ge^{-\sigma}, \tag{4.23}$$

with positive real part. The values of A and G are

$$A = \frac{1}{\varepsilon\delta}, \quad G = \frac{\nu a|v_0'|}{\varepsilon\delta}. \tag{4.24}$$

Equation (4.23) is well understood; see, for example, Mackey [11] or Murray [15, pp. 23–26]. It is a transcendental equation with an infinite number of complex roots which accumulate at the essential singularity at $\sigma = -\infty$. It follows that the set of $\operatorname{Re}\sigma$ is bounded above. Consequently, there is an instability criterion which determines when *all* the roots σ have negative real part, and this is indicated in Figure 10. The three curves in the figure are given by $G = -A$, $G = \exp[-(1 + A)]$, and the Hopf bifurcation curve $G = G_0(A)$, which is given parametrically by

$$A = -\frac{\Omega}{\tan\Omega}, \quad G_0 = \frac{\Omega}{\sin\Omega}, \tag{4.25}$$

where $\Omega \in [0, \pi]$. Since G and A are positive, oscillatory instability occurs precisely if $G > G_0(A)$. Since $G_0 \sim A$ as $A \to \infty$, the instability criterion for large A is simply $G \gtrsim A$, i.e.,

$$\nu|v_0'| \exp\left[\frac{\mu bh}{1 + bh}\right] > 1. \tag{4.26}$$

Instability is promoted by increasing $|v_0'|$, for example, as indicated in Figure 9.

5. Conclusions. In this paper we have studied the onset of oscillations in a model of blood cell production which includes a description of cell cycling and proliferation, and also of differentiation and maturation. The model formulation extends the work of previous authors by correcting an apparent inconsistency in the description of the primitive stem cell population, and also by including the simultaneous control of stem cell proliferation, stem cell committal, and blast cell proliferation. All three controls can cause oscillations for appropriate values of control parameters.

Previous results concerning stem cell oscillations are reproduced (see Figure 5), but these oscillations are harder to obtain when the parameter ν is small, and in addition they hardly affect the mature blood cell population without additional destabilization of the blast cell proliferation. The reason for this is that an $O(1)$ oscillation in the stem cell population causes only an $O(\nu)$ oscillation in the blast cell committal rate, and this amplitude propagates through the differentiating cells. Thus one consequence of stem cell paucity is that any instability in the stem cell population itself is hardly manifested in the blood cell production. From the point of view of survival and control, this is, of course, a positive result.

Instability in the proliferation of blast cells due to enhancement of the proliferative controller $h(\bar{M})$ causes oscillations which propagate down the maturation axis and are amplified as they progress. The result of this is shown in Figures 6, 7, and 8. The oscillations have a period equal to the cell cycling time. The mechanism for these oscillations appears to be a destabilization of the maturing cell amplification, together with a type of resonance which ties the period to the delay.

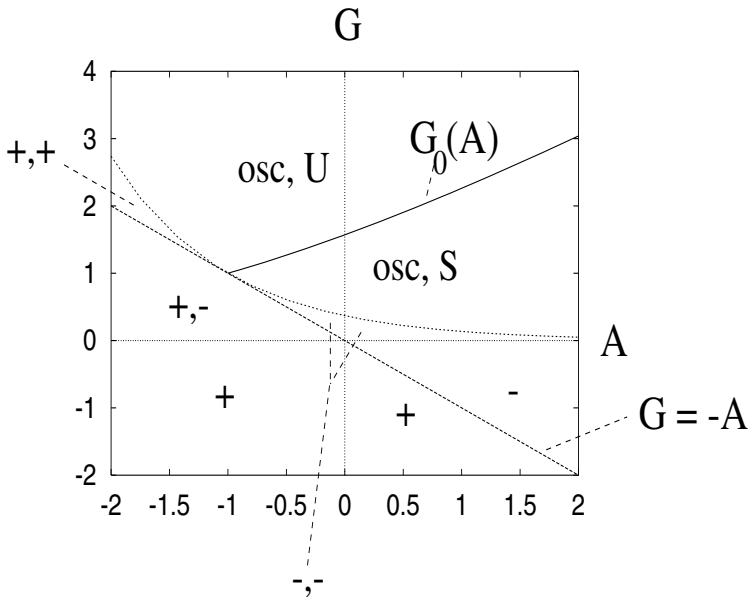


FIG. 10. *Stability map for* (4.23). *The plus and minus signs indicate the sign of real values of the growth rate* σ, *when these exist. A Hopf bifurcation occurs as* G *increases through* $G_0(A)$, *and* $G_0 \approx A$ *for large* A.

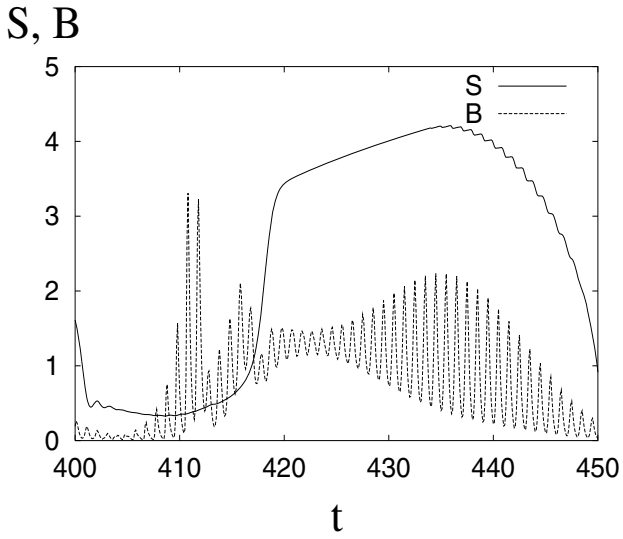


FIG. 11. *The effect of switching on all three instability mechanisms. Default parameters are used, except that* $v^* = v' = 2$, $\lambda_0 = 0.05$, *and* $\lambda = 0.6$.

The final kind of oscillation is induced by enhanced peripheral control, as seen in Figure 9. Stem cell paucity implies that $\nu \ll 1$ and, consequently, that $\xi_f \gg 1$, and thus that the oscillation period (controlled by the delay ξ_f) is long. This allows an approximate reduction of the partial differential delay equation to a simple first order differential delay equation, which is simply analyzed. In particular, if a threshold form of peripheral controller is used, blood cell counts can decrease to zero, as can be the case in cyclical neutropenia.

Finally, and as shown in Figure 11, a combination of all three destabilizing mech-

anisms can lead to oscillations which operate on both the slow, peripherally controlled time scale and the fast, proliferatively controlled one. We consider this observation to be a possible explanation of the apparent fact in Figure 2 that both reticulocytes and platelets appear to oscillate on a fast as well as a slow time scale. Further study of this behavior requires the extension of this model to accommodate multiple cell lineages.

REFERENCES

[1] S. Bernard, J. Bélair, and M. C. Mackey, *Oscillations in cyclical neutropenia: New evidence based on mathematical modeling*, J. Theoret. Biol., 223 (2003), pp. 283–298.

[2] G. Chikkappa, G. Borner, H. Burlington, A. D. Chanana, E. P. Cronkite, S. Öhl, M. Pavelec, and J. S. Robertson, *Periodic oscillation of blood leukocytes, platelets, and reticulocytes in a patient with chronic myelocytic leukemia*, Blood, 47 (1976), pp. 1023–1030.

[3] R. Crabb, M. C. Mackey, and A. D. Rey, *Propagating fronts, chaos, and multistability in a cell replication model*, Chaos, 6 (1996), pp. 477–492.

[4] D. C. Dale and W. P. Hammond IV, *Cyclical neutropenia: A clinical review*, Blood Rev., 2 (1988), pp. 178–185.

[5] J. Dyson, R. Villella-Bresson, and G. F. Webb, *A singular transport equation modelling a proliferating maturity structured cell population*, Can. Appl. Math. Q., 4 (1996), pp. 65–95.

[6] A. S. Fokas, J. B. Keller, and B. D. Clarkson, *Mathematical model of granulocytopoiesis and chronic myelogenous leukemia*, Cancer Res., 51 (1991), pp. 2084–2091.

[7] P. Fortin and M. C. Mackey, *Periodic chronic myelogenous leukemia: Spectral analysis of blood cell counts and etiological implications*, Br. J. Haematol., 104 (1999), pp. 336–345.

[8] A. C. Fowler and M. C. Mackey, *Relaxation oscillations in a class of delay differential equations*, SIAM J. Appl. Math., 63 (2002), pp. 299–323.

[9] D. Guerry IV, D. C. Dale, M. Omine, S. Perry, and S. M. Wolff, *Periodic hematopoiesis in human cyclic neutropenia*, J. Clin. Invest., 52 (1973), pp. 3220–3230.

[10] C. Haurie, D. C. Dale, and M. C. Mackey, *Cyclical neutropenia and other periodic hematological disorders: A review of mechanisms and mathematical models*, Blood, 92 (1998), pp. 2629–2640.

[11] M. C. Mackey, *A unified hypothesis for the origin of aplastic anaemia and periodic haematopoesis*, Blood, 51 (1978), pp. 941–956.

[12] M. C. Mackey, *Cell kinetic status of haematopoietic stem cells*, Cell. Prolif., 34 (2001), pp. 71–83.

[13] M. C. Mackey and R. Rudnicki, *Global stability in a delayed partial differential equation describing cellular replication*, J. Math. Biol., 33 (1994), pp. 89–109.

[14] M. C. Mackey and R. Rudnicki, *A new criterion for the global stability of simultaneous cell replication and maturation processes*, J. Math. Biol., 38 (1999), pp. 195–219.

[15] J. D. Murray, *Mathematical Biology.* I. *An Introduction*, Springer-Verlag, New York, 2002.

[16] A. D. Rey and M. C. Mackey, *Multistability and boundary layer development in a transport equation with delayed arguments*, Can. Appl. Math. Q., 1 (1993), pp. 61–81.

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 2049–2060

CELLULAR TRACTION AS AN INVERSE PROBLEM*

D. AMBROSI†

Abstract. The evaluation of the traction exerted by a cell on a planar substrate is here considered as an inverse problem: shear stress is calculated on the basis of the measurement of the deformation of the underlying gel layer. The adjoint problem of the direct two-dimensional plain stress operator is derived by a suitable minimization requirement. The resulting coupled systems of elliptic partial differential equations (the direct and the adjoint problem) are solved by a finite element method and tested against experimental measures of displacement induced by a fibroblast cell's traction.

Key words. cellular traction, linear elasticity, inverse problem, finite element

AMS subject classifications. 92C10, 92C17, 92-08, 74G75

DOI. 10.1137/060657121

Introduction. The study of the basic mechanisms of cell migration has greatly increased in the last few years. Cell locomotion occurs through a very complex interaction that involves, among others processes, actin polymerization, matrix degradation, chemical signaling, adhesion, and pulling on substrate and fibers [11]. All these ingredients occur not only in single cell migration but also in collective morphogenetic behaviors [15].

When focusing on mechanical aspects only, a major issue is the determination of the dynamical action of the cells on the environment during migration: the cells adhere, pull the surrounding matrix, and move. As a cell can have more than one hundred focal adhesion sites, each with thousands of integrins, it is quite difficult to obtain a pointwise description of the forces exerted by moving cells on a direct basis. Nevertheless, the striking improvement of nanotechnology has very recently lead to *direct* measurements of cell traction: cells are deposed on a bed of microneedles and, on the basis of the Young modulus and the moment of inertia, length, and displacement of the microneedles, one can directly obtain the exerted deflective force [4, 14]. However, these very recent experimental achievements still provide partial information on a very special configuration only: the resolution of the displacement field is at most the distance between two microneedles (2 microns), and, more important, this setting is far from being a natural migration environment.

These kinds of considerations suggests that the dynamics of cell locomotion can be fruitfully studied as an *inverse* problem, an idea that dates back to the seminal paper of Harris, Wild, and Stopak [8]. A thin elastic film over a fluid is deformed under cell traction in a wrinkled pattern, and the size of the crimps is correlated to the shear load. Unfortunately, buckling of a thin film is an essentially nonlinear phenomenon and a quantitative reconstruction of the exerted traction would call for a nontrivial stability analysis in nonlinear elasticity.

A quantitative methodology was proposed in 1996 by Dembo et al. [3], using prestressed silicone rubber, an approach further improved by Dembo and Wang in 1999

*Received by the editors April 12, 2006; accepted for publication (in revised form) June 5, 2006; published electronically October 12, 2006.
http://www.siam.org/journals/siap/66-6/65712.html
†Dipartimento di Matematica, Politecnico di Torino, corso Duca degli Abruzzi 24, 10129 Torino, Italy (davide.ambrosi@polito.it). This research has been partially supported by the FIRB 2001 Project "Metodi dell'Analisi Matematica in Biologia, Medicina e Ambiente."

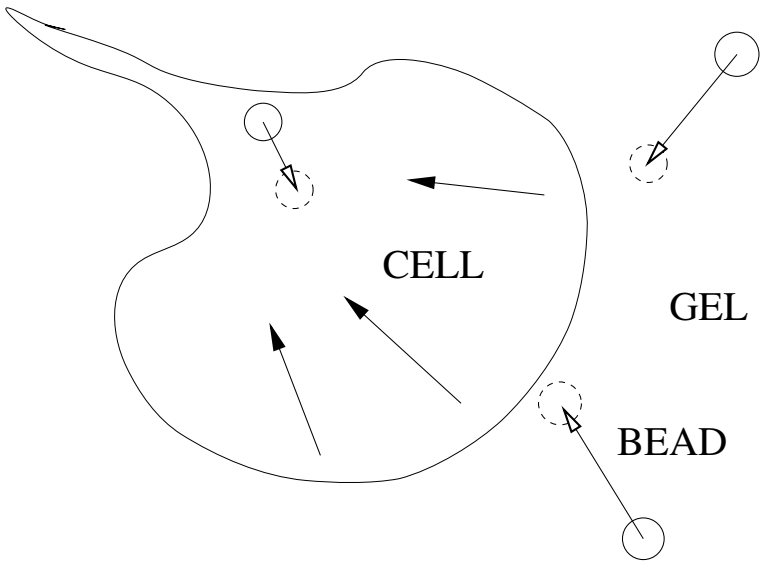


FIG. 1. *The experiment by Dembo and Wang. The cell exerts traction (filled arrows) on the gel. The beads, solidal with the substrate, move from the former position (continuous-line circle) to the new one (dashed circles). The difference in these positions gives the displacement of the gel (open arrows).*

[2]; see Figure 1. They deduce the traction exerted by a fibroblast on a polyacrilamide substrate from the measured displacement of several fluorescent beads merged into the upper layer of the gel. The gel is soft enough to remain in a linear elasticity regime, and no wrinkles form. The cellular traction is then computed by maximizing the total Bayesian likelihood of the markers displacement, predicted on the basis of the Boussinesq solution for the linear elastic half-plane with pointwise traction.

The same approach (in two spatial dimensions) is followed by Schwarz et al. [13], who numerically invert the Boussinesq integral operator, thus expressing the displacement in terms of the traction. They point out the strong dependence of the solution on small variations in the data, in particular those depending on experimental uncertainty. They therefore propose a regularization of the original problem according to the Tichonov method [7].

In this paper the same biomechanical issue studied by Dembo and Wang is addressed by a different mathematical approach, based on the classical functional analysis framework due to Lions [10]. On the basis of dimensional arguments, the three-dimensional elasticity system of equations is first reduced to a two-dimensional one. The inverse problem is then stated as minimization of the distance between the measured and the computed displacement under penalization of the force magnitude [12]. Standard derivation of the cost function leads to two sets of elastic-type problems: the direct and the adjoint. The unknown of the adjoint equation is just the shear force exerted by the cells we are looking for. The two systems of equations are then numerically solved by a coupled finite element discretization.

The paper is organized as follows. In section 1 the abstract formulation of the minimization of a cost functional is stated. The resulting adjoint problem is specified to the case of linear elasticity with an unknown body force in section 2. The methods and results of the numerical approximation of the system of equations are illustrated

in section 3, where the specific force field exerted by a fibroblast on a flat surface is calculated and qualitatively compared with the numerical results obtained by different approaches. The appendix details the assumptions that yield a two-dimensional *plain stress* system of equations in linear elasticity. Analogies and differences between the present approach and the ones reported in the relevant literature are discussed in the concluding remarks of section 4.

1. Abstract formulation. Let V be a Hilbert space equipped with the internal product $(\cdot,\cdot)$. Let $U \subset V$, and let $U = U_0 \otimes U_1$ be an orthogonal (nontrivial) decomposition of U in V. Consider the linear operator $A : U \to V$ and the problem

$$A\mathbf{u} = \mathbf{f}, \tag{1.1}$$

where $\mathbf{u} \in U$. We call the direct problem the determination of $\mathbf{u} \in U$ for a given $\mathbf{f} \in V$. If $\mathbf{u} \in U$ is given, by straightforward application of the operator A we get an $\mathbf{f} \in V$ and, in this sense, we denote $\mathbf{u} = \mathbf{u}(\mathbf{f})$. This is the inverse problem. Often only a partial knowledge of $\mathbf{u}$ is available. As a member of V, any $\mathbf{u} \in U$ can be orthogonally decomposed into two components $\mathbf{u} = \mathbf{u}_0 + \mathbf{u}_1$, where $\mathbf{u}_0 \in U_0$ and $\mathbf{u}_1 \in U_1$. If only the component $\mathbf{u}_0$ is known, the association $\mathbf{u} \to \mathbf{f}$ illustrated above cannot be carried out. Even worse, in general $\mathbf{u}_0$ and $\mathbf{u}_1$ are not in U, the domain of the operator A. These reasons make the problem ill-posed and call for a suitably supplementary condition to determine an optimal $\mathbf{f}$ on the basis of a partial knowledge of $\mathbf{u}$.

Let $\mathbf{u}_0 \in U_0$, and define the projector $P : U \to U_0$. We introduce the functional $J : V \to \mathbb{R}$:

$$J(\mathbf{f}) = (P\mathbf{u}(\mathbf{f}) - \mathbf{u}_0, \mathbf{u}(\mathbf{f}) - \mathbf{u}_0) + \varepsilon(\mathbf{f}, \mathbf{f}), \tag{1.2}$$

where $\varepsilon > 0$ is a penalty parameter.

We look for $\mathbf{g} \in V_c$ such that

$$J(\mathbf{g}) \le J(\mathbf{f}) \qquad \forall \mathbf{f} \in V_c, \tag{1.3}$$

where V_c is a convex and closed subset of V. In other words, we look for a value of $\mathbf{f}$ minimizing the distance between $\mathbf{u}$ and $\mathbf{u}_0$ under penalty of the norm of $\mathbf{f}$.

Taking the Gateaux derivative of J evaluated in $\mathbf{g}$, we obtain the equivalent condition,

$$J'[\mathbf{g}](\mathbf{f} - \mathbf{g}) \ge 0 \qquad \forall \mathbf{f} \in V_c; \tag{1.4}$$

that is,

$$(P\mathbf{u}(\mathbf{g}) - \mathbf{u}_0, \mathbf{u}(\mathbf{f}) - \mathbf{u}(\mathbf{g})) + \varepsilon(\mathbf{g}, \mathbf{f} - \mathbf{g}) \ge 0 \qquad \forall \mathbf{f} \in V_c. \tag{1.5}$$

Introduce the adjoint problem

$$A^*\mathbf{p} = P\mathbf{u}(\mathbf{g}) - \mathbf{u}_0, \tag{1.6}$$

where $A^* : V \to U_0$. Using the adjoint problem, (1.5) rewrites to

$$(\mathbf{p}, A(\mathbf{u}(\mathbf{f}) - \mathbf{u}(\mathbf{g}))) + \varepsilon(\mathbf{g}, \mathbf{f} - \mathbf{g}) \ge 0 \qquad \forall \mathbf{f} \in V_c \tag{1.7}$$

or

$$(\mathbf{p}, \mathbf{f} - \mathbf{g}) + \varepsilon(\mathbf{g}, \mathbf{f} - \mathbf{g}) \ge 0 \qquad \forall \mathbf{f} \in V_c. \tag{1.8}$$

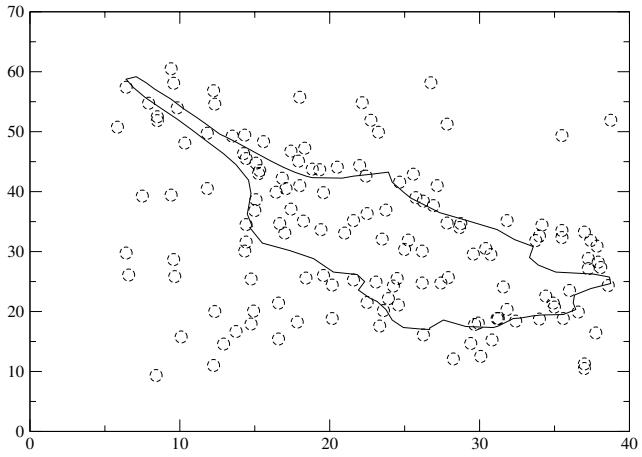


FIG. 2. *The domain Ω of the elasticity equation contains the subdomain Ω_c, the area covered by the cell and where the force applies: in the figure it is enclosed by the continuous line. The dashed circles are centered in the beads location, and their collection represents the Ω_0 subdomain where the displacement is known.*

Therefore we obtain

$$\mathbf{g} = -\frac{1}{\varepsilon}\mathbf{p}, \tag{1.9}$$

where $\mathbf{g} \in V_c$.

2. Linear elasticity. In this section we apply the general theory illustrated above to the specific problem of the small deformation of a homogeneous elastic body subject to body forces only. Let $\mathbf{u}(\mathbf{x})$ be the displacement vector field, $\mathbf{x} \in \Omega \subset \mathbb{R}^3$. Suppose that the displacement is known in a subset $\Omega_0 \subset \Omega$; the target function $\mathbf{u}_0(\mathbf{x})$ has support in Ω_0. In the problem at hand the force is exerted just on the portion of the domain where the cell lies; let us call this subdomain $\Omega_c \subset \Omega$ (see Figure 2). The cell actually adheres to the substrate just in specific small regions called focal adhesion sites, which can be experimentally localized [1]. Nothing prevents restricting the force support to these areas; this assumption is not applied here just because the information is missing for the experiment numerically reproduced in the final section.

The linear elasticity operator in strong form is

$$A\mathbf{u} = -\mu\Delta\mathbf{u} - (\mu + \lambda)\nabla\left(\nabla \cdot \mathbf{u}\right), \tag{2.1}$$

where μ and λ are the Lamé constants that characterize the material. The elasticity problem reads

$$A\mathbf{u} = \mathbf{f}, \qquad \mathbf{u}|_{\partial\Omega} = 0. \tag{2.2}$$

The direct problem consists in solving (2.2) for a given force field $\mathbf{f}$; the inverse problem is to determine the force that produces a known displacement. Here, as in all cases of practical interest, the displacement is not known in all the domain Ω, but just in Ω_0. In this case the inverse problem is ill-posed, and uniqueness of the solution has to be recovered by supplementing one more condition. Note that no issue of regularization in the sense of smooth dependence on initial data is addressed here directly: regularity will follow a posteriori.

After definition of the bilinear form

$$\sigma(\mathbf{u},\mathbf{v}) = \mu \int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}\, d\Omega + (\mu+\lambda)\int_\Omega (\nabla\cdot\mathbf{u})\,(\nabla\cdot\mathbf{v})\, d\Omega, \qquad \mathbf{u},\mathbf{v}\in H_0^1(\Omega), \tag{2.3}$$

the weak form of the problem (2.1) can be stated as follows: for a given function $\mathbf{f}\in L^2(\Omega)$, find the solution $\mathbf{u}\in H_0^1(\Omega)$ such that

$$\sigma(\mathbf{u},\mathbf{v}) = (\mathbf{f},\mathbf{v}) \qquad \forall\mathbf{v}\in H_0^1(\Omega). \tag{2.4}$$

The unique solution for the given $\mathbf{f}$ will be denoted by $\mathbf{u}(\mathbf{f})$ [5].

We note that any $\mathbf{u}\in H_0^1(\Omega)$ can be trivially rewritten as $\mathbf{u} = \chi_0\mathbf{u} + (1-\chi_0)\mathbf{u}$, where χ_o is the characteristic function of the Ω_0 set. This form provides an orthogonal decomposition of $\mathbf{u}$ as a member of $L^2(\Omega)$. Therefore we can define the projector P as follows:

$$P\mathbf{u} = \chi_o\mathbf{u}. \tag{2.5}$$

The mere knowledge of $\mathbf{u}_0$ does not allow us to obtain $\mathbf{f}$ straightforwardly. We therefore define the functional J as

$$\begin{aligned} J(\mathbf{f}) &= \int_{\Omega_0} |\mathbf{u}-\mathbf{u}_0|^2\, dV + \varepsilon\int_\Omega |\mathbf{f}|^2 dV \\ &= \int_\Omega (P\mathbf{u}-\mathbf{u}_0)\cdot(\mathbf{u}-\mathbf{u}_0)\, dV + \varepsilon\int_\Omega |\mathbf{f}|^2 dV, \end{aligned} \tag{2.6}$$

where ε is a real positive number. We look for $\mathbf{g}$ minimizing J:

$$J(\mathbf{g}) \le J(\mathbf{f}) \qquad \forall\mathbf{f}\in V_c, \tag{2.7}$$

where $V_c\subset L^2(\Omega)$ is the space of the finite energy functions with support in Ω_c. The minimization of J accomplishes the minimization of the distance of the solution from the measured value $\mathbf{u}_0$ under penalization of the magnitude of the associated force field $\mathbf{f}$. The penalty parameter ε balances the two requirements.

The minimum of J occurs in $\mathbf{g}$, where the Gateaux derivative satisfies

$$J'[\mathbf{g}](\mathbf{f}-\mathbf{g}) \ge 0 \qquad \forall\mathbf{f}\in V_c; \tag{2.8}$$

that is,

$$\int_\Omega (P\mathbf{u}(\mathbf{g})-\mathbf{u}_0)\cdot(\mathbf{u}(\mathbf{f})-\mathbf{u}(\mathbf{g}))\, dV + \varepsilon\int_\Omega \mathbf{g}\cdot(\mathbf{f}-\mathbf{g})\, dV \ge 0 \qquad \forall\mathbf{f}\in V_c. \tag{2.9}$$

Introduce the adjoint problem

$$\begin{aligned} A^*\mathbf{p} &= P\mathbf{u}-\mathbf{u}_0, \\ \mathbf{p}|_{\partial\Omega} &= 0. \end{aligned} \tag{2.10}$$

Because here A is self-adjoint, back-substitution into (2.8) gives

$$\int_\Omega A^*\mathbf{p}\cdot(\mathbf{u}(\mathbf{f})-\mathbf{u}(\mathbf{g}))\, dV + \varepsilon\int_\Omega \mathbf{g}\cdot(\mathbf{f}-\mathbf{g})\, dV \ge 0 \qquad \forall\mathbf{f}\in V_c \tag{2.11}$$

or

$$\int_\Omega \mathbf{p}\cdot(\mathbf{f}-\mathbf{g})\,dV+\varepsilon\int_\Omega \mathbf{g}\cdot(\mathbf{f}-\mathbf{g})\,dV\geq 0 \qquad \forall \mathbf{f}\in V_c, \tag{2.12}$$

and, finally, one finds the optimal body force in the sense of (2.7):

$$\mathbf{g}=-\frac{\chi_c}{\varepsilon}\mathbf{p}, \tag{2.13}$$

where χ_c is the characteristic function of the Ω_c set. The function $\mathbf{g}$ vanishes outside Ω_c because of our characterization of the set of admissible force fields V_c.

3. Numerical methods and results. For the specific application addressed in the present paper, the general three-dimensional theory illustrated above is restricted to two dimensions. The three-dimensional elasticity system of equations is approximated to a two-dimensional plain-stress one by vertical averaging along an *effective thickness* (see the appendix). The direct and inverse systems of partial differential equations deduced in section 2 now rewrite to

$$-\hat{\mu}\Delta\mathbf{u}-(\hat{\mu}+\hat{\lambda})\nabla(\nabla\cdot\mathbf{u})=-\frac{\chi_c}{\varepsilon}\mathbf{p}, \qquad \mathbf{u}|_{\partial\Omega}=0, \tag{3.1}$$

$$-\hat{\mu}\Delta\mathbf{p}-(\hat{\mu}+\hat{\lambda})\nabla(\nabla\cdot\mathbf{p})=\chi_o\mathbf{u}-\mathbf{u}_0, \qquad \mathbf{p}|_{\partial\Omega}=0. \tag{3.2}$$

Equations (3.1) and (3.2) have been discretized by a finite element method using linear basis functions on an unstructured mesh. The two resulting linear systems are solved by a global conjugate gradient method, thus avoiding any unnecessary iterative coupling [12].

The computational domain is a square with side of 100 microns. The effective Lamé constants are taken from [2], thus resulting in $\hat{\mu}=2100$ and $\hat{\lambda}=4150$ picoNewtons per micrometer. The measured displacement $\mathbf{u}_0$ and the cell contour are obtained from the same paper.

The value of the penalty parameter ε can be suitably chosen when reinterpreting the system of equations (3.1)–(3.2) as a filter. In fact, suppose for a moment that $\Omega_0=\Omega$ under periodic boundary conditions. The amplitude of the Fourier components of the solution u_k, p_k satisfies

$$\hat{\mu}k^2u_k\simeq-\frac{1}{\varepsilon}p_k, \tag{3.3}$$

$$\hat{\mu}k^2p_k\simeq u_k-u_{0,_k}; \tag{3.4}$$

that is,

$$u_k=\frac{u_{0,k}}{1+\varepsilon\hat{\mu}^2k^4}. \tag{3.5}$$

Equation (3.5) points out that the system of equations (3.1)–(3.2) acts as a filter damping the modes corresponding to wavenumbers $\gtrsim\varepsilon^{-1/4}\hat{\mu}^{-1/2}$. The penalty parameter used in the calculations illustrated below is $\varepsilon=10^{-6}$, thus algebraically damping wavelengths larger than 9 μm.

In Figure 3 a portion of the computational domain is shown: the cell contour, the displacement of the beads, and the computational mesh are plotted. The cell contour represents a boundary between internal and external elements. Note that some nodes of the mesh correspond to the original bead locations, while others do

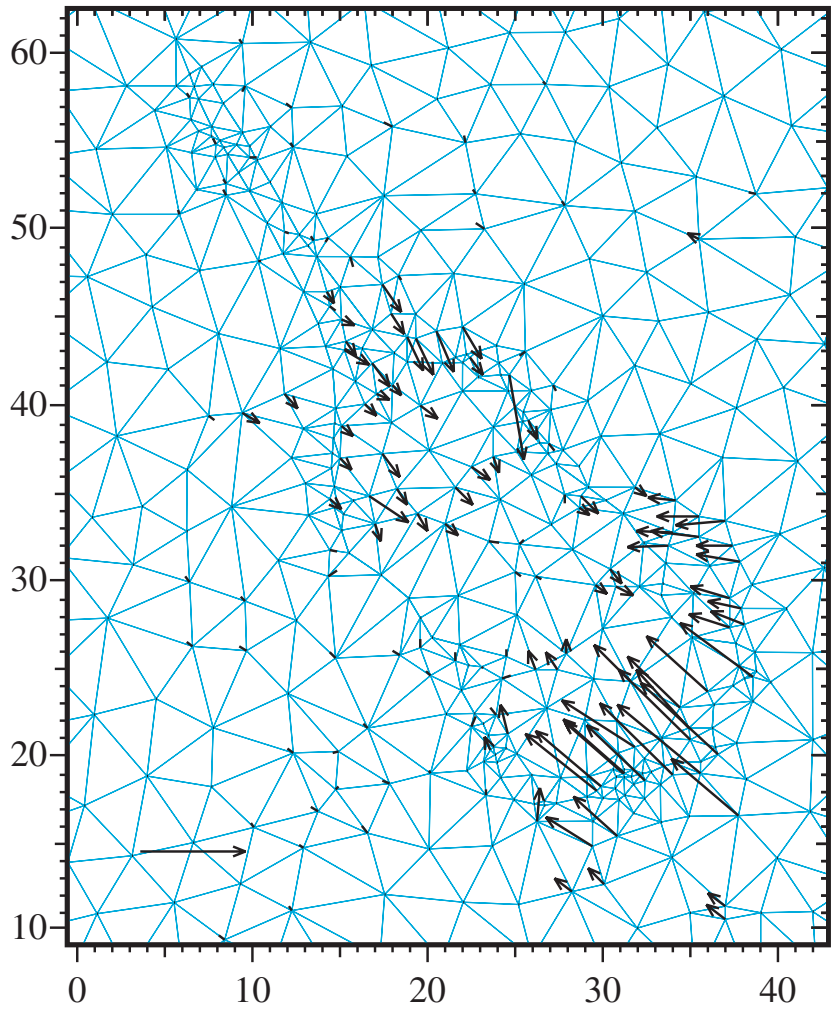


FIG. 3. *Experimental displacement of the beads merged in the upper layer of the gel, as taken from* [2]. *The computational mesh is represented in grey. The mesh satisfies two constraints: it has a node in every point where displacement has been measured, and a sequence of element sides coincides with the boundary of the cell. The reference vector at the bottom left corner is* 6 *microns long.*

not: they have been created for the sake of regularity of the computational grid. The present approach ensures full flexibility in this respect. According to the notation introduced in the preceding sections, the cell contour defines Ω_c, while the collection of the elements that have at least one node with measured displacement defines Ω_0. The computed displacement is shown in Figure 4. The mean difference between the calculated and the measured solution is

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{(\mathbf{u}_i-\mathbf{u}_{0,i})^2}=8.8\ 10^{-2}\,\mu\mathrm{m}, \tag{3.6}$$

where the sum runs over all the nodes at which $\mathbf{u}_0$ is known. The displacement of the gel matrix essentially occurs around the cell edge and, secondly, at the tail, as

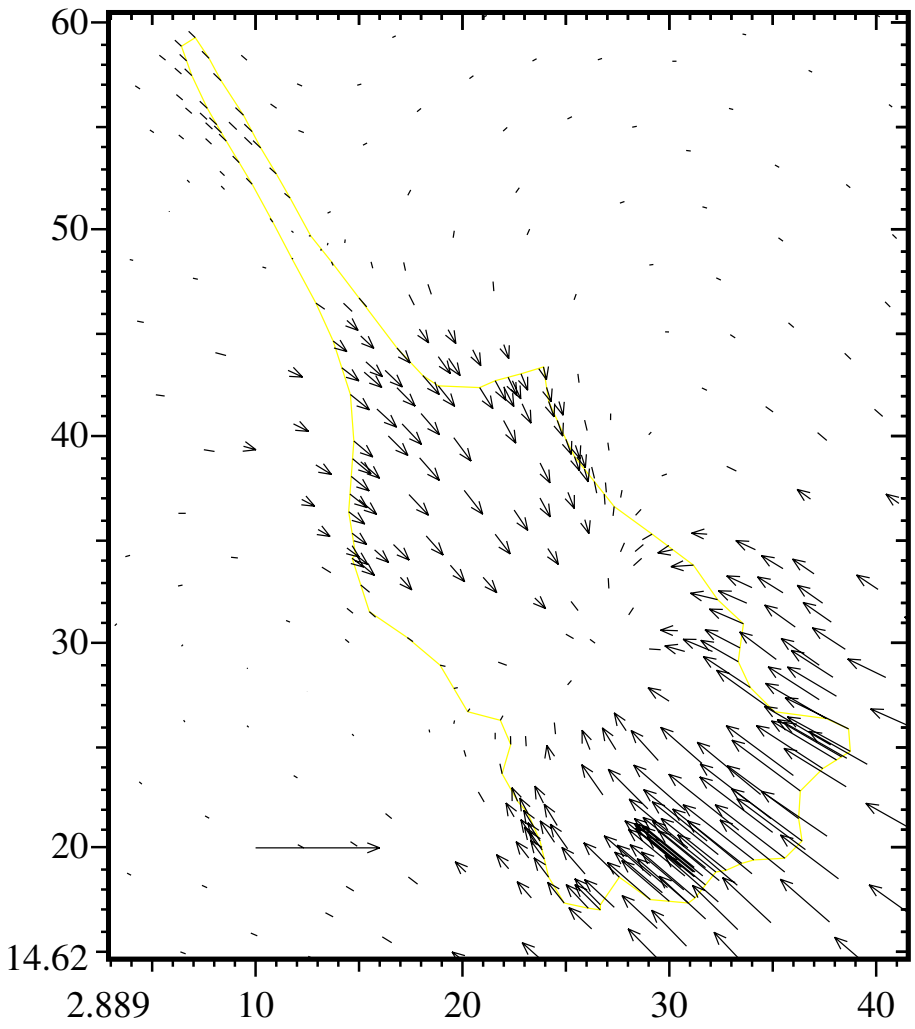


FIG. 4. *Computed displacement of the gel layer. The reference vector at the bottom left corner is* 6 *microns long.*

can be observed also in Figure 5, where the magnitude of the displacement field is represented. The computed force field is plotted in Figure 6. The qualitative behavior is very near to results shown by Dembo and Wang. The exerted force reaches the maximum value of some thousands of picoNewtons, corresponding to the remarkable stress of thousands of Pascals. A striking feature of the plot is that the cell is pulling both at the edge and at the tail, although the latter with less intensity. This result is in agreement with [2].

4. Final remarks. In this paper a novel method has been proposed for solving the inverse problem to obtain forces from displacement of an elastic body. The statement and functional derivation of a cost function yields two coupled systems of partial differential equations. The numerical solution of the inverse problem deduces the force acting on a surface on the basis of a partial knowledge of the deformation. As a specific application, the method has been applied to obtain a quantitative plot

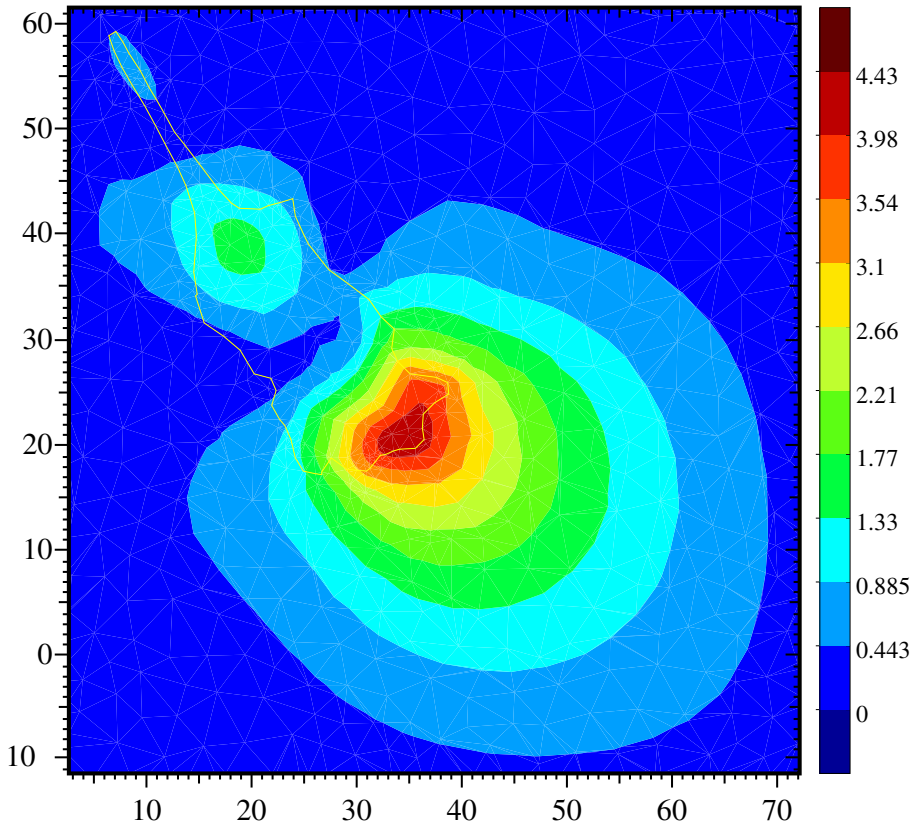


FIG. 5. *Color map of the magnitude of the computed displacement of the gel layer. The color scale is in microns.*

of the forces exerted by a fibroblast cell on a gel substratum.

The proposed approach is quite general in not assuming a pointwise nature of the surface force. The direct solution of the elasticity equations makes it possible, in principle, to apply the proposed methodology to a variety of situations for which the Boussinesq solution does not apply: for instance, nonhomogeneous substrate or nonisotropic prestress of the gel.

The deduction of the system of equations from a minimum principle provides a clear meaning to the regularization procedure, while the statement of the equations in precise functional spaces ensures the well-posedness of the direct and inverse problems. Finally, the numerical integration of the equations by a finite element discretization ensures full geometrical flexibility to account for the complicated contour of a cell or to operate local mesh refinements suggested by accuracy arguments.

Appendix. Two-dimensional modeling. The extracellular matrix is deformed by the traction exerted by the cell, and in principle the strain of the gel can be predicted solving the force balance equation for an elastic three-dimensional body with suitable boundary conditions. However, an a priori consideration of the characteristics of the displacement field in the elastic substrate suggests some suitable simplification of the full three-dimensional model.

As a matter of fact, the substrate layer is very shallow: its horizontal extension

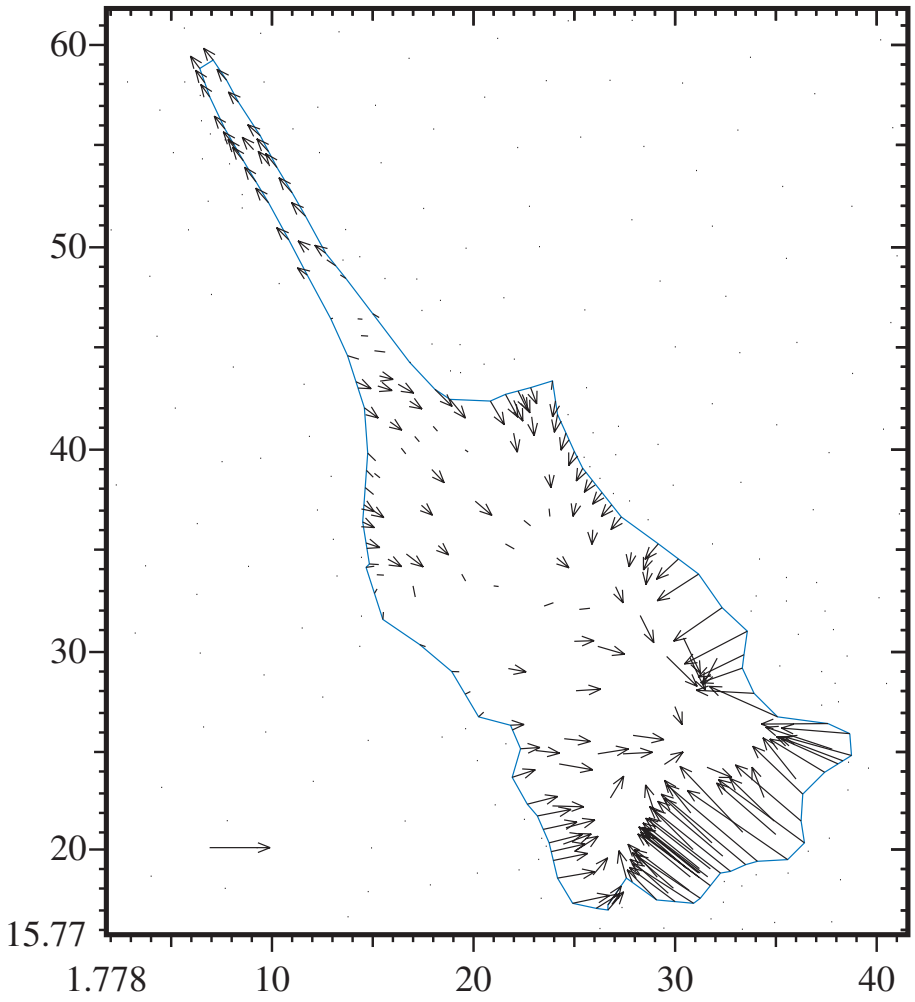


FIG. 6. *Computed force field exerted by the fibroblast on the gel layer. The magnitude of the reference vector at the bottom left corner corresponds to* 10^3 *picoNewtons.*

($\sim$ 1mm) is much larger than the vertical one ($\sim 10-100\mu$m). However, this geometrical scale ratio by itself does not allow any estimate of the behavior of the solution along the vertical direction, a consideration that could be helpful in addressing a suitable approximate model.

The useful observation is instead that the vertical displacement is very small when compared to the horizontal one. Let us denote by X and Z the typical scale lengths of the horizontal and vertical displacements, respectively. The length X is of the order of the diameter of a cell ($\sim 20\mu$m), while Z is to be determined. In the following we indicate by U and W the typical horizontal and vertical displacements.

The boundary conditions of the three-dimensional force balance equation at the cell-gel interface read

$$\mu(u_z + w_x) = \tau \qquad \text{at } z = 0, \tag{A.1}$$

$$(\mu + \lambda)w_z + \lambda u_x = 0 \qquad \text{at } z = 0. \tag{A.2}$$

As $\mu \sim \lambda$, inspection of (A.2) yields that it must be $W/Z \sim U/X$. According to the experimental evidence, the vertical displacement is much smaller than the horizontal one: $U \gg W$. Putting this observation into (A.1), one deduces that $|w_x| \ll |u_z|$, and therefore the horizontal derivative of the vertical displacement can be neglected therein.

It follows that to first order the horizontal displacement vanishes in the layer at the depth $Z \sim \mu U/\tau$. In the mentioned experiments

$$U_{max} = 6\,\mu\text{m}, \qquad \mu = \frac{E}{2(1+\nu)} = \frac{E}{3} = 2000\,p\text{N}/\mu\text{m}^2, \qquad \tau_{max} = 10^4\,p\text{N}/\mu\text{m}^2,$$

thus yielding $Z_{max} \sim 1\mu$m, much smaller than the height of the matrigel layer (70μm). According to the observations above, we adopt a two-dimensional "plain stress" model, obtained by vertical integration of the three-dimensional equation of the linearized elasticity along the effective thickness Z_{max}. Therefore one gets

$$-\hat{\mu}\Delta\mathbf{u} - (\hat{\mu} + \hat{\lambda})\nabla\,(\nabla \cdot \mathbf{u}) = \boldsymbol{\tau}, \tag{A.3}$$

where $\boldsymbol{\tau}$ is the shear stress at the surface. The quantities

$$\hat{\mu} = Z_{max}\frac{E}{2(1+\nu)}, \qquad \hat{\lambda} = Z_{max}\frac{E\nu}{1-\nu^2} \tag{A.4}$$

are the effective Lamé constants of the two-dimensional model and have the dimension of force per unit length [3, 9]. It is to be remarked that the above determination of Z_{max} is based on dimensional arguments only.

Acknowledgments. The author is indebted to Pierluigi Rozza, Enrico Serra, and Luigi Preziosi for fruitful discussions about the content of this paper.

REFERENCES

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*, 3rd edition, Garland, New York, 1994.

[2] M. Dembo and Y. L. Wang, *Stresses at the cell-to-substrate interface during locomotion of fibroblasts*, Biophys. J., 76 (1999), pp. 2307–2316.

[3] M. Dembo, T. Oliver, A. Ishihara, and K. Jacobson, *Imaging the traction stresses exerted by locomoting cells with elastic substratum method*, Biophys. J., 70 (1996), pp. 2008–2022.

[4] O. du Roure, A. Saez, A. Buguin, R. H. Austin, P. Chavrier, P. Siberzan, and B. Ladoux, *Force mapping in endothelial cell migration*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 2390–2395.

[5] G. Fichera, *Existence theorems in elasticity*, in Handbuch der Physik, Band VIa/2, C. Truesdell, ed., Springer-Verlag, Berlin, 1972, pp. 347–389.

[6] G. G. Galbraith and M. P. Sheetz, *A micromachined device provides a new bend on fibroblast traction forces*, Proc. Natl. Acad. Sci. Cell Biol., 94 (1997), pp. 9114–9118.

[7] P. C. Hansen, *Rank-Deficient and Discrete Ill-posed Problems: Numerical Aspects of Linear Inversion*, SIAM Monogr. Math. Model. Comput., SIAM, Philadelphia, 1997.

[8] A. K. Harris, P. Wild, and D. Stopak, *Silicone rubber substrata: A new wrinkle in the study of cell locomotion*, Science, 208 (1980), pp. 177–179.

[9] L. Landau and E. Lisfchitz, *Théorie de l'Élasticité*, Éditions Mir, Moscow, 1967.

[10] J. L. Lions, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod et Gauthier–Villard, Paris, 1968.

[11] A. J. Ridley, M. A. Schwartz, K. Burridge, R. A. Firtel, M. H. Ginsberg, G. Borisy, J. T. Parsons, and A. R. Horwitz, *Cell migration: Integrating signals from front to back*, Science, 302 (2003), pp. 1704–1709.

[12] A. Rincon and I. S. Liu, *On numerical approximation of an optimal control problem in linear elasticity*, Divulg. Mat., 11 (2003), pp. 91–107.

[13] U. S. Schwarz, N. Q. Balaban, D. Riveline, A. Bershadsky, B. Geiger, and S. A. Safran, *Calculation of forces at focal adhesions from elastic substrate data: The effect of localized force and the need for regularization*, Biophys. J, 83 (2002), pp. 1380–1394.

[14] J. L. Tan, J. Tien, D. M. Pirone, D. S. Gray, K. Bhadriraju, and C. S. Chen, *Cells lying on a bed of microneedles: An approach to isolate mechanical forces*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 1484–1489.

[15] A. Tosin, D. Ambrosi, and L. Preziosi, *Mechanics and chemotaxis in the morphogenesis of vascular networks*, Bull. Math. Biol., 68 (2006), pp. 1819–1836.

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 2061–2084

HOMOGENIZATION OF A WIRE PHOTONIC CRYSTAL: THE CASE OF SMALL VOLUME FRACTION*

GUY BOUCHITTÉ† AND DIDIER FELBACQ‡

Abstract. We consider the diffraction of a monochromatic incident electromagnetic field by a bounded obstacle made of parallel high conducting metallic fibers of finite length which are periodically disposed. Our goal is to study the asymptotics of this three-dimensional Maxwell problem as the period, the thickness of the rods, and the resistivity are simultaneously small. We solve this problem in the case where the filling ratio of fibers vanishes but their capacity remains positive. We find a limit volumic density of current which results in a nonlocal constitutive relation between the electric and displacement fields. This extends previous results obtained in the polarized case where a two-dimensional effective local equation was found with a possibly negative effective permittivity.

Key words. Maxwell equations, metamaterials, homogenization, thin structures, two-scale convergence, capacitary potentials, nonlocal effects

AMS subject classifications. 28A33, 31C15, 35B27, 35Q60

DOI. 10.1137/050633147

1. Introduction.

Photonic crystals are artificial periodic structures which possess a photonic band gap: the propagation of electromagnetic waves is impossible in them over some intervals of frequencies [11]. This phenomenon is very close to that of the band structure for electrons in natural crystals. Many studies have been devoted to dielectric devices, that is, photonic crystals made with a dielectric material. However, there is a growing interest in metallic crystals, which now form an emerging field of research, because of their particular behavior in the low-frequency domain, i.e., their effective properties. One of these properties is the existence of a plasma frequency for structures made of wires: for frequencies below some cut frequency the electric field is evanescent in the structure. One of the pioneering studies in this domain was that of Pendry et al. [20], where the problem was analyzed in terms of plasmons and the renormalized effective mass of electrons due to their confinement in very thin wires. This work used a crude model for the metallic properties of the medium. We have presented in [14] another approach based on a homogenization method involving Maxwell equations only.

In this paper, we consider the diffraction of a given monochromatic incident electromagnetic field (wavenumber $k_0 = \frac{2\pi}{\lambda}$) by a bounded obstacle made of parallel high conducting metallic fibers of finite length. These fibers of cross section r are x_3-parallel and placed periodically in (x_1, x_2) (period η) in a cylinder $\Omega = \mathcal{D} \times (0, L)$, where $\mathcal{D}$ denotes a connected bounded open subset of $\mathbb{R}^2$. The fibers share the same conductivity represented by parameter σ.

We are interested in all possible asymptotics of the diffraction problem as the small parameters η, r, σ^{-1} tend simultaneously to zero. The situation is well understood in the case where Ω is an infinite cylinder $\mathcal{D} \times \mathbb{R}$ and the incident field is $E||$

*Received by the editors June 6, 2005; accepted for publication (in revised form) July 26, 2006; published electronically October 16, 2006.
http://www.siam.org/journals/siap/66-6/63314.html
†Département de Mathématiques, Université de Toulon et du Var, BP 132, F-83957 La Garde Cedex, France (bouchitte@univ-tln.fr).
‡GES UMR-CNRS 5650, Université de Montpellier II, CC074, Place E. Bataillon, 34095 Montpellier Cedex 05, France (felbacq@GES.univ-montp2.fr).

or $H||$ polarized: if, for simplicity we take $\sigma = +\infty$ (perfectly conducting fibers), the diffraction problem reduces to a scalar Helmholtz equation with homogeneous Dirichlet ($E||$ case) or Neumann ($H||$ case) conditions on a perforated domain. Then the following results can be proved using the following classical homogenization techniques (see [7, 8, 9] in the case of diffusion equations). Two critical scales of the radius r_η of the scatterers appear to be relevant:

(1) $E||$ *case:* $\log r_\eta \sim -\eta^{-2}$. Denoting by γ^{-1} the limit of $\eta^2|\log r_\eta|$, the limit problem is equivalent to the diffraction by the whole set Ω filled up with a homogeneous medium of relative permittivity $\varepsilon^{\text{eff}} = 1 - 2\pi\gamma/k_0^2$. This simple fact has striking applications, namely the existence of a suitable frequency (called plasma frequency) under which ε^{eff} becomes negative and the field decreases exponentially inside Ω. When γ is infinite (in particular, when the size of the rods is of order η), the limit field vanishes and Ω behaves like a perfectly conducting medium. As shown by numerical experiments (see [14]), these behaviors are observed even if the ratio $\frac{\eta}{\lambda}$ is not so small.

(2) $H||$ *case:* $r_\eta \sim \eta$. In this case, Ω behaves like a homogeneous medium whose effective permittivity ε^{eff} is greater than 1 through a corrective term which is computed by solving as usual an elementary periodic Neumann problem on a unit cell. Additionally, the magnetic field induces a small turning current on the surface of each small cylinder, which as $\eta \to 0$ becomes a surface current on the boundary of Ω (accordingly the magnetic field becomes discontinuous across $\partial\Omega$). In case $r_\eta \ll \eta$ (as was assumed in the $E||$ case), the equivalent structure is transparent (i.e., $\varepsilon^{\text{eff}} = 1$).

Insofar as the Maxwell problem can be decomposed into two problems of type $E||$ and $H||$ (for instance, if perfectly conducting fibers with infinite length are considered), we can draw the conclusion that for an intermediate size of rods r_η (i.e., $r_\eta \ll \eta$ and $\gamma = +\infty$, for example $r_\eta \sim \eta^2$), the homogenized medium is perfectly reflective for the vertical component of the electric field, whereas it is transparent for that of the magnetic field. Hence very strong polarizing properties are obtained at any incidence.

However, when the diffracting domain Ω is bounded, it is no longer possible to split the original problem into independent problems of type $E||$ and $H||$. As a consequence we have to deal with the full three-dimensional (3D) Maxwell system, which is mathematically much more involved. In particular we will show that nonlocal effects appear in the limit as $\eta \to 0$.

In this paper, we deal only with the case where $r_\eta \ll \eta$, that is, the filling ratio θ_η of rods vanishes. The high conductivity is modeled as a stiff parameter depending on η:

$$\sigma_\eta = \frac{\kappa\,\varepsilon_0\omega}{\theta_\eta}, \tag{1.1}$$

where $\kappa > 0$ is a fixed (dimensionless) constant. Our model includes the case $\kappa = +\infty$, which corresponds to infinitely conducting fibers (see Remark 3.4).

Denoting by T_η the union of the rods, the relative permittivity is then defined as follows:

$$\varepsilon_\eta(x) = \begin{cases} 1 + i\,\dfrac{\kappa}{\theta_\eta} & \text{if } x \in T_\eta, \\ 1 & \text{if } x \in \mathbb{R}^3 \setminus T_\eta. \end{cases}$$

Let (E^i, H^i) denote a given incident electromagnetic wave. As it satisfies the harmonic Maxwell system in the vacuum (that is, $\operatorname{curl} E^i = i\,\omega\mu_0\,H^i$, $\operatorname{curl} H^i = -i\,\omega\varepsilon_0\,E^i$ in all

$\mathbb{R}^3$), it is smooth and solves the homogeneous Helmholtz equation on $\mathbb{R}^3$ with wave number k_0.

For every value of the parameter η, the total electromagnetic field (E_η, H_η) is then characterized as the unique solution of the time harmonic Maxwell system

$$\begin{cases} \operatorname{curl} E_\eta = i\omega\mu_0 H_\eta, \\ \operatorname{curl} H_\eta = -i\omega\varepsilon_0\varepsilon_\eta E_\eta \end{cases} \tag{1.2}$$

such that the diffracted field $(E_\eta^d, H_\eta^d) := (E_\eta - E^i,\ H_\eta - H^i)$ satisfies the so-called outgoing wave condition at infinity. This condition can be written in the following way due to Silver and Müller (see [22]):

$$(E_\eta^d, H_\eta^d) = O\left(\frac{1}{|x|}\right), \qquad \omega\varepsilon_0\left(\frac{x}{|x|} \wedge E_\eta^d\right) - k_0 H_\eta^d = o\left(\frac{1}{|x|}\right). \tag{1.3}$$

For the existence and the regularity of the solution (E_η, H_η), we refer, for instance, to [6].

The main mathematical issue of the paper consists in proving that (E_η, H_η) does converge weakly in $(L^2_{loc}(\mathbb{R}^3))^3$ and that the limit is characterized as the unique solution of a suitable diffraction problem.

It turns out that, like in the $E||$ case mentioned before (see [14]), the limit equation will depend on the *average capacity* of the fibers, which is described by the following limit:

$$\gamma := \lim_{\eta\to 0} (\log r_\eta\ \eta^2)^{-1}. \tag{1.4}$$

Clearly, assuming that such a limit is finite forces $r_\eta \ll \eta$, so that the fibers seem to disappear as $\eta \to 0$. However, it is crucial to record the information on the behavior of the electric field E_η in the fibers. To that aim, in the same spirit as in [3], we introduce the rescaled field

$$F_\eta = \kappa\, E_\eta \frac{1_{\mathbf{T}_\eta}}{\theta_\eta}.$$

This vector function F_η represents the average displacement field in the fibers. It can be seen (up to a multiplicative constant) as a density of current. We may rewrite the second equation in (1.2) as

$$\operatorname{curl} H_\eta = -i\omega\varepsilon_0(E_\eta + i\,\kappa F_\eta). \tag{1.5}$$

We are going to prove that the sequence $\{F_\eta\}$ is bounded in $L^1(\Omega)$ and that its weak limit F as a vector measure is parallel to the fibers and absolutely continuous with respect to the Lebesgue measure. Consequently we will write F in the form $F = J\,e_3$, where $J \in L^1(\Omega)$ can be seen as an effective density of current.

Assuming that (E_η, H_η) converges weakly to (E, H) in L^2_{loc}, we may write down the limit Maxwell system in the form

$$\begin{cases} \operatorname{curl} E = i\omega\mu_0 H, \\ \operatorname{curl} H = -i\omega\varepsilon_0(E + i\,Je_3). \end{cases} \tag{1.6}$$

The fundamental point is to derive a relation between the current density $J(x)$ and the macroscopic electric field $E(x)$. Our results show evidence of the three following important features:

– For $0 < \gamma < +\infty$, the constitutive relation found between J and E is nonlocal: J satisfies on Ω a propagation equation in the vertical direction with a source term depending on $E(x)$. It is then possible to express J in terms of E_3 as an integral of a suitable kernel with respect to x_3 (see (3.5)). Substituting this expression for J into (1.6) results in a nonlocal homogenized equation for E. This kind of situation has already been observed in the simpler scalar case of the heat equation (see [3]).

– The boundary condition satisfied by J on the bases of the cylinder Ω is not a Dirichlet condition as is expected by common knowledge in antenna theory. By a careful analysis involving the second order derivative of J (in the distributional sense), it is shown (see Lemma 2.8) that the correct boundary condition is of Neumann type, i.e., $\frac{\partial J}{\partial x_3} = 0$. In turn J has an internal trace on $\partial\Omega$ which balances the jump of the vertical component of E.

– In our diffraction problem, none of the fields E_η, H_η converges strongly in $L^2_{\rm loc}$. The defect of strong convergence is directly related to the dissipation of the system by Joule's effect in the fibers.

The paper is organized as follows. In section 2, we fix some notation and prove some preliminary estimates. In section 3 we state and prove the main convergence result. In section 4 we apply our results to the case of fibers with infinite length and present numerical results.

2. Notation and preliminary estimates. We denote by $\{e_1, e_2, e_3\}$ the canonical basis of $\mathbb{R}^3$. We consider a set of e_3-parallel fibers of finite length L disposed in a cylindrical domain $\Omega := \mathcal{D} \times]0, L[$, where $\mathcal{D}$ is a bounded connected open subset of $\mathbb{R}^2$ with smooth boundary. We denote the bases of this cylinder $\mathcal{D}_0 := \mathcal{D} \times \{0\}$ and $\mathcal{D}_L := \mathcal{D} \times \{L\}$.

For every $\eta > 0$, we consider a partition of $\mathcal{D}$ into a set of periodically distributed cells of size η:

$$Y^i_\eta = (\eta\, i_1, \eta\, i_2) + \left]-\frac{\eta}{2}, \frac{\eta}{2}\right[^2, \qquad i = (i_1, i_2) \in I_\eta \subset Z^2,$$

where $I_\eta := \{i \in Z^2;\ Y^i_\eta \subset \mathcal{D}\}$.

Given a small parameter r_η such that $r_\eta \ll \eta$, we define the following:

– $D^i_\eta :=$ two-dimensional (2D) disk centered at $(\eta\, i_1, \eta\, i_2)$ of radius r_η,

– $T^i_\eta := D^i_\eta \times]0, L[,\ T_\eta := \bigcup_{i \in I_\eta} T^i_\eta$.

The fibers are represented by the set of thin parallel cylinders T_η (see Figure 1) and are filled up with a metallic medium. This medium is assumed to be homogeneous with a very high conductivity, which we model through a stiff parameter $\sigma_\eta \to \infty$. Note that, by the definition of I_η, we have $D^i_\eta \subset \mathcal{D}$, so that the fibers do not intersect the lateral part of the boundary of Ω. The subset $\Omega \setminus T_\eta$ as well as the complement of Ω is assumed to have the electromagnetic parameters of the vacuum ε_0 (permittivity) and μ_0 (permeability). Recall that the wave number of a monochromatic wave of angular frequency ω is given by

$$k_0 = \omega\sqrt{\varepsilon_0 \mu_0}. \tag{2.1}$$

In order to simplify the mathematical presentation, the permeability will be kept constant and equal to μ_0 in the metal, and the relative permittivity ε_η is taken as follows:

$$\varepsilon_\eta(x) = 1 \quad \text{if } x \in \Omega \setminus T_\eta, \qquad \varepsilon_\eta(x) = 1 + i\frac{\sigma_\eta}{\varepsilon_0 \omega} \quad \text{if } x \in T_\eta. \tag{2.2}$$

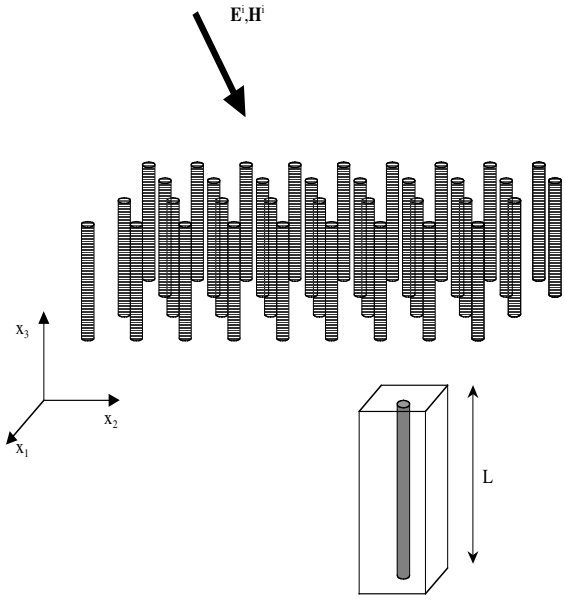


FIG. 1. *Schematic of the diffractive structure.*

Clearly the asymptotic behavior of (ε_η) in $L^1(\Omega)$ is characterized in terms of the filling ratio θ_η by the parameter

$$\kappa := \lim_\eta \theta_\eta \frac{\sigma_\eta}{\omega\varepsilon_0}, \qquad \text{where } \theta_\eta := \pi \frac{r_\eta^2}{\eta^2}. \tag{2.3}$$

If $0 \le \kappa < +\infty$, (ε_η) is bounded in $L^1(\Omega)$ and converges weakly star in the sense of measures to the constant function $1 + i\kappa$. (Notice that the sequence $\{\varepsilon_\eta\}$ is not equi-integrable in $L^1(\Omega)$ unless $\kappa = 0$.)

To simplify we keep constant the mean value of ε_η over Ω so that

$$\sigma_\eta = \frac{\kappa\omega\varepsilon_0}{\theta_\eta}. \tag{2.4}$$

Now we define a rescaled electric field in the fibers by setting

$$F_\eta = \kappa\, E_\eta \frac{1_{\mathbf{T}_\eta}}{\theta_\eta}. \tag{2.5}$$

Then taking into account (2.4), the Maxwell system (1.2) can be rewritten as

$$\begin{cases} \operatorname{curl} E_\eta = i\omega\mu_0 H_\eta, \\ \operatorname{curl} H_\eta = -i\omega\epsilon_0 (E_\eta + i\, F_\eta). \end{cases} \tag{2.6}$$

From now on, we choose a reference ball B of sufficiently large diameter such that $\Omega \subset\subset B$, and we start our analysis with a sequence (E_η, H_η) solving (2.6) such that

$$\sup_\eta \int_B (|E_\eta|^2 + |H_\eta|^2)\, dx \ < +\infty. \tag{2.7}$$

This estimate will be proved in section 3 in the case when (E_η, H_η) is associated with the diffraction of some incident wave. Possibly passing to a subsequence, we may assume that $(E_\eta, H_\eta) \rightharpoonup (E, H)$ weakly in L^2_{loc}. It is rather intuitive that the oscillations of (E_η, H_η) will take place only in a neighborhood of Ω. This fact is confirmed in a rigorous way as in the following claim.

LEMMA 2.1. *Let (E_η, H_η) be a solution of (1.2), and assume that (E_η, H_η) w converges weakly in $L^2(B)$ and that $(E_\eta - E^i_\eta,\ H_\eta - H^i_\eta)$ satisfies the outgoing wave condition (1.3) for a suitable sequence of incident waves (E^i_η, H^i_η) converging uniformly to (E^i, H^i). Then the convergence of (E_η, H_η) holds in $\mathcal{C}^\infty(K)$ for every compact subset $K \subset \mathbb{R}^3 \setminus \overline{\Omega}$. Furthermore, the limit field (E, H) solves the Helmholtz equation in $\mathbb{R}^3 \setminus \overline{\Omega}$, whereas $(E - E^i,\ H - H^i)$ satisfies (1.3).*

Proof. Since $\varepsilon_\eta = 1$ is constant in the complement of Ω, we infer from (1.2) that all the components of the fields E_η and H_η satisfy the Helmholtz equation $\Delta u + k_0^2 u = 0$ on the open set $\mathbb{R}^3 \setminus \overline{\Omega}$. By the properties of hypoelliptic operators (see, for example, [21]), the weak convergence of (E_η, H_η) to (E, H) in $L^2(B)$ implies its uniform convergence, as well as that of all its derivatives, on every compact subset of $B \setminus \overline{\Omega}$. We extend this convergence to all $\mathbb{R}^3 \setminus \overline{\Omega}$ as follows: we consider a sphere $S_R = \{|x| = R\}$, where R is chosen so large that $\Omega \subset\subset \{|x| \le R\} \subset\subset B$. Then we apply to the diffracted field $(E\eta^d, H^d_\eta) := (E_\eta - E^i_\eta,\ H_\eta - H^i_\eta)$ the following Stratton–Chu integral identities, which are in fact equivalent to (1.3) (see [6, 22]):

$$
\begin{aligned}
E^d_\eta(x) &= \int_{|y|=r} \left[i\omega\mu_0\, \Phi(x-y) \left(\frac{y}{|y|} \wedge H^d_\eta \right) + \nabla\Phi(x-y) \wedge \left(\frac{y}{|y|} \wedge E^d_\eta \right) \right] d\sigma, \\
H^d_\eta(x) &= \int_{|y|=r} \left[-i\omega\varepsilon_0\, \Phi(x-y) \left(\frac{y}{|y|} \wedge E^d_\eta \right) + \nabla\Phi(x-y) \wedge \left(\frac{y}{|y|} \wedge H^d_\eta \right) \right] d\sigma,
\end{aligned}
\tag{2.8}
$$

where $\Phi(x) = \frac{e^{ik_0 x}}{4\pi x}$ (recall that $\omega = \frac{k_0}{\sqrt{\varepsilon_0\mu_0}}$ is the angular frequency) and $d\sigma$ denotes the surface integral. Passing to the limit in above relations as $\eta \to 0$, we extend the definition of (E, H) outside B and obtain the convergence of (E_η, H_η) to (E, H) in $\mathcal{C}^\infty(K)$ for every compact subset $K \subset \mathbb{R}^3 \setminus \overline{\Omega}$. Clearly the integral relations (2.8) will hold for the limit diffracted field $(E - E^i,\ H - H^i)$, and therefore it satisfies the Silver–Müller conditions (1.3). ∎

Now we need a precise analysis of the oscillations of the electromagnetic wave in Ω. To that aim, we use a variant of the double scale convergence (see [1]) in a framework of periodic varying measures. This notion (introduced in [4, 5]) will allow us to account simultaneously for the oscillations at scale η and the 3D-1D reduction at the level of each period (recall that $r_\eta \ll \eta$). Although the oscillations of the system described above are expected only in the (x_1, x_2) variables, we will consider periodic test functions in $\mathbb{R}^3$ which oscillate in the directions of the three axes.

We need some further notation: let $Q = [-1/2, 1/2[^3$ and let us denote by $\mathbf{T}$ the 3D torus $\mathbb{R}^3/Q$. In what follows, Q-periodic functions will be systematically identified with functions on $\mathbf{T}$. For every $d \in \mathbb{N}$, we denote by $\mathbb{C}^\infty(\mathbf{T}; \mathbb{R}^d)$ the space of smooth Q-periodic functions from $\mathbb{R}^3$ to $\mathbb{R}^d$, and by $(L^2_\mu(\mathbf{T}))^d$ the space of Q-periodic functions belonging to $L^2_{\mu,\mathrm{loc}}(\mathbb{R}^3; \mathbb{C}^d)$. A Q-periodic measure can be seen as a linear form on the space of continuous periodic functions $C^0(\mathbf{T})$ or equivalently as a Radon measure on $\mathbb{R}^d$ such that $\mu(B + Q) = \mu(B)$ for every Borel subset $B \subset \mathbb{R}^d$. Given such a Q-periodic measure, we indicate by $\mu(x/\eta)$ the measure on $\mathbb{R}^3$ defined by $\langle \mu(x/\eta), \varphi \rangle = \eta^3 \int \varphi(\eta y) d\mu(y)$. Notice that the sequence $\{\mu(\frac{x}{\eta})\}$ converges

weakly star to the Lebesgue measure on $\mathbb{R}^3$ up to the multiplicative constant $c = \mu(Q)$, meaning that $\lim_{\eta\to 0} \langle \mu(x/\eta), \varphi\rangle = c \int_{\mathbb{R}^3} \varphi\, dy$ for every continuous compactly supported test function φ.

DEFINITION 2.2. *Let $\{\mu_\eta\}$ be a sequence of Q-periodic measures converging weakly star to μ. Then a sequence $\{v_\eta\} \subset L^2_{\mu_\eta}(B;\mathbb{R}^d)$ is said to be two-scale convergent to $v_0 \in L^2_{dx\otimes\mu}(B\times\mathbf{T};\mathbb{R}^d)$, and we write $v_\eta \rightharpoonup\!\!\!\!\rightharpoonup v_0$ if, for every test function $\varphi \in \mathcal{C}^\infty_0(B;\mathcal{C}^\infty(\mathbf{T};\mathbb{R}^d))$, the following holds:*

$$\lim_{\eta\to 0} \int_B v_\eta(x)\varphi\left(x,\frac{x}{\eta}\right) d\mu_\eta\left(\frac{x}{\eta}\right) = \iint_{B\times Q} v_0(x,y)\varphi(x,y)\,dx\otimes d\mu(y). \tag{2.9}$$

The sequence $\{v_\eta\}$ is said to be two-scale strongly convergent (denoted $v_\eta \rightarrow\!\!\!\!\rightarrow v_0$) if in addition

$$\operatorname*{limsup}_{\eta\to 0} \int_B |v_\eta(x)|^2 \left(x,\frac{x}{\eta}\right) d\mu_\eta\left(\frac{x}{\eta}\right) = \iint_{B\times Q} |v_0(x,y)|^2\,dx\otimes d\mu(y). \tag{2.10}$$

A straightforward generalization of the results of [4] (see [1] in the case when μ_η is the Lebesgue measure) is summarized in the following result.

PROPOSITION 2.3. *Let $\{v_\eta\}$ be a sequence of Borel functions from B to $\mathbb{R}^d$ such that*

$$\sup_\eta \int_B |v_\eta|^2 d\mu_\eta\left(\frac{x}{\eta}\right) < +\infty.$$

(i) *(compactness) There exists a subsequence $\{v_{\eta_k}\}$ and $v_0 \in L^2_{dx\otimes\mu}(B\times\mathbf{T};\mathbb{R}^d)$ such that $v_{\eta_k} \rightharpoonup\!\!\!\!\rightharpoonup v_0$. In addition, the sequence of measures $\{v_\eta\mu_\eta(\frac{x}{\eta})\}$ is uniformly bounded in total variation, and $v_{\eta_k}\mu_{\eta_k}(x/\eta_k)$ converges weakly star to $v dx$, where $v(x) = \int_Q v_0(x,y)d\mu(y)$.*

(ii) *(lower semicontinuity) If $v_\eta \rightharpoonup\!\!\!\!\rightharpoonup v_0$, then*

$$\liminf_\eta \int_B |v_\eta|^2 d\mu_\eta\left(\frac{x}{\eta}\right) \geq \iint_{B\times Q} |v_0(x,y)|^2\,dx\otimes d\mu(y).$$

(iii) *(weak-strong convergence) Let $\{v_\eta\}, \{w_\eta\}$ such that $v_\eta \rightharpoonup\!\!\!\!\rightharpoonup v_0$ and $w_\eta \rightarrow\!\!\!\!\rightarrow w_0$. Then, for every test function $\varphi \in \mathcal{C}^\infty_0(B;\mathcal{C}^\infty(\mathbf{T}))$, there holds*

$$\lim_{\eta\to 0} \int_B v_\eta \cdot w_\eta\ \varphi\left(x,\frac{x}{\eta}\right) d\mu\left(\frac{x}{\eta}\right) = \iint_{B\times D} v_0(x,y)\cdot w_0(x,y)\,\varphi(x,y)\,dx\otimes d\mu(y).$$

In particular, $v_\eta \cdot w_\eta \mu_\eta(\frac{x}{\eta}) \overset{*}{\rightharpoonup} p(x)dx$, where $p(x) = \int_Q v_0(x,y)\cdot w_0(x,y)\,d\mu(y)$.

We will apply Proposition 2.3 several times, choosing μ_η to be

$$\mu_\eta(dy) = a_\eta(y)dy, \qquad a_\eta(y) := \begin{cases} 1 & \text{if } y \notin S_\eta, \\ \frac{\kappa}{\theta_\eta} & \text{if } y \in S_\eta, \end{cases} \tag{2.11}$$

where S_η denotes the vertical cylinder $S_\eta := \{y_1^2 + y_2^2 \leq \frac{r_\eta^2}{\eta^2},\ |y_3| \leq 1/2\}$. It is easy to check that μ_η converges weakly star to the periodic measure μ defined by

$$\langle \mu, \varphi\rangle := \int_Q \varphi dy + \kappa \int_{S_0} \varphi\, \mathcal{H}^1(dy), \quad \varphi \in \mathcal{C}^0(\mathbf{T}), \tag{2.12}$$

$\mathcal{H}^1(dy)$ being the curvilinear abscissa and $S_0 = \{(0,0,y_3)\ :\ -1/2 \leq y_3 \leq 1/2\}$.

The fact that the rescaled vector field F^η introduced in (2.5) converges weakly to a vertical vector field in $L^2(\Omega)$ is a consequence of the following result.

PROPOSITION 2.4. *Let (E_η, H_η) be a solution of (1.2) such that $(E_\eta, H_\eta) \rightharpoonup (E, H)$ weakly in $L^2_{loc}(B;\mathbb{R}^3)$. Let μ_η be defined in (2.11). Then, we have*

(i) $\sup_\eta \ \frac{1}{\theta_\eta}\int_{T_\eta} |E_\eta|^2\,dx < +\infty$;

(ii) *there exist suitable subsequences of $\{E_\eta\}, \{\varepsilon_\eta E_\eta\}$ (still denoted $\{E_\eta\}, \{\varepsilon_\eta E_\eta\}$) and an element $J \in L^2(\Omega)$ such that*

(2.13)
$$E_\eta \rightharpoonup E_0(x,y), \qquad E_0(x,y) = \begin{cases} E(x) & \text{if } (x,y) \in B \times Q \setminus \Omega \times S_0, \\ \frac{J(x)}{\kappa}\, e_3 & \text{if } (x,y) \in \Omega \times S_0, \end{cases}$$

(2.14)
$$\varepsilon_\eta\, E_\eta \rightharpoonup D_0(x,y), \qquad D_0(x,y) = \begin{cases} E(x) & \text{if } (x,y) \in B \times Q \setminus \Omega \times S_0, \\ i\,\frac{J(x)}{\kappa}\, e_3 & \text{if } (x,y) \in \Omega \times S_0; \end{cases}$$

(iii) *the sequences $\{F_\eta\}$ and $\{\varepsilon_\eta E_\eta\}$ are bounded in $(L^1(B))^3$, and for suitable subsequences one has*

$$F_\eta \overset{*}{\rightharpoonup} J(x)\,1_\Omega(x)\,e_3, \qquad \varepsilon_\eta\, E_\eta \overset{*}{\rightharpoonup} D := E + i\,J(x)\,1_\Omega(x)\,e_3. \tag{2.15}$$

Proof. (i) Recalling (2.4) (where $\kappa < +\infty$), we have to show that $I_\eta := \frac{1}{\theta_\eta}\int_{\mathbf{T}_\eta} |E_\eta|^2$ remains uniformly bounded. Notice that this quantity represents up to a multiplicative factor the energy dissipated by Joule's effect. By using Maxwell equations (1.2) and by integrating by parts ($n(x)$ denotes the exterior normal on ∂B), we derive that

$$\begin{aligned} \int_{\partial B} (E_\eta \wedge \overline{H_\eta}) \cdot n(x) &= \int_B \left(\operatorname{curl} E_\eta \cdot \overline{H_\eta} - \operatorname{curl} \overline{H_\eta} \cdot E_\eta\right) \\ &= i\omega \int_B \left(\mu_0 |H_\eta|^2 - \varepsilon_0 |E_\eta|^2\right) - \frac{\omega \varepsilon_0 \kappa}{\theta_\eta} \int_{\mathbf{T}_\eta} |E_\eta|^2. \end{aligned} \tag{2.16}$$

The left-hand side of (2.16) is well defined, by Lemma 2.1, and converges to $\int_{\partial B} E \wedge \overline{H}$. Therefore by passing to the limit in the real parts of (2.16), we obtain that $\{I_\eta\}$ is uniformly bounded and eventually satisfies the following balance relation:

$$\Re\left(\int_{\partial B} E \wedge \overline{H}\right) = -\omega \varepsilon_0 \kappa \ \lim_{\eta \to 0} I_\eta. \tag{2.17}$$

(ii) and (iii). By (i), we have the uniform upperbound $\sup_\eta \int_B |E_\eta|^2 d\mu_\eta(\frac{x}{\eta}) < +\infty$. Therefore, up to a subsequence, we may assume that $E_\eta \rightharpoonup E_0$, where $E_0(x,y)$ is an element of $L^2_{dx\otimes\mu}(B \times \mathbf{T};\mathbb{R}^3)$ (see Definition 2.2). In addition, since θ_η goes to zero, we have

$$\int_{T_\eta} |E_\eta|^2 dx \to 0. \tag{2.18}$$

From Lemma 2.1 below, we know that the convergence of E_η to E holds with respect to the uniform norm on every compact subset of $B \setminus \overline{\Omega}$; it is then easy to check that $E_0(x,y) = E(x)$ for $x \in B \setminus \Omega$.

We now focus our attention on the set Ω. Due to the particular form of μ given by (2.12), for $x \in \Omega$, we may split E_0 as

$$E_0(x,y) = \tilde{E}_0(x,y) \quad \text{if } y \in Q \setminus S_0, \quad E_0(x,y) = \frac{1}{\kappa} J_0(x,y) \quad \text{if } y \in S_0, \tag{2.19}$$

where for a.e. $x \in \Omega$, $E_0(x,\cdot)$ and $J_0(x,\cdot)$ are respectively elements of $L^2(Q)$ and $L^2(S_0)$.

By (2.18), $E_\eta 1_{T_\eta} \to 0$ strongly in $L^2(\Omega)$, and the weak limit of $\{E_\eta 1_{\Omega\setminus T_\eta}\}$ on Ω coincides with E. Let us apply Proposition 2.3(iii) with $v_\eta = E_\eta$ and $w_\eta := 1_{\Omega\setminus T_\eta}$. We clearly have $w_\eta \rightharpoonup\!\!\!\rightharpoonup w_0(x,y) := 1_{\Omega\times Q\setminus S_0}$. Since $v_\eta w_\eta \mu_\eta(\frac{x}{\eta})$ coincides with $E_\eta 1_{\Omega\setminus T_\eta} dx$, we infer that

$$E(x) = \int_{Q\setminus S_0} E_0(x,y)\, \mu(dy) = \int_Q \tilde{E}_0(x,y)\, dy. \tag{2.20}$$

Now substituting w_η with $1 - w_\eta$, we obtain in a similar way that $E_\eta 1_{T_\eta}\, \mu_\eta(\frac{x}{\eta})$ (which by (2.5) coincides with $F_\eta\, dx$) converges weakly star to $F(x)dx$, where in view of (2.12) and (2.19),

$$F(x) = 1_\Omega(x) \int_{S_0} E_0(x,y)\, \mu(dy) = 1_\Omega(x) \int_{S_0} J_0(x,y)\, \mathcal{H}^1(dy). \tag{2.21}$$

In particular the sequence $\{F_\eta\}$ is bounded in $L^1(\Omega)$.

Eventually we apply Proposition 2.3(iii) with $v_\eta = E_\eta$ and $w_\eta := \frac{\varepsilon_\eta}{a_\eta}$. We find that $w_\eta \rightharpoonup\!\!\!\rightharpoonup w_0$, where $w_0(x,\cdot) := 1_{Q\setminus S_0} + i\, 1_{S_0}$. We deduce that $\varepsilon_\eta\, E_\eta \rightharpoonup\!\!\!\rightharpoonup D_0$, where

$$D_0(x,y) = \tilde{E}_0(x,y) \quad \text{if } y \in Q \setminus S_0, \quad E_0(x,y) = \frac{i}{\kappa}\, J_0(x,y) \quad \text{if } y \in S_0. \tag{2.22}$$

Furthermore $\{\varepsilon_\eta\, E_\eta\}$ is bounded in $L^1(\Omega)$ and converges weakly star to a limit $D(x)$ which, according to (2.20), (2.21), and (2.22), is given by

$$D(x) = \int_Q D_0(x,y)\ \mu(dy) = E(x) + i\, F(x).$$

Summarizing, we have proved the assertions (ii) and (iii) of Proposition 2.3, provided that we can show that $\tilde{E}_0(x,\cdot)$ and $J_0(x,\cdot)$ are constants and that $J_0(x,\cdot)$ is e_3-parallel. This is a consequence of the following claims: for Lebesgue almost all $x \in \Omega$, the vector fields $\tilde{E}_0(x,\cdot)$ and $J_0(x,\cdot)$ defined in (2.19) satisfy as distributions on the torus:

$$\operatorname{curl}_y \tilde{E}_0(x,\cdot) = 0 \quad \text{on } Q, \tag{2.23}$$

$$\operatorname{div}_y \tilde{E}_0(x,\cdot) = 0 \quad \text{on } Q \setminus S_0, \tag{2.24}$$

$$\operatorname{div}_y \big(J_0(x,\cdot)\delta_{S_0}\big) = 0 \quad \text{on } Q \tag{2.25}$$

(where in (2.25) δ_{S_0} stands for the measure associated with the curvilinear integral on S_0).

Indeed the curl-free condition in (2.23) implies that $\tilde{E}_0(x,\cdot)$ as an element of $L^2(\mathbf{T};\mathbb{R}^3)$ reads as $\tilde{E}_0(x,\cdot) = E(x) + \nabla_y w$, where w is a periodic potential in $W^{1,2}(Q)$. Now by the divergence-free condition (2.24), w is harmonic on $Q \setminus S_0$, and thus on the whole cube Q, since the segment S_0 has a vanishing capacity in $W^{1,2}$ (see, for

instance, [13, Theorem 3, p. 154]). Then the periodicity condition yields that w is constant, and thus by (2.20), we have $\tilde{E}_0(x,y) = E(x)$ for almost all $y \in Q$. On the other hand, by applying (2.25) to test functions of the kind $\varphi(y) = \beta(y_3)\Psi(y_1, y_2)$, we deduce that

$$\int_0^1 \left(a\,\beta'(s) J_0(x,s)\cdot e_3 + \beta(s) J_0(x,s)\cdot z\right)\, ds = 0,$$

where $a := \Psi(0,0)$; the horizontal direction $z := \nabla\Psi(0,0)$ and the function $\beta(s)$ as well can be chosen arbitrarily. It is then straightforward to infer that $J_0(x,\cdot)$ is e_3-parallel and constant along S_0, and thus of the form given in (2.13) with a suitable $J \in L^2(\Omega)$.

Proof of claim (2.23). Let us apply again Proposition 2.3(iii) with $v_\eta = E_\eta$ and $w_\eta := 1_{\Omega\setminus T_\eta} + \frac{\theta_\eta}{\kappa} 1_{T_\eta}$. As $\theta_\eta \to 0$, we have $w_\eta \twoheadrightarrow w_0(x,y) := 1_{\Omega\times Q\setminus S_0}$. Therefore, for every test function $\Psi \in \mathcal{C}_0^\infty(\Omega; \mathcal{C}^\infty(\mathbf{T};\mathbb{R}^3))$, we have in view of (2.19)

$$\begin{aligned}\lim_\eta \int_\Omega \Psi\left(x, \frac{x}{\eta}\right)\cdot E_\eta\, dx &= \lim_\eta \int_\Omega \Psi\left(x,\frac{x}{\eta}\right)\cdot v_\eta(x)\; w_\eta(x)\,\mu_\eta\left(\frac{x}{\eta}\right)\\ &= \iint_{\Omega\times(Q\setminus S_0)} \Psi(x,y)\cdot V_0(x,y)\, dx\otimes d\mu(y)\\ &= \iint_{\Omega\times Q} \Psi(x,y)\cdot \tilde{E}_0(x,y)\, dx\, dy.\end{aligned} \tag{2.26}$$

By the first Maxwell equation in (1.2) and (2.7), the sequence $\{\operatorname{curl} E_\eta\}$ is bounded in $L^2(\Omega)$ and $\eta \operatorname{curl} E_\eta \to 0$ in $L^2(\Omega)$. Taking $\Psi = \operatorname{curl}_y \varphi$ in (2.26), where $\varphi \in \mathcal{C}_0^\infty(\Omega; \mathcal{C}^\infty(\mathbf{T};\mathbb{R}^3))$, and integrating by parts, we deduce that

$$\begin{aligned} 0 = \lim_\eta \int_\Omega \eta\, E_\eta\cdot \operatorname{curl}_y \varphi\left(x,\frac{x}{\eta}\right) &= \lim_\eta \int_\Omega E_\eta\cdot \operatorname{curl}\left(\eta\,\varphi\left(x,\frac{x}{\eta}\right)\right)\, dx\\ &= \lim_\eta \int_O E_\eta\cdot (\operatorname{curl}_y\varphi)\left(x,\frac{x}{\eta}\right)\, dx\\ &= \iint_{\Omega\times Q} \tilde{E}_0(x,y)\cdot \operatorname{curl}_y\varphi(x,y)\, dx\, dy.\end{aligned} \tag{2.27}$$

The conclusion follows from (2.27) by choosing a test function φ of the kind $\varphi(x,y) = \beta(x)\varphi_0(y)$, where $\beta \in \mathcal{C}_0^\infty(\Omega)$ and $\varphi_0 \in \mathcal{C}_0^\infty(Q;\mathbb{R}^3)$, and by localizing with respect to x.

Proof of claim (2.24). By the second Maxwell equation in (1.2), the vector field $\varepsilon_\eta E_\eta$ is divergence-free. Thus by taking $\Psi = \nabla_y\varphi$ as a test function, where $\varphi \in \mathcal{C}_0^\infty(\Omega; \mathcal{C}^\infty(\mathbf{T}))$, we obtain (taking into account (2.22))

$$\begin{aligned} 0 &= \lim_\eta\ \eta \int_\Omega \nabla\left(\varphi\left(x,\frac{x}{\eta}\right)\right)\cdot \varepsilon_\eta E_\eta\, dx = \lim_\eta \int_\Omega \nabla_y\varphi\left(x,\frac{x}{\eta}\right)\cdot \varepsilon_\eta E_\eta\, dx\\ &= \iint_{\Omega\times Q} \nabla_y\varphi(x,y)\cdot D_0(x,y)\, dx\otimes d\mu(y)\\ &= \iint_{\Omega\times Q} \nabla_y\varphi(x,y)\cdot \tilde{E}_0(x,y)\, dx\, dy + \iint_{\Omega\times S_0} \nabla_y\varphi(x,y)\cdot J_0(x,y)\; dx\otimes \mathcal{H}^1(dy).\end{aligned} \tag{2.28}$$

Choosing φ of the kind $\varphi(x,y) = \beta(x)\varphi_0(y)$, where $\beta \in \mathcal{C}_0^\infty(\Omega)$ and φ_0 is smooth compactly supported in $Q \setminus S_0$, we easily recover the condition (2.24).

Proof of claim (2.25). From (2.23), (2.24), we have already deduced that $\tilde{E}_0(x,\cdot)$ is constant. Thus the first integral in the last line of (2.28) vanishes, and we are reduced to

$$0 = \int\!\!\int_{\Omega\times S_0} \nabla_y \varphi(x,y)\cdot J_0(x,y)\; dx \otimes \mathcal{H}^1(dy),$$

for every $\varphi \in \mathcal{C}_0^\infty(\Omega;\mathcal{C}^\infty(\mathbf{T}))$. The claim follows by localization. ∎

The aim of the end of this section is to recover a relation between J and E. It turns out that this relation is highly dependent on the parameter γ defined in (1.4). From now on we assume that

$$0 \;<\; \gamma := \lim_\eta (\eta^2 |\log r_\eta|)^{-1} \;< +\infty. \tag{2.29}$$

The borderline cases $\gamma = 0$ and $\gamma = +\infty$ will be discussed in section 3 (see Remark 3.4). We introduce a kind of capacitary potential w_η, which renders the transition between F_η on the fibers and the electric field E_η. Let $Y_\eta = (-\eta/2,\eta/2)^2$ and θ_η be the filling ratio parameter defined in (2.3). Then there exists a unique solution w_η in $W_0^{1,2}(Y_\eta)$ of

$$\frac{1}{2\pi\gamma}\Delta w_\eta = \begin{cases} -\theta_\eta^{-1} & \text{if } |x| \le r_\eta, \\ \left(\dfrac{\pi}{4} - \theta_\eta\right)^{-1} & \text{if } r_\eta < |x| \le \dfrac{\eta}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.30}$$

From now on we consider the η-periodization of w_η on the whole plane, and we consider $w_\eta = w_\eta(x_1,x_2)$ as a function on $\mathbb{R}^3$. The following lemma shows that w_η is very close to 1 (but nonconstant) on the fibers and also that it converges weakly to 0 in $W^{1,2}_{loc}$.

LEMMA 2.5. *The function $w_\eta(x_1,x_2)$ satisfies the following conditions:*

$$w_\eta \rightharpoonup 0 \;\; \textit{in } W^{1,2}_{loc}, \qquad \sup_{|x|\le r_\eta} |1 - w_\eta| \to 0, \quad |\nabla w_\eta| \le C\frac{r_\eta}{\theta_\eta}, \tag{2.31}$$

$$w_\eta \ge c_\eta \quad \textit{on } T_\eta, \qquad w_\eta \le c_\eta \quad \textit{on } \Omega \setminus T_\eta, \tag{2.32}$$

where C and c_η are suitable constants such that $C > 0$ and $c_\eta \to 1$.

Proof. w_η can be computed explicitly as a radial function. For every $x \in [-\eta/2,\eta/2[^3$, we set $w_\eta(x) = f_\eta(|x|)$, where

$$f_\eta(r) = \begin{cases} \dfrac{-\pi\gamma}{2\theta_\eta} r^2 + B_\eta & \text{if } r < r_\eta, \\ \dfrac{2\gamma}{1 - \frac{4\theta_\eta}{\pi}}\left(r^2 - \dfrac{\eta^2}{4}\right) - C_\eta \log\dfrac{2r}{\eta} & \text{if } r_\eta < r < \frac{\eta}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.33}$$

Here the constants B_η and C_η are chosen in such a way that f_η and f_η' are continuous at $r = r_\eta$ and $r = \eta/2$ (in particular, $f_\eta(\eta/2) = f_\eta'(\eta/2) = 0$); we find

$$C_\eta := \frac{\gamma\eta^2}{1 - \frac{4\theta_\eta}{\pi}}, \quad B_\eta := -\frac{\gamma}{1 - \frac{4\theta_\eta}{\pi}}\eta^2 \log r_\eta.$$

Set $c_\eta = f_\eta(r_\eta)$. Thanks to (2.29) we clearly have that $B_\eta \to 1$, $C_\eta \to 0$, and $c_\eta \to 1$. Then it is easy to check that w_η satisfies all the conditions (2.30), (2.31), and (2.32). ∎

A consequence of the positivity of γ (see (2.29)) is the following important estimate.

LEMMA 2.6. *There exists a constant $C = C(\gamma) > 0$ such that, for every $\varphi \in H^1(\Omega)$,*

$$\int_{T_\eta} |\varphi|^2 \;\le\; C\,\theta_\eta \int_\Omega \left(|\varphi|^2 + |\nabla\varphi|^2\right)\,dx. \tag{2.34}$$

Proof. By using Fubini's theorem, it is enough to prove the 2D version of (2.34); that is, setting $D_\eta := \bigcup_{i\in I_\eta} D^i_\eta$,

$$\int_{D_\eta} |\varphi|^2 \;\le\; C\,\theta_\eta \int_{\mathcal{D}} \left(|\varphi|^2 + |\nabla\varphi|^2\right)\,dx \quad \forall\varphi \in H^1(\mathcal{D}). \tag{2.35}$$

Let us define for every $\delta > 0$ (here $Y := (-1/2, 1/2)^2$)

$$A(\delta) := \inf\left\{ \int_Y |\nabla\varphi|^2 \quad : \quad \int_Y \varphi\,dy = 0, \quad \int_{|y|<\delta} |\varphi|^2 = 1 \right\}. \tag{2.36}$$

It turns out that $A(\delta)$ tends increasingly to infinity as $\delta \to 0$. More precisely we can prove the following estimate:

$$\liminf_{\delta\to 0}\; A(\delta)\,\delta^2 |\log\delta| \;\ge\; \frac{2}{\gamma}. \tag{2.37}$$

Let now fix $\varphi \in W^{1,2}(\mathcal{D})$. We consider the piecewise constant function φ_η on D defined by setting $\varphi_\eta = [\varphi]^i_\eta$ on each Y^i_η, where $[\varphi]^i_\eta$ denotes the mean value of φ. After an easy rescaling of (2.36), we obtain

$$\int_{Y^i_\eta} |\nabla\varphi|^2\,dx \;\ge\; \frac{1}{\eta^2}\,A\left(\frac{r_\eta}{\eta}\right) \int_{D^i_\eta} |\varphi - \varphi_\eta|^2\,dx.$$

After summing the previous inequality over $i \in I_\eta$, we derive

$$\begin{aligned}
\int_{D\eta} |\varphi|^2\,dx &\le 2\int_{D\eta} |\varphi_\eta|^2\,dx + 2\int_{D\eta} |\varphi - \varphi_\eta|^2\,dx \\
&\le 2\sum_i ([\varphi]^i_\eta)^2 |D^i_\eta| + 2\frac{\eta^2}{A(\frac{r_\eta}{\eta})} \sum_i \int_{Y^i_\eta} |\nabla\varphi|^2\,dx \\
&\le 2\,\theta_\eta \int_{\mathcal{D}} |\varphi|^2\,dx + 2\frac{\eta^2}{A(\frac{r_\eta}{\eta})} \int_{\mathcal{D}} |\nabla\varphi|^2\,dx.
\end{aligned}$$

Recalling (2.3), the estimate (2.35) follows by taking into account (2.37). Thus the proof of Lemma 2.6 is achieved if we show (2.37).

Proof of claim (2.37). We consider first a competitor $\varphi = v(r,\theta)$ in (2.36) such that v is compactly supported in the ball $\{r < 1/2\}$. For almost all $\theta \in [0, 2\pi)$, $v(\cdot,\theta)$ belongs to $W^{1,2}_{\rm loc}(0, 1/2)$ and vanishes at $r = 1/2$. Therefore, for every $\rho \in (0, \delta)$, we have

$$\int_0^{1/2} \left|\frac{\partial v}{\partial r}\right|^2 r dr \ge\; \inf\left\{ \int_\rho^{1/2} r\,{w'}^2\,dr \;:\; w(\rho) = v(\rho,\theta),\; w\left(\frac{1}{2}\right) = 0 \right\} = \frac{|v(\rho,\theta)|^2}{|\log(2\rho)|}.$$

Multiplying by $\rho|\log(2\rho)|$ and integrating with respect to (ρ,θ) over $(0,\delta)\times(0,2\pi)$, we derive the following estimate for $\varphi = v(r,\theta)$:

$$\int_Y |\nabla\varphi|^2\,dy \geq \frac{\int_{|y|<\delta}|\varphi|^2\,dy}{\int_0^\delta \rho|\log(2\rho)|\,d\rho} = \frac{2-\circ(\delta)}{\delta^2|\log(\delta)|}\int_{|y|<\delta}|\varphi|^2\,dy. \tag{2.38}$$

Then we conclude (2.37) by contradiction: assume that there exists a sequence $\{\varphi_\delta\}$ in $H^1(Y)$ such that

$$\int_Y |\nabla\varphi_\delta|^2\,dy \ \to 0, \quad \int_Y \varphi_\delta\,dy = 0, \quad \frac{1}{\delta^2|\log(\delta)|}\int_{|y|<\delta}|\varphi_\delta|^2\,dy = 1.$$

Clearly φ_δ converges strongly to 0 in $H^1(Y)$, and for every smooth cut-off function $\alpha\in C^\infty(Y;[0,1])$ compactly supported in $\{|y|<1/2\}$ such that $\alpha=1$ in $\{|y|<1/4\}$, the truncated function $\tilde\varphi_\delta := \alpha\varphi_\delta$ satisfies

$$\int_Y |\nabla\tilde\varphi_\delta|^2\,dy \ \to 0, \quad \frac{1}{\delta^2|\log(\delta)|}\int_{|y|<\delta}|\tilde\varphi_\delta|^2\,dy \to 1.$$

This is impossible, since $\tilde\varphi_\delta$ needs to verify (2.38). ∎

The sequences $\{E_{3,\eta}\Delta w_\eta\}$, $\{E_{3,\eta}\partial_{x_\alpha} w_\eta\}$, and $\{E_\eta\cdot\nabla w_\eta\}$ are bounded in $L^1(B)$. The next lemma is related to their weak star convergence in the sense of measures on B.

LEMMA 2.7. *For every $\varphi\in\mathcal{C}_0^\infty(B)$, there hold*

$$\lim_{\eta\to0}\int_B E_{3,\eta}\,\Delta w_\eta\,\varphi\,dx = 2\pi\gamma\int_\Omega\left(E_3-\frac{J}{\kappa}\right)\varphi\,dx, \tag{2.39}$$

$$\lim_{\eta\to0}\int_B E_{3,\eta}\,\frac{\partial w_\eta}{\partial x_\alpha}\,\varphi\,dx = 0 \qquad \textit{for } \alpha\in\{1,2\}, \tag{2.40}$$

$$\lim_{\eta\to0}\int_B E_\eta\cdot\nabla w_\eta\,\varphi\,dx = -i\int_\Omega J\frac{\partial\varphi}{\partial x_3}\,dx + \int_{\partial\Omega}\varphi\,d\nu, \tag{2.41}$$

where ν is a suitable bounded Radon measure on $\partial\Omega$.

Here we point out that the third limit, which is achieved a priori for a subsequence, will be found later to be unique (see the explicit expression of ν given in (2.51)). Therefore the whole sequence actually converges.

Proof. (i) We write (2.30) as $\Delta w_\eta = 2\pi\gamma\ g_\eta(x)\,a_\eta(\frac{x}{\eta})$, where a_η is defined in (2.11) and $g_\eta = g_\eta(x_1,x_2)$ is the η-periodization of

$$g_\eta(x) = \begin{cases} -\kappa^{-1} & \text{if } |x'|<r_\eta,\\ (\frac{\pi}{4}-\theta_\eta)^{-1} & \text{if } r_\eta<|x'|<\frac{\eta}{2},\\ 0 & \text{otherwise}\end{cases} \qquad (x'=(x_1,x_2)).$$

It is easy to check that $g_\eta \twoheadrightarrow g_0$, where

$$g_0(x,y) := \begin{cases} -\frac{1}{\kappa} & \text{if } y\in S_0,\\ \frac{4}{\pi} & \text{if } 0\leq|y|<\frac12,\\ 0 & \text{otherwise.}\end{cases}$$

By applying Proposition 2.3(iii) to the pair $(E_\eta, g_\eta\, e_3)$ and recalling (2.13), we obtain (2.39) as follows:

$$\begin{aligned}\lim_{\eta\to 0}\int_B E_{3,\eta}\ \Delta w_\eta\,\varphi\,dx &= 2\pi\gamma\ \lim_{\eta\to 0}\int_B g_\eta\, e_3\cdot E_\eta\,\varphi\ \mu_\eta\left(\frac{x}{\eta}\right)\\ &= 2\pi\gamma\ \iint_{B\times Q} g_0(x,y)\, e_3\cdot E_0(x,y)\,\varphi(x)\ dx\otimes d\mu(y)\\ &= 2\pi\gamma\ \int_\Omega\left(E_3(x)-\frac{J(x)}{\kappa}\right)\varphi(x)\,dx.\end{aligned}$$

(ii) Let us prove (2.40). Given $\psi=(\psi_1,\psi_2,0)$ a transverse test function in $\mathcal{C}_0^\infty(B)$, we multiply the first equation of (2.6) by Ψw_η and integrate by parts on B. We obtain

$$\begin{aligned}i\omega\mu_0\int_B H_\eta\cdot(\Psi w_\eta)\,dx &= \int_B \operatorname{curl} E_\eta\cdot(\Psi w_\eta)\,dx\\ &= \int_B E_\eta\cdot\operatorname{curl}(\Psi w_\eta)\,dx\\ &= \int_B (E_\eta\cdot\operatorname{curl}\Psi) w_\eta\,dx+\int_B E_\eta\cdot(\nabla w_\eta\wedge\Psi)\,dx.\end{aligned}$$

By the strong convergence of w_η to 0 in $L^2(B)$, we infer that

$$0=\lim_{\eta\to 0}\int_B E_\eta\cdot(\nabla w_\eta\wedge\Psi)\,dx=\lim_{\eta\to 0}\int_B E_\eta\cdot\left(\Psi_2\frac{\partial w_\eta}{\partial x_1}-\Psi_1\frac{\partial w_\eta}{\partial x_2}\right)dx,$$

and hence the conclusion by taking $\Psi=(\varphi,0,0)$ or $\Psi=(0,\varphi,0)$.

(iii) We consider the sequence $\{E_\eta\cdot\nabla w_\eta\}$. It is bounded in $L^1(B)$ and possibly passing to a subsequence; we may assume that it converges weakly star to some bounded measure m on B. By the uniform convergence of E_η on every compact subset of $B\setminus\overline{\Omega}$ obtained in Lemma 2.1 and the weak convergence of ∇w_η to 0 in $L^2(B)$, we find that m vanishes on $B\setminus\overline{\Omega}$ and therefore can be written in the form $m=m_\Omega+\nu$, where m_Ω and ν are Radon measures supported respectively in Ω and in $\partial\Omega$. Then proving (2.41) reduces to showing that, for every test function φ compactly supported in Ω, there holds

$$\langle m_\Omega,\varphi\rangle=\lim_{\eta\to 0}\int_\Omega E_\eta\cdot\nabla w_\eta\,\varphi\,dx=-i\int_\Omega J\frac{\partial\varphi}{\partial x_3}\,dx. \tag{2.42}$$

Let $\overline{w}_\eta$ denote the truncature of w_η defined by setting $\overline{w}_\eta:=\inf\{w_\eta,c_\eta\}$. Then by (2.31) we have $\overline{w}_\eta=c_\eta$ on T_η and $\nabla\overline{w}_\eta=1_{\Omega\setminus T_\eta}\nabla w_\eta$. Noticing that $\varepsilon_\eta E_\eta=E_\eta$ on $\Omega\setminus T_\eta$, we infer that

$$\begin{aligned}\int_\Omega E_\eta\cdot\nabla w_\eta\,\varphi\,dx\ &=\ A_\eta+B_\eta,\\ A_\eta&:=\int_\Omega\varepsilon_\eta E_\eta\cdot\nabla\overline{w}_\eta\,\varphi\,dx,\\ B_\eta&:=\int_{T_\eta} E_\eta\cdot\nabla w_\eta\,\varphi\,dx.\end{aligned} \tag{2.43}$$

Taking into account that $\varepsilon_\eta E_\eta$ is divergence-free, we may rewrite A_η as

$$\begin{aligned} A_\eta &= -\int_\Omega \overline{w}_\eta \varepsilon_\eta E_\eta \cdot \nabla\varphi \, dx \\ &= -\int_{\Omega\setminus T_\eta} w_\eta E_\eta \cdot \nabla\varphi \, dx - c_\eta \int_{T_\eta} \varepsilon_\eta E_\eta \cdot \nabla\varphi \, dx. \end{aligned}$$

By (2.15), we know that $\varepsilon_\eta E_\eta 1_{T_\eta} = i\kappa F_\eta$ does converge weakly star to $i\,J\,e_3$. Since $w_\eta \to 0$ in $L^2(\Omega)$, $c_\eta \to 1$, we deduce that

$$\lim_{\eta\to 0} A_\eta = -i \int_\Omega J(x)\, \frac{\partial\varphi}{\partial x_3}\, dx. \tag{2.44}$$

On the other hand, by the last inequality in (2.31), we have

$$|B_\eta| \;\le\; C\frac{r_\eta}{\theta_\eta} \int_{T_\eta} |E_\eta|\, dx = C r_\eta \int_\Omega |F_\eta|\, dx. \tag{2.45}$$

Since F_η is bounded in $L^1(\Omega)$, the claim (2.42) follows from (2.43), (2.44), and (2.45). □

LEMMA 2.8. *Under the assumptions of Proposition* 2.4*, the current density J defined in* (2.13) *belongs to $L^2(\mathcal{D}; W^{2,2}(0,L))$ and satisfies the boundary value problem*

$$\begin{cases} \frac{\partial^2 J}{\partial x_3^2} + (k_0^2 + \frac{2i\pi\gamma}{\kappa})\; J = 2i\,\pi\,\gamma\,E_3 & \text{on} \quad \Omega, \\ \qquad\qquad \frac{\partial J}{\partial x_3} = 0 & \text{on} \quad \mathcal{D}_0 \cup \mathcal{D}_L. \end{cases} \tag{2.46}$$

Proof. In order to recover boundary conditions over $\partial\Omega$, we consider a test function in $\varphi \in \mathcal{C}_0^\infty(B)$ and integrate over all $\mathbb{R}^3$. We multiply the second equation of (2.6) by $w_\eta\,\varphi(x)\,e_3$ and integrate by parts to obtain

$$\begin{aligned} -i\,\omega\,\varepsilon_0 \int (E_{3,\eta} + i\,F_{3,\eta})\, w_\eta \varphi \, dx &= \int \operatorname{curl} H_\eta \cdot (w_\eta \varphi\, e_3)\, dx \\ &= \int H_\eta \cdot [(\nabla w_\eta \wedge e_3)\varphi + (\nabla\varphi \wedge \varepsilon_3) w_\eta]\; dx. \end{aligned} \tag{2.47}$$

Recalling that, by Lemma 2.5, w_η does converge to 0 in L^2_{loc} and is uniformly close to 1 on the subset T_η, we can identify the limit of the integral in the left-hand member by applying Proposition 2.4(iii). We are led to

$$\omega\varepsilon_0 \int_\Omega J\,\varphi dx = \lim_{\eta\to 0} I_\eta, \qquad \text{where} \quad I_\eta := \int H_\eta \cdot (\nabla w_\eta \wedge \overrightarrow{e_3})\; \varphi\; dx. \tag{2.48}$$

Now we express I_η in terms of E_η by using the first equation in (2.6). By integrating once more by parts, we obtain

$$\begin{aligned} I_\eta &= \frac{1}{i\omega\mu_0} \int_\Omega E_\eta \cdot \operatorname{curl}\left[\varphi\,(\nabla w_\eta \wedge e_3)\right]\; dx \\ &= \frac{1}{i\omega\mu_0} \int \left[E_\eta \cdot (\nabla\varphi \wedge (\nabla w_\eta \wedge e_3)) + \varphi \operatorname{curl}(\nabla w_\eta \wedge\, e_3)\right]. \end{aligned}$$

Recalling that $w_\eta = w_\eta(x_1, x_2)$, we compute $\operatorname{curl}(\nabla w_\eta \wedge e_3) = -\Delta w_\eta\, e_3$. Thus we may write I_η as the sum of three terms:

(2.49)
$$I_\eta = \frac{1}{i\omega\mu_0}\left[\int E_\eta \cdot \nabla w_\eta \left(\frac{\partial\varphi}{\partial x_3}\right) dx - \int E_{3,\eta}\, \Delta w_\eta\, \varphi\, dx - \int E_{3,\eta}\, \nabla w_\eta \cdot \nabla\varphi\, dx\right].$$

The limits of the three integrals above are deduced from (2.41), (2.39), and (2.40), respectively. Therefore, by (2.48) and (2.49), we are led to

$$k_0^2 \int_\Omega J\,\varphi\, dx = -\int_\Omega J\,\frac{\partial^2\varphi}{\partial x_3^2}\, dx - i\int_{\partial\Omega} J\,\frac{\partial\varphi}{\partial x_3}\, d\nu + 2i\pi\gamma \int_\Omega \left(E_3 - \frac{J}{\kappa}\right)\varphi\, dx, \tag{2.50}$$

where we have multiplied by $\omega\mu_0$ and used the relation $k_0^2 = \varepsilon_0\mu_0\omega^2$. Choosing first φ compactly supported in Ω leads to the propagation equation in (2.46), which is satisfied in the distributional sense in Ω. Since E_3 belongs to $L^2(\Omega)$, this implies that J is an element of $L^2(\mathcal{D}; W^{2,2}(0,L))$. Therefore J and $\frac{\partial J}{dx_3}$ have traces on $\mathcal{D}_0 \cup \mathcal{D}_L$. Let us denote by j_L and j_0 respectively the traces of J on $\mathcal{D}_L$ and $\mathcal{D}_0$. Then by the integration by parts of the volume integrals in (2.50) and by taking into account (2.46), we are led to

$$-\int_{D_L}\left(J\frac{\partial\varphi}{\partial x_3} - \frac{\partial J}{\partial x_3}\right) d\sigma + \int_{D_0}\left(J\frac{\partial\varphi}{\partial x_3} - \frac{\partial J}{\partial x_3}\varphi\right) d\sigma - i\int_{\partial\Omega}\frac{\partial\varphi}{\partial x_3}\, d\nu = 0.$$

Since this relation holds for every $\varphi \in \mathcal{C}_0^\infty(B)$, we derive that

$$\frac{\partial J}{\partial x_3} = 0 \quad \text{on} \quad \mathcal{D}_0 \cup \mathcal{D}_L \qquad \text{and} \qquad \nu = i\,(j_L\,\Sigma_L - j_0\,\Sigma_0), \tag{2.51}$$

with Σ_0, Σ_L the surface integrals on $\mathcal{D}_0$ and $\mathcal{D}_L$, respectively. □

3. The main homogenization result. Thanks to the results of section 2, we are now in position to describe asymptotically the diffraction by the fibered structure depicted in Figure 1. Our main convergence theorem states that passing to the limit as the period η tends to zero leads to a coupled system of partial differential equations with suitable transmission conditions on the boundary of the structure.

In the following, (E^i, H^i) denotes a given incident electromagnetic wave. It satisfies the harmonic Maxwell system in the vacuum (i.e., $\operatorname{curl} E^i = i\,\omega\mu_0\, H^i$, $\operatorname{curl} H^i = -i\,\omega\varepsilon_0\, E^i$ in all $\mathbb{R}^3$). Recall that he total electromagnetic field (E_η, H_η) is then completely determined as the unique solution of (1.2) such that the diffracted field $(E_\eta^d, H_\eta^d) := (E_\eta - E^i,\, H_\eta - H^i)$ satisfies the outgoing wave condition at infinity (1.3).

THEOREM 3.1. *Let γ and κ be given by* (1.4) *and* (2.3), *respectively. Then, as $\eta \to 0$, the following convergences hold:*

$$(E_\eta, H_\eta) \rightharpoonup (E, H) \quad \textit{in } L^2_{loc} \quad , \quad F_\eta := \kappa\, E_\eta \frac{1_{T_\eta}}{\theta_\eta} \overset{*}{\rightharpoonup} J(x)\overrightarrow{e_3} \quad \textit{in } (L^1(\Omega))^3,$$

where E, H, J are the unique solutions (in the distributional sense) of the system

(3.1)
$$\begin{cases} \operatorname{curl} E &= i\omega\mu_0 H & \text{on } \mathbb{R}^3, \\ \operatorname{curl} H &= -i\omega\varepsilon_0(E + i\,J\,1_\Omega \overrightarrow{e_3}) & \text{on } \mathbb{R}^3, \\ \frac{\partial^2 J}{\partial x_3^2} + (k_0^2 + \frac{2i\pi\gamma}{\kappa})\,J &= 2i\pi\gamma\,E_3 & \text{on } \Omega, \\ \frac{\partial J}{\partial x_3} &= 0 & \text{on } \mathcal{D}_0 \cup \mathcal{D}_L, \\ (E - E^i, H - H^i) & \text{ satisfies the outgoing wave condition.} \end{cases}$$

Remark 3.2. In view of the improved regularity of (E, H, J) obtained in Lemma 3.5 below, it is possible to split (3.1) into two systems of equations (one on Ω, the other on $\mathbb{R}^3 \setminus \Omega$), to which we add suitable transmission conditions along the boundary of Ω:

(3.2)
$$\begin{cases} \operatorname{curl} E &= i\omega\mu_0 H \\ \operatorname{curl} H &= -i\omega\varepsilon_0(E + i\,J\,e_3) \qquad \text{on } \Omega, \\ \frac{\partial^2 J}{\partial x_3^2} + (k_0^2 + \frac{2i\pi\gamma}{\kappa})\,J &= 2i\pi\gamma\,E_3 \end{cases}$$

(3.3)
$$\begin{cases} \Delta E + k_0^2 E = \Delta H + k_0^2 H = 0 \qquad \text{on } \mathbb{R}^3 \setminus \Omega, \\ (E - E^i, H - H^i) \text{ satisfies the outgoing wave condition,} \end{cases}$$

(3.4)
$$\begin{cases} H & \text{continuous across} & \partial\Omega, \\ E & \text{continuous across} & \partial\mathcal{D} \times (0, L), \\ \frac{\partial J}{\partial x_3} = 0 & \text{on} & \mathcal{D}_0 \cup \mathcal{D}_L, \\ E_3 + i\,J\,1_\Omega\,e_3 & \text{continuous across} & \mathcal{D}_0 \cup \mathcal{D}_L. \end{cases}$$

Notice that, in (3.4), the continuity has to be understood in the sense of traces in the related Sobolev spaces (we refer to Duvaut and Lions [12] for further details). The last condition is a concise rewriting of the continuity of the normal trace of the divergence-free vector field $D := E + i\,J\,1_\Omega\,e_3$.

Remark 3.3. By using the third equation in (3.1) together with the Neumann condition, it is possible to express J in terms of E_3 as follows: let $K(s, \cdot)$ be, for every $s \in (0, L)$, the solution of the 1D problem

$$\begin{cases} y'' + (k_0^2 + \frac{2i\pi\gamma}{\kappa})\,y = \frac{2i\pi\gamma}{\kappa}\,\delta_s, \\ y'(0) = y'(L) \quad = 0. \end{cases}$$

Then there holds, for almost every $(x_1, x_2) \in \mathcal{D}$,

(3.5)
$$J(x_1, x_2, x_3) = \int_0^L K(x_3, s)\,E_3(s)\,ds, \quad x_3 \in (0, L).$$

Substituting this expression into the second equation in (3.1) leads to a nonlocal constitutive relation between E and the displacement field $D := E + i\,J\,1_\Omega\,e_3$.

Remark 3.4. The limit model is described by the two parameters γ (*capacity*) and κ (*stiffness*). It is clear from (1.4) and (2.3) that these constants can be tuned at will by playing with the size r_η of the fibers and the conductivity coefficient σ_η. In other words, *any problem of the kind* (3.1) *can be obtained as a limit as* $\eta \to 0$ *of classical diffraction problems.*

To complete the picture, some comments related to the extreme values of κ and γ are in order. Note that the conclusions we are drawing below are based on heuristical arguments.

(a) *Cases* $\kappa = 0$ *and* $\kappa = \infty$. Throughout the paper we assumed (1.1) with $\kappa > 0$. If we substitute κ with a sequence $\kappa_\eta \to 0$, it is clear that the fibers disappear in the limit process so that the structure becomes transparent. In other words, the conclusions of Theorem 3.1 hold with $J = 0$. In contrast, when $\kappa = \infty$ and if k_0^2 is not an eigenvalue of the Neumann problem on $(0, L)$ (i.e., if $k_0 \notin \{\frac{\pi n}{L} : n \in \mathbb{N}\}$), we find a nonvanishing limit current density J, and in (3.1) the propagation equation satisfied by J becomes $\frac{\partial^2 J}{\partial x_3^2} + k_0^2\, J = 2i\pi\gamma E_3$. Notice also that the dissipation of energy by Joule's effect given by (3.6) vanishes. This case will occur either when considering a sequence $\kappa_\eta \to \infty$ in (1.1) or when starting, for every η, with infinitely conducting fibers. In the latter situation, (E_η, H_η) vanishes in T_η as well as the tangential component of the trace of the electric field E_η, whereas the jump of the tangential component of the magnetic field H_η induces a microscopic current on ∂T_η which, as $\eta \to 0$, is responsible for the presence of J.

(b) *Cases* $\gamma = 0$ *and* $\gamma = \infty$. For $\gamma = 0$, the source term in the propagation equation (2.46) disappears. Thus J vanishes, and we are led to the same conclusions as for $\kappa = 0$. If $\gamma = +\infty$, the strong interaction between fibers and matrix forces the equality $J = \kappa\, E_3$. As a consequence, the relation between E and D on Ω remains local with an effective tensor ε^{eff} given by $\varepsilon^{\text{eff}} = \operatorname{diag}\{1, 1, 1 + i\,\kappa\}$. However, E_3 inherits the homogeneous Neumann boundary condition of J, and a jump of E_3 will still occur on $\mathcal{D}_0 \cup \mathcal{D}_L$.

The end of this section is devoted to the proof of Theorem 3.1, which crucially makes use of the uniqueness of the limit system.

LEMMA 3.5.

(i) *The solution of* (E, H, J) *of* (3.1)*, if it exists, is unique and satisfies* $E \in W^{1,2}_{\text{loc}}(\mathbb{R}^3 \setminus \mathcal{D}_0 \cup \mathcal{D}_L)$*,* $H \in W^{1,2}_{\text{loc}}(\mathbb{R}^3)$*, and* $J \in L^2(\mathcal{D}; W^{2,2}(0, L))$.

(ii) *Furthermore, denoting by* $\Re$*,* $\Im$ *the real and imaginary parts, we have*

$$-\Re\left(\int_{\partial B} \left(E \wedge \overline{H}\right) \cdot n(x)\right) = \frac{\omega\varepsilon_0}{\kappa} \int_\Omega |J|^2\, dx \quad (\textit{Joule's dissipation}), \tag{3.6}$$

$$\lim_{\eta\to 0} \frac{1}{\theta_\eta} \int_{T_\eta} |E_\eta|^2 = \frac{1}{\kappa} \int_\Omega |J|^2, \tag{3.7}$$

$$\lim_{\eta\to 0} \int_B \left(\mu_0\, |H_\eta|^2 - \varepsilon_0\, |E_\eta|^2\right) = \int_B \left(\mu_0\, |H|^2 - \varepsilon_0\, |E|^2\right) - \varepsilon_0\, \Im\left(\int_\Omega \overline{J} E_3\right). \tag{3.8}$$

Proof. (i) *Regularity.* We begin by noticing that for every open subset $\mathcal{O} \subset \mathbb{R}^3$ we have the identity (see [12] for a more precise statement including traces)

$$W^{1,2}_{\text{loc}}(\mathcal{O}; \mathbb{C}^3) = \left\{u \in L^2_{\text{loc}}(\mathcal{O}; \mathbb{C}^3) \ :\ \operatorname{curl} u \in L^2_{\text{loc}}(\mathcal{O}),\ \operatorname{div} u \in L^2_{\text{loc}}(\mathcal{O})\right\}. \tag{3.9}$$

By (3.1), H is divergence-free, and $\operatorname{curl} H$ belongs to $L^2_{\text{loc}}(\mathbb{R}^3)$. Therefore by (3.9), we have $H \in W^{1,2}_{\text{loc}}(\mathbb{R}^3)$. From the 1D propagation equation satisfied by J,

we infer easily that J belongs to $L^2(\mathcal{D}, W^{1,2}(0,L))$. In particular, the traces j_0, j_L of J on the basis belong to $L^2(\mathcal{D}_L)$ and $L^2(\mathcal{D}_0)$, respectively. On the other hand, $\operatorname{curl} E \in L^2_{\mathrm{loc}}(\mathbb{R}^3)$, and from the second equation in (3.1) the distributional divergence of E as a measure satisfies

$$\operatorname{div} E = -i\kappa \left(\frac{\partial J}{\partial x_3} 1_\Omega \, dx - j_L \, \mathcal{H}^2 \llcorner \mathcal{D}_L + j_0 \, \mathcal{H}^2 \llcorner \mathcal{D}_0 \right).$$

In particular, the trace of $\operatorname{div} E$ on the open set $\mathbb{R}^3 \setminus (\mathcal{D}_0 \cup \mathcal{D}_L)$ belongs to L^2_{loc}. It then follows from (3.9) that $E \in W^{1,2}_{\mathrm{loc}}(\mathbb{R}^3 \setminus \mathcal{D}_0 \cup \mathcal{D}_L)$.

Uniqueness. This will be a consequence of (3.6). Indeed, by linearity, we are reduced to showing that any solution (E, H, J) for (3.1) vanishes over all $\mathbb{R}^3$, provided that (E^i, H^i) is identically 0. In this case, by Silver–Müller radiation conditions, we know that $\Re \left(\int_{\partial B} E \wedge \overline{H} \right) \cdot n(x)$ is nonnegative. Then by (3.6), J is identically 0 on Ω, and (E, H) satisfies the equation $\Delta w + k_0^2 w = 0$ in all $\mathbb{R}^3$ and the outgoing wave condition at infinity. The conclusion $E = H = 0$ follows classically (see, for example, [6] or [10]).

(ii) By using the propagation equation for J in (3.1) together with the Neumann boundary condition, and after integrating by parts with respect to x_3, we are led to

$$-\int_\Omega \left| \frac{\partial J}{\partial x_3} \right|^2 dx + \int_\Omega \left(k_0^2 + \frac{2i\pi\gamma}{\kappa} \right) |J|^2 \, dx = \int_\Omega 2i\pi\gamma \, E_3 \overline{J} dx.$$

Thus, taking the imaginary parts,

$$\Re \left(\int_\Omega E_3 \overline{J} dx \right) = \frac{1}{\kappa} \int_\Omega |J|^2 \, dx. \tag{3.10}$$

Now from (3.1) we have

$$\int_B \left(\operatorname{curl} E \cdot \overline{H} - \operatorname{curl} \overline{H} \cdot E \right) dx = i\omega \int_B \left(\mu_0 |H|^2 - \varepsilon_0 |E|^2 \right) - \omega\varepsilon_0 \int_\Omega \overline{J} E_3 \, dx.$$

By integrating the left-hand side of the previous equality by parts and by taking into account (3.10), we deduce (3.6) together with the relation

$$\Im \left(\int_{\partial B} E \wedge \overline{H} \cdot n(x) \right) = \omega \int_B \left(\mu_0 |H|^2 - \varepsilon_0 |E|^2 \right) - \omega \, \varepsilon_0 \, \Im \left(\int_\Omega \overline{J} E_3 \right). \tag{3.11}$$

In a similar way, from (2.16), we derive that

$$\Im \left(\int_{\partial B} \left(E_\eta \wedge \overline{H_\eta} \right) \cdot n(x) \right) = \omega \int_B \left(\mu_0 |H_\eta|^2 - \varepsilon_0 |E_\eta|^2 \right). \tag{3.12}$$

The convergences (3.7), (3.8) are then a straightforward consequence of (3.6), (3.11) and of the convergence of the left-hand expression of (3.12) towards $\int_{\partial B} \left(E \wedge \overline{H} \right) \cdot n(x)$. □

Proof of Theorem 3.1. The proof proceeds in two steps.

Step 1. We assume that the boundedness condition (2.7) is satisfied. Then by Proposition 2.4, we deduce that there exists a triplet (E, H, J) $(L^2(B))^3 \times (L^2(B))^3 \times L^1(\Omega)$ such that a suitable subsequence of $\{(E_\eta, H_\eta, F_\eta)\}$ does converge (weakly and weakly star) to $(E, H, J\, e_3)$. Moreover, by Lemma 2.1, (E, H) can be extended to

all $\mathbb{R}^3$ so that (E,H) solves (3.3) and the convergence $(E_\eta,H_\eta)\to(E,H)$ holds in $C^\infty(K)$ for all compact $K\subset\mathbb{R}^3\setminus\Omega$. Then, by Proposition 2.8, passing to the limit in (2.6) with the help of Proposition 2.4(iii), we find that (E,H,J) solves the system (3.1). Owing to the uniqueness property proved in Lemma 3.5, we conclude that the whole sequence $\{(E_\eta,H_\eta,F_\eta)\}$ converges and that all the conclusions of Theorem 3.1 hold true.

Step 2. We normalize the electromagnetic fields as follows,

$$u_\eta:=\frac{E_\eta}{t_\eta},\quad v_\eta:=\frac{H_\eta}{t_\eta},\quad w_\eta:=\frac{F_\eta}{t_\eta},\qquad\text{where}\quad t_\eta:=\sqrt{\int_B(|E_\eta|^2+|H_\eta|^2)dx},$$

and prove (2.7) by contradiction, assuming that $t_\eta\to+\infty$. By linearity, the triplet (u_η,v_η,w_η) satisfies the diffraction problem (2.6) and the outgoing wave condition with respect to the rescaled incident wave $(\frac{E^i}{t_\eta},\frac{H^i}{t_\eta})$. As it is bounded in $L^2(B)$, we may apply the conclusions of Theorem 3.1, where we have substituted (E^i,H^i) with the vanishing limit of $(\frac{E^i}{t_\eta},\frac{H^i}{t_\eta})$. By the uniqueness property proved in Lemma 3.5, we find that $(u_\eta,v_\eta)\rightharpoonup 0$ in $L^2_{\rm loc}$ whereas $w_\eta\overset{*}{\rightharpoonup}0$ in $L^1(\Omega)$. Furthermore, by the second assertion of Lemma 3.5, we obtain

$$\lim_\eta\ \frac{1}{\theta_\eta}\int_{T_\eta}|u_\eta|^2=0,\quad \lim_\eta\int_B\left(\mu_0\,|v_\eta|^2-\varepsilon_0\,|u_\eta|^2\right)\,dx=0. \tag{3.13}$$

In particular, recalling that $\int_B(|u_\eta|^2+|v_\eta|^2)=1$, we infer that

$$\lim_\eta\int_B\mu_0\,|v_\eta|^2=\lim_\eta\int_B\varepsilon_0\,|v_\eta|^2=\frac{\varepsilon_0\mu_0}{\varepsilon_0+\mu_0}>0.$$

Therefore, in order to find a contradiction, we need only to prove that $v_\eta\to0$ strongly in $L^2(B)$. To that aim, we consider a bounded open subset $B'\supset\supset B$ and use the following claim:

$$w_\eta:=\frac{1}{\theta_\eta}E_\eta 1_{T\eta}\ \to0\qquad\text{strongly in } W^{-1,2}(B'). \tag{3.14}$$

Then by the weak convergence of u_η to 0 in $L^2_{\rm loc}$, we infer that

$$\operatorname{curl}v_\eta=-i\omega\varepsilon_0\,(u_\eta+iw_\eta)\ \to0\qquad\text{strongly in } W^{-1,2}(B').$$

Since v_η is divergence-free, by the div-curl lemma [23], we deduce that $|v_\eta|^2$ converges to 0 weakly star in $L^1(B')$. By localizing this convergence with a continuous test function $\varphi\in\mathcal{D}(B',[0,1])$ such that $\varphi=1$ on B, we arrive at $\int_B|v_\eta|^2\to0$, and hence the contradiction.

It remains to prove (3.14), which is a consequence of estimate (2.34). Indeed, for every $\varphi\in(\mathcal{D}(B'))^3$, we have

$$\begin{aligned}\int_{B'}w_\eta\cdot\varphi\,dx=\frac{\kappa}{\theta_\eta}\int_{T_\eta}u_\eta\cdot\varphi\,dx&\le\kappa\left[\frac{1}{\theta_\eta}\int_{T_\eta}|u_\eta|^2\,dx\right]^{1/2}\left[\frac{1}{\theta_\eta}\int_{T_\eta}|\varphi|^2\,dx\right]^{1/2}\\&\le C\left[\frac{1}{\theta_\eta}\int_{T_\eta}|u_\eta|^2\,dx\right]^{1/2}\|\varphi\|_{W^{1,2}(B')}.\end{aligned}$$

The claim (3.14) follows from the left-hand-side convergence statement in (3.13). The proof of Theorem 3.1 is finished. ∎

4. Fibers with infinite length and numerical examples. In this section we sketch the natural extension of our previous results to the case of fibers of infinite length and present some features of the associated limit model.

4.1. Fibers of infinite length. We wish to model the diffraction by e_3-parallel fibers of infinite length. To that aim, we choose $\Omega_L := \mathcal{D} \times (-L/2, L/2)$ to be our reference obstacle and consider the limit as $L \to \infty$. We expect some limit equations on the infinite cylinder $\Omega_\infty := \mathcal{D} \times (-\infty, +\infty)$ with suitable transmission conditions. In fact, there exist many ways to proceed, namely, the following:

(a) First we fix L and pass to the limit as $\eta \to 0$, exploiting the results of section 3. Then we pass to the limit as $L \to \infty$.

(b) First we consider for fixed η the diffraction problem associated with Ω_∞. This allows us to reduce the problem to a scalar diffraction problem in dimension two. Indeed, we may decompose the incident wave by means of a Fourier transform in x_3, and then we look for solutions with a multiplicative x_3-dependence in $\exp(i\beta x_3)$. In a second step, we pass to the limit as $\eta \to 0$.

Of course all intermediate situations between (a) and (b) could also be considered, that is, taking a length $L(\eta)$ tending to infinity as $\eta \to 0$. The case (b) has been investigated in [14], leading for $\kappa = +\infty$ to a homogenized medium characterized by a diagonal effective permittivity tensor with a possibly negative eigenvalue in the x_3-direction. Let us sketch out how the same result can be obtained by following the strategy (a).

Owing to Theorem 3.1 (see (3.1)), we have to find the equation satisfied by the limit as $L \to \infty$ of the solution (E_L, H_L, j_L) of

$$
\left\{
\begin{array}{lll}
\operatorname{curl} E_L & = i\omega\mu_0\, H_L & \text{on } \mathbb{R}^3,\\
\operatorname{curl} H_L & = -i\omega\varepsilon_0\,(E_L + i\kappa J_L\, 1_{\Omega_L} e_3) & \text{on } \mathbb{R}^3,\\
\frac{\partial^2 J_L}{\partial x_3^2} + (k_0^2 + \frac{2i\pi\gamma}{\kappa})\, J_L & = 2i\pi\gamma E_L \cdot e_3 & \text{on } \Omega_L,\\
\frac{\partial J_L}{\partial x_3} & = 0 & \text{on } \mathcal{D}_{\pm L/2},\\
(E_L - E^i, H_L - H^i) & \text{satisfies the outgoing wave condition.} &
\end{array}
\right. \tag{4.1}
$$

Let us denote by (E, H, J) such a limit in L^2_{loc} (possibly obtained after extracting a subsequence). Passing to the limit in the three first equations of (4.1), we are led to

$$
\left\{
\begin{array}{lll}
\operatorname{curl} E & = i\omega\mu_0\, H & \text{on } \mathbb{R}^3,\\
\operatorname{curl} H & = -i\omega\varepsilon_0\,(E + i\ J\, e_3) & \text{on } \mathbb{R}^3,\\
\frac{\partial^2 J}{\partial x_3^2} + (k_0^2 + \frac{2i\pi\gamma}{\kappa})\, J & = 2i\pi\,\gamma E_3 & \text{on } (-\infty, +\infty).
\end{array}
\right. \tag{4.2}
$$

The difficult task is now to determine the radiation condition satisfied by (E, H) at infinity. By Fourier transform with respect to x_3, it is possible to reduce to an incident wave with a multiplicative x_3-dependence in $\exp(i\beta x_3)$. Then the L^2_{loc} solutions (E, H, J) of (4.2) can be shown to have the same x_3-dependence, and therefore the solution J of the third equation in (4.2) reads as

$$
J = \frac{2i\pi\,\gamma}{k_0^2 - \beta^2 + \frac{2i\pi\gamma}{\kappa}}\, E_3. \tag{4.3}
$$

Writing $E_3^i(x) = u^i(x_1, x_2)\exp(i\beta x_3)$, $E_3(x) = u(x_1, x_2)\exp(i\beta x_3)$ and substituting into (4.2), we obtain that u is solution of the following bidimensional diffraction problem:

$$\begin{cases} \Delta u + (k_0^2\,\varepsilon(x) - \beta^2)\,u = 0 & \text{on } \mathbb{R}^2, \\ (u - u^i) & \text{satisfies the outgoing wave condition,} \end{cases} \tag{4.4}$$

where $\varepsilon(x)$ is defined on $\mathbb{R}^2$ by

$$\varepsilon(x) = \begin{cases} \varepsilon^{\text{eff}}(\beta) & \text{if } x \in \mathcal{D}, \\ \varepsilon_0 & \text{otherwise,} \end{cases} \qquad \varepsilon^{\text{eff}}(\beta) = 1 - \frac{2\pi\gamma}{k_0^2}\left[\frac{k_0^2 - \beta^2}{k_0^2 - \beta^2 + \frac{2i\pi\gamma}{\kappa}}\right], \tag{4.5}$$

and the outgoing wave condition has to be understood in the sense of Sommerfeld (see, for instance, [6] or [10]). The magnetic field H and the horizontal components of E are then deduced in a standard way.

To summarize, we have obtained that, for an off-plane incident wave (described by an $\exp(i\beta x_3)$ dependence), the infinite wire mesh photonic crystal behaves like an anisotropic homogeneous medium characterized by the effective permittivity tensor $\operatorname{diag}\{1, 1, \varepsilon^{\text{eff}}(\beta)\}$. The dependence of $\varepsilon^{\text{eff}}(\beta)$ with respect to β indicates that the limit behavior is spatially nonlocal: it involves a convolution with respect to the inverse Fourier transform of $\varepsilon^{\text{eff}}(\beta)$. However, this effect disappears as $\kappa \to \infty$. Indeed, we observe that, for $\kappa = +\infty$ (infinite conductivity), $\varepsilon^{\text{eff}}(\beta) = 1 - \frac{2\pi\gamma}{k_0^2}$ is independent of β. We recover the same results as in [14], where it was pointed out that ε^{eff} becomes negative below a cut-off wave number $k_p = \sqrt{2\pi\gamma}$.

4.2. Numerical results. In this section, we present some numerical results obtained for infinitely long and perfectly conducting fibers; that is, the effective permittivity is given by (4.5), where $\kappa = +\infty$. We consider a stack of ten diffraction gratings infinite in extent in the horizontal direction. The gratings are made of parallel, infinitely long and perfectly conducting fibers, and the period is denoted by d. The distance between each grating is equal to d. The radius of the fibers is $r = d/200$. The structure is illuminated by a plane wave under normal incidence (see Figure 2) and linearly polarized in $E_{||}$; that is, the incident field reads as $E^i(y) = e^{-ik_0 y}e^{-ik_0 ct}e_3$, where $k_0 = 2\pi/\lambda$, λ is the wavelength, and c is the speed of light in a vacuum. Below the stack, the electric field can be expanded in a Rayleigh series [18]:

$$E^t(y) = \Sigma_n a_n e^{-ik_n y}e_3, \tag{4.6}$$

where $k_n = \sqrt{k_0^2 - (n\pi/d)^2}$, the square root is defined with a cut along $i\mathbb{R}^-$, and the determination $\sqrt{-1} = i$. The sum in (4.6) splits into two parts: one finite sum for which k_n is real, and an infinite one for which k_n is imaginary. The latter corresponds to evanescent waves, that do not contribute to the far field. The transmitted field is computed numerically by means of a rigorous modal method [19]. We are interested in the behavior of the field when the wavelength is large with respect to the period. In normal incidence and for $\lambda/d > 1$ there is only one propagating wave in the finite sum above, and we plot the function $\lambda \to |a_0(\lambda)|^2$, the curve being given by the dashed line in Figure 3. It can be seen that for $\lambda/d > 5$ the transmission is strongly damped, corresponding to the forbidden band. In the homogenization result, it is stated that the structure behaves for large wavelengths as a homogeneous medium with permittivity $\varepsilon^{\text{eff}}(\lambda) = 1 - 2\pi\gamma/k_0^2$ (because of the polarization we have to make $\beta = 0$ in (4.5)). Our point is now to check numerically to what extent the real structure can be considered as homogeneous. To that aim we consider a homogeneous slab of

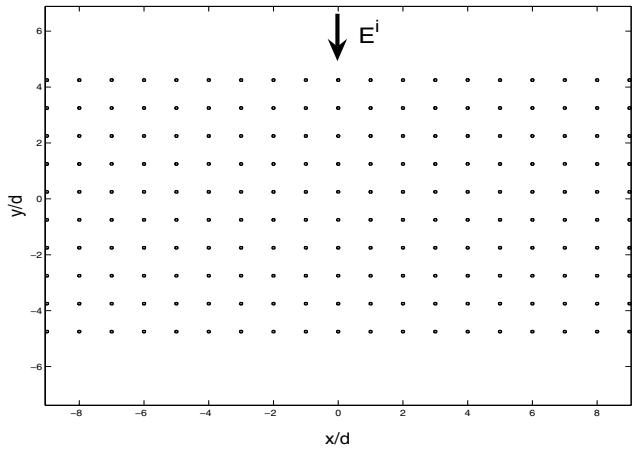


FIG. 2. *Cross section of the metallic photonic crystal.*

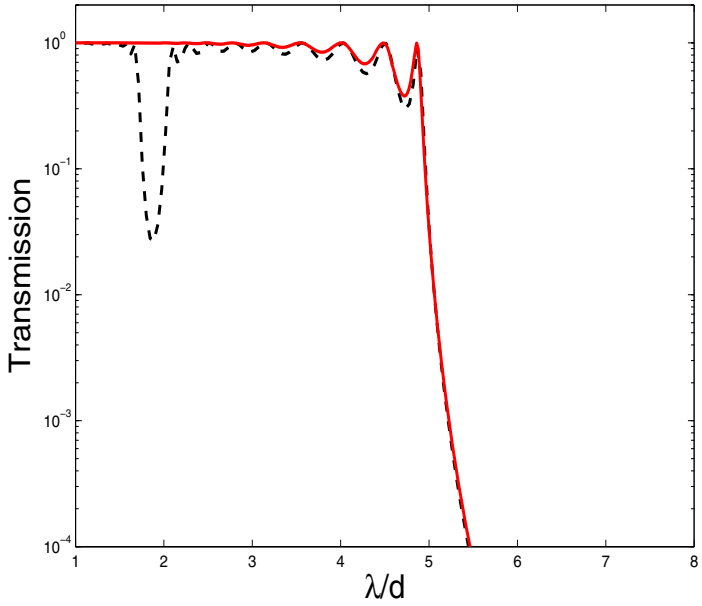


FIG. 3. *Transmitted energy through the device of Figure 2 (dashed line) and through the homogenized structure (solid line).*

height $10d$ with permittivity ε^{eff} illuminated under normal incidence by a plane wave, i.e., the same set-up as in the case of the heterogeneous structure. Below the slab, the electric field reads as $E^t(y) = t(\lambda)e^{-ik_0 y)}$, and we plot the function $\lambda \to |t(\lambda)|^2$. We expect a good fit between the curves at least for large wavelengths. One important parameter here is γ. A direct numerical application of (2.29) gives $\gamma \sim 0.2$, but we

must stress that this value is obtained as a limit as η tends to 0. The value of γ can be determined more precisely by studying correctors [16]. A numerical test shows that the value $\gamma = 0.25$ provides an excellent fit between the real transmission and that obtained from the slab (solid line in Figure 3). This shows that, up to a minor correction in the numerical value of γ, the homogeneous behavior reproduces very precisely that of the real device. Moreover, it can be seen in Figure 3 that the curves coincide for values of the wavelengths that are not very high, i.e., for $\lambda/d > 4$.

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[2] M. Artola, *Homogenization and electromagnetic wave propagation in composite media with high conductive inclusions*, in Composite Media Homogenization Theory, G. Dal Maso and G. Dell'Antonio, eds., World Scientific, River Edge, NJ, 1993, pp. 1–17.

[3] M. Bellieud and G. Bouchitté, *Homogenization of elliptic problems in a fiber reinforced structure. Nonlocal effects*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 26 (1998), pp. 407–436.

[4] G. Bouchitté and I. Fragalà, *Homogenization of thin structures by two-scale method with respect to measures*, SIAM J. Math. Anal., 32 (2001), pp. 1198–1226.

[5] G. Bouchitté and I. Fragalà, *Homogenization of elastic problems on thin structures: A measure-fattening approach*, J. Convex Anal., 9 (2002), pp. 339–362.

[6] M. Cessenat, *Mathematical Methods in Electromagnetism. Linear Theory and Applications*, Ser. Adv. Math. Appl. Sci. 41, World Scientific, River Edge, NJ, 1996.

[7] D. Cioranescu and F. Murat, *Un terme étrange venu d'ailleurs*, I, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. II, Res. Notes in Math. 60, Pitman, Boston, 1982, pp. 98–138.

[8] D. Cioranescu and F. Murat, *Un terme étrange venu d'ailleurs*, II, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. III, Res. Notes in Math. 60, Pitman, Boston, 1982, pp. 389–390.

[9] D. Cioranescu and J. Saint Jean Paulin, *Homogenization in open sets with holes*, J. Math. Anal. Appl., 71 (1979), pp. 590–607.

[10] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.

[11] J. P. Dowling, *Photonic & Sonic Band-Gap Bibliography*, Louisiana State University, Baton Rouge, LA, 2004, online at http://phys.lsu.edu/~jdowling/pbgbib.html.

[12] G. Duvaut and J.-L. Lions, *Les inéquations en mécanique et en physique*, Travaux et Recherches Mathématiques 21, Dunod, Paris, 1972.

[13] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.

[14] D. Felbacq and G. Bouchitté, *Homogenization of a set of parallel fibers*, Waves in Random Media, 7 (1997), pp. 1–12.

[15] D. Felbacq and G. Bouchitté, *Low Frequency Scattering by a Set of Parallel Metallic Fibers:* II—*Homogenized Limit for a Non Vanishing Filling Ratio*, in progress.

[16] D. Felbacq, G. Bouchitté, and F. Zolla, *Bloch vector dependence of the plasma frequency in metallic photonic crystals*, Phys. Rev. E, submitted.

[17] C. Müller, *Foundations of the Mathematical Theory of Electromagnetic Waves*, Springer-Verlag, Berlin, 1969.

[18] M. Neviere and E. Popov, *Light Propagation in Periodic Media: Differential Theory and Design*, Marcel Dekker, New York, 2002.

[19] N. A. Nicorovici, R. C. McPhedran, and R. Petit, *Efficient calculation of the Green's function for electromagnetic scattering by gratings*, Phys. Rev. E, 49 (1994), p. 4563–4573.

[20] J. B. Pendry, A. J. Holden, W. J. Stewart, and I. Youngs, *Extremely low frequency plasmons in metallic mesostructures*, Phys. Rev. Lett., 76 (1996), pp. 4773–4776.

[21] L. Schwartz, *Théorie des distributions*, Hermann, Paris, 1966.

[22] J. A. Stratton, *Electromagnetic Theory*, McGraw-Hill, New York, 1941.

[23] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics (Heriot-Watt Symposium), Vol. IV, Pitman, Boston, MA, 1979, pp. 136–212.

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 2085–2106

TWO-DIMENSIONAL HIGH-ACCURACY SIMULATION OF RESISTIVITY LOGGING-WHILE-DRILLING (LWD) MEASUREMENTS USING A SELF-ADAPTIVE GOAL-ORIENTED *hp* FINITE ELEMENT METHOD*

D. PARDO†, L. DEMKOWICZ‡, C. TORRES-VERDÍN§, AND M. PASZYNSKI¶

Abstract. We simulate electromagnetic (EM) measurements acquired with a logging-while-drilling (LWD) instrument in a borehole environment. The measurements are used to assess electrical properties of rock formations. Logging instruments as well as rock formation properties are assumed to exhibit axial symmetry around the axis of a vertical borehole. The simulations are performed with a self-adaptive goal-oriented *hp*-finite element method that delivers exponential convergence rates in terms of the quantity of interest (for example, the difference in the electrical current measured at two receiver antennas) against the CPU time. Goal-oriented adaptivity allows for accurate approximations of the quantity of interest without the need to obtain an accurate solution in the entire computational domain. In particular, goal-oriented *hp*-adaptivity becomes essential to simulating LWD instruments, since it reduces the computational cost by several orders of magnitude with respect to the global energy-norm-based *hp*-adaptivity. Numerical results illustrate the efficiency and high accuracy of the method, and provide physical interpretation of resistivity measurements obtained with LWD instruments. These results also describe the advantages of using magnetic buffers in combination with solenoidal antennas for strengthening the measured EM signal so that the "signal-to-noise" ratio is minimized.

Key words. *hp*-finite elements, exponential convergence, goal-oriented adaptivity, computational electromagnetics, Maxwell's equations, through casing resistivity tools (TCRT)

AMS subject classifications. 78A25, 78A55, 78M10, 65N50

DOI. 10.1137/050631732

1. Introduction.

A plethora of energy-norm-based algorithms intended to generate *optimal* grids have been developed throughout recent decades (see, for example, [10, 23] and references therein) to accurately solve a large class of engineering problems. However, the energy-norm is a quantity of limited relevance for most engineering applications, especially when a particular objective is pursued, such as simulating the electromagnetic response of geophysical resistivity logging instruments in a borehole environment. In these instruments, the amplitude of the measurement (for example, the electric field) is typically several orders of magnitude smaller at the receiver antennas than at the transmitter antennas. Thus, small relative errors of the solution in the energy-norm *do not* imply small relative errors of the solution at the receiver

*Received by the editors May 17, 2005; accepted for publication (in revised form) February 14, 2006; published electronically October 16, 2006. This work was financially supported by Baker-Atlas and the *Joint Industry Research Consortium on Formation Evaluation* supervised by Prof. C. Torres-Verdin.

http://www.siam.org/journals/siap/66-6/63173.html

†Institute for Computational Engineering and Sciences (ICES) and Department of Petroleum and Geosystems Engineering, The University of Texas at Austin, Austin, TX 78712 (dzubiaur@yahoo.es).

‡Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, Austin, TX 78712 (leszek@ices.utexas.edu).

§Department of Petroleum and Geosystems Engineering, The University of Texas at Austin, Austin, TX 78712 (cverdin@uts.cc.utexas.edu).

¶Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, Austin, TX 78712 (maciek@ices.utexas.edu). On leave from Department of Computer Methods in Metallurgy, AGH University of Science and Technology, Cracow, Poland.

antennas. Indeed, it is not uncommon to construct adaptive grids delivering a relative error in the energy-norm below 1% while the solution at the receiver antennas still exhibits a relative error above 1000% (see [18]).

Consequently, in order to accurately simulate logging-while-drilling (LWD) resistivity measurements in this paper, we develop a self-adaptive strategy to approximate a specific feature of the solution. Refinement strategies of this type are called *goal-oriented* adaptive algorithms [16, 22], and are based on minimizing the error of a prescribed *quantity of interest* mathematically expressed in terms of a linear functional (see [5, 12, 17, 16, 22, 24] for details).

In this paper, we formulate, implement, and study (both theoretically and numerically) a self-adaptive *hp* goal-oriented algorithm intended to solve electrodynamic problems. This algorithm is an extension of the fully automatic (energy-norm-based) *hp*-adaptive strategy described in [10, 23], and a continuation of concepts presented in [19, 25] for elliptic problems.

We apply the self-adaptive *hp* goal-oriented algorithm to accurately simulate induction LWD instruments in a borehole environment with axial symmetry. These instruments are widely used by the geophysical logging industry, and their simulation requires resolution of electromagnetic (EM) singularities generated by the LWD geometry and rock formation materials [28], as well as resolution of high material contrasts that occur between the mandrel and the borehole.

Other methods for simulation of LWD measurements include the transmission line matrix method [14], fast Fourier transform [29], and finite differences [26, 13]. In contrast to previous contributions, here we consider a detailed geometry of the logging instrument, which requires the resolution of strong singularities in the EM fields, we account for the finite conductivity of the mandrel, we incorporate magnetic buffers in both transmitter and receiver antennas, we consider the effect of the magnetic permeability of the mandrel, and we provide extremely accurate results with guaranteed relative error bounds below 0.1% (0.001% if desired). We also consider a high contrast in conductivity among different layers in the formation, and we present a comparison between using two and three receiver antennas.

The organization of this paper is as follows. In section 2, we describe the main characteristics of induction logging instruments. We also describe our problem of interest, composed of an induction LWD instrument in a borehole environment, and used for the assessment of the rock formation electrical properties. In section 3, we introduce Maxwell's equations, governing the EM phenomena and explaining the physics of resistivity measurements. We also derive the corresponding variational formulation for axisymmetric problems. A self-adaptive goal-oriented *hp* algorithm for electrodynamic problems is described in section 4. The corresponding details of implementation are discussed in the same section. Simulations and numerical results concerning the response of LWD instruments in a borehole environment are shown in section 5. Section 6 draws the main conclusions and outlines future lines of research. Finally, in the appendix, we compare numerical results with a semianalytical solution obtained using Bessel functions for a simplified LWD model problem. The comparison is intended to verify the code as well as to illustrate the high-accuracy results obtained with the self-adaptive goal-oriented *hp*-finite element method (FEM).

2. Alternate current (AC) logging applications. In this article, we consider an induction[1] LWD instrument operating at 2 MHz. The instrument makes use of one of the following two types of source antennas/coils:

- solenoidal coils (Figure 1, left panel), and
- toroidal coils (Figure 1, right panel).

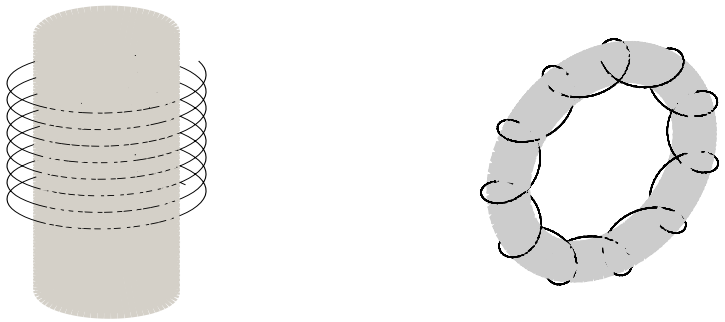


FIG. 1. *Two coil antennas: a solenoid antenna (left panel) composed of a wire wrapped around a cylinder, and a toroid antenna (right panel) composed of a wire wrapped around a toroid.*

2.1. Induction LWD instrumentsbased on solenoidal coils. For axisymmetric problems, these logging instruments generate a TM_ϕ field; i.e., the only nonzero components of the EM fields are E_ϕ, H_ρ, and H_z, where (ρ, ϕ, z) denote the cylindrical system of coordinates.

A solenoidal coil (Figure 1) produces an impressed current $\mathbf{J}^{imp}$ that we mathematically describe as

$$\mathbf{J}^{imp}(\mathbf{r}) = \hat{\phi} I \delta(\rho - a)\delta(z), \tag{2.1}$$

where I is the electric current measured in Amperes (A), δ is the Dirac's delta function, and a is the radius of the solenoid. In the numerical computations, we replace function $\delta(\rho - a)\delta(z)$ with an approximate function U_F that considers the finite dimensions of the coil, and such that $\int U_F d\rho dz = 1$.

The analytical electric far-field solution excited by a solenoidal coil of radius a radiating in homogeneous media is given in terms of the electric field by (see [15])

$$\mathbf{E} = \hat{\phi}\omega\mu k I \pi a^2 \frac{e^{-jkd}}{4\pi d}\left[1 - \frac{j}{kd}\right]\frac{\rho}{d}\,, \tag{2.2}$$

where $k = \sqrt{\omega^2\epsilon - j\omega\sigma}$ is the wave number; $j = \sqrt{-1}$ is the imaginary unit; ω is angular frequency; ϵ, μ, and σ stand for dielectric permittivity, magnetic permeability, and electrical conductivity of the medium, respectively; and d is the distance between the source coil and the receiver coil.

In order to avoid the dependence upon the dimensions of the solenoid, we impose a current on the solenoidal coil equal to $1/(\pi a^2)$A, i.e., equivalent to that of 1A with a

[1]Induction logging instruments are characterized by the fact that impressed current $\mathbf{J}^{imp}$ is divergence-free (i.e., $\nabla \cdot \mathbf{J}^{imp} = \mathbf{0}$).

vertical magnetic dipole (VMD). The corresponding far-field solution in homogeneous media is given by (see [15])

$$\mathbf{E} = \hat{\phi}\omega\mu k I \frac{e^{-jkd}}{4\pi d}\left[1 - \frac{j}{kd}\right]\frac{\rho}{d}\,. \tag{2.3}$$

Thus, solution (2.3) is independent of the dimensions of the coil.[2]

2.2. Induction LWD instruments based on toroidal coils. For axisymmetric problems, these logging instruments generate a TE_ϕ field; i.e., the only nonzero components of the EM fields are H_ϕ, E_ρ, and E_z.

A toroidal coil induces a *magnetic current* I_M in the azimuthal direction. If we place a toroid of radius a radiating in homogeneous media, the resulting magnetic far-field is given by (see [15])

$$\mathbf{H} = \hat{\phi}(\sigma + j\omega\epsilon)\pi a^2 I_M jk \frac{e^{-jkd}}{4\pi d}\left[1 - \frac{j}{kd}\right]\frac{\rho}{d}\,. \tag{2.4}$$

In order to avoid the dependence upon the dimensions of the toroid, we impose a magnetic current on the toroidal coil equal to that induced by a $(\sigma + j\omega\epsilon)$A electric current excitation with a vertical electrical dipole (VED), also known as a Hertzian dipole. The corresponding magnetic far-field solution in homogeneous media is given by (see [15])

$$\mathbf{H} = \hat{\phi}(\sigma + j\omega\epsilon) I jk \frac{e^{-jkd}}{4\pi d}\left[1 - \frac{j}{kd}\right]\frac{\rho}{d}\,. \tag{2.5}$$

In this case, $I_M = I/(\pi a^2)$.

2.2.1. Goal of the computations. We are interested in simulating the EM response of an induction LWD instrument in a borehole environment.

For a *solenoidal coil*, the main objective of our simulation is to compute the first difference of the voltage between the two receiving coils of radius a divided by the (vertical) distance Δz between them, i.e.,

$$\frac{V_1 - V_2}{\Delta z} = \left(\oint_{\mathbf{l}_1} \mathbf{E}(l)\, dl - \oint_{\mathbf{l}_2} \mathbf{E}(l)\, dl\right)/(\Delta z) = \frac{2\pi a}{\Delta z}(\mathbf{E}(l_1) - \mathbf{E}(l_2))\,, \tag{2.6}$$

where $\mathbf{l}_1$ and $\mathbf{l}_2$ are the first and second receiving coils, respectively, and $l_1 \in \mathbf{l}_1$, $l_2 \in \mathbf{l}_2$ are two arbitrary points located at the receiving coils. Notice that, due to the axisymmetry of the electric field, $\mathbf{E}(l_i^j) = \mathbf{E}(l_i^k)$ for all $l_i^j, l_i^k \in \mathbf{l}_i$.

This quantity of interest (first difference of voltage) is widely used in resistivity logging applications. Indeed, a first-order asymptotic approximation of the electric field response at low frequencies (Born's approximation) shows that the voltage at a receiver coil is proportional to the rock formation resistivity in the proximity of such a coil (see [15] for details). At higher frequencies (> 20 kHz), asymptotic approximations (see [3] for details) also indicate the dependence of the voltage upon the rock formation conductivity. Thus, an adequate approximation of the rock formation

[2] In resistivity logging applications, it is customary to consider solutions that have been divided by the *geometrical factor* (also called K-factor) [3], so that results are independent (as much as possible) of the logging instrument's geometry. Thus, solutions obtained from different logging instruments can be readily compared.

conductivity (which is unknown a priori in practical applications) can be estimated from the voltage measured at the receiving coils. Computing the first difference of the voltage between two receivers (rather than the voltage at one receiver) is convenient for improving the vertical resolution of the measurements. This well-known fact among well-logging practitioners will be illustrated here with numerical experiments.

For a *toroidal coil*, the main objective of these simulations is to compute the first difference of the electric current at the two receiving coils of radius a divided by the (vertical) distance Δz between them, i.e.,

$$\frac{I_1 - I_2}{\Delta z} = \left(\oint_{l_1} \mathbf{H}(l)\, dl - \oint_{l_2} \mathbf{H}(l)\, dl \right) / (\Delta z) = \frac{2\pi a}{\Delta z} (\mathbf{H}(l_1) - \mathbf{H}(l_2)) \,. \tag{2.7}$$

Notice that the main difference between a toroidal and a solenoidal coil is that the former generates an impressed magnetic current, while the latter produces an impressed electric current. This fact leads to the physical consideration that, if the voltage due to a solenoidal coil is proportional to the rock formation conductivity, then the electric current enforced by a toroidal coil is also proportional to the rock formation resistivity. Thus, the selection of the quantity of interest for toroidal coils (first difference of electric current) is dictated by the physical relation between solenoidal and toroidal coils and by the previous choice of a quantity of interest for solenoidal coils (first difference of voltage).

2.3. Description of an LWD instrument in a borehole environment. We consider an LWD instrument composed of the following axisymmetric materials (all dimensions are given in cm):

- one transmitter and two receiver coils defined on
  1. $\Omega_{C_1} = \{(\rho, \phi, z) : 7.1 < \rho < 7.3,\ -2.5 < z < 2.5\}$,
  2. $\Omega_{C_2} = \{(\rho, \phi, z) : 7.1 < \rho < 7.3,\ 98.75 < z < 101.25\}$, and,
  3. $\Omega_{C_3} = \{(\rho, \phi, z) : 7.1 < \rho < 7.3,\ 113.75 < z < 116.25\}$, respectively;
- three magnetic buffers with resistivity $10^4\ \Omega \cdot$m and relative permeability 10^4, defined on
  1. $\Omega_{B_1} = \{(\rho, \phi, z) : 6.675 < \rho < 6.985,\ -5 < z < 5\}$,
  2. $\Omega_{B_2} = \{(\rho, \phi, z) : 6.675 < \rho < 6.985,\ 97.5 < z < 102.5\}$, and,
  3. $\Omega_{B_3} = \{(\rho, \phi, z) : 6.675 < \rho < 6.985,\ 112.5 < z < 117.5\}$, respectively; and
- a metallic mandrel with resistivity $10^{-6}\ \Omega \cdot$m defined on $\Omega_M = \{(\rho, \phi, z) : \rho < 7.6\} - (\{(\rho, \phi, z) : 6.675 < \rho < 7.6,\ -5 < z < 5\} \cup \{(\rho, \phi, z) : 6.675 < \rho < 7.6,\ 97.5 < z < 102.5\} \cup \{(\rho, \phi, z) : 6.675 < \rho < 7.6,\ 112.5 < z < 117.5\})$.

This LWD instrument moves along the vertical direction (z-axis) in a subsurface borehole environment composed of

- a borehole mud with resistivity $0.1\ \Omega \cdot$m defined on
  1. $\Omega_{BH} = \{(\rho, \phi, z) : \rho < 10.795\} - (\cup_i \Omega_{B_i} \cup \Omega_M)$, and
- three formation materials of resistivities $100\ \Omega \cdot$m, $10000\ \Omega \cdot$m, and $1\ \Omega \cdot$m, defined on
  1. $\Omega_{M_1} = \{(\rho, \phi, z) : \rho \geq 10.795,\ (z < -50 \text{ or } z > 100)\}$,
  2. $\Omega_{M_2} = \{(\rho, \phi, z) : \rho \geq 10.795,\ -50 \leq z < 0\}$, and,
  3. $\Omega_{M_3} = \{(\rho, \phi, z) : \rho \geq 10.795,\ 0 \leq z \leq 100\}$, respectively.

Figure 2 shows the geometry of the described logging instrument and borehole environment.

3. Maxwell's equations. In this section, we first introduce the time-harmonic Maxwell equations in the frequency domain. They form a set of first-order partial

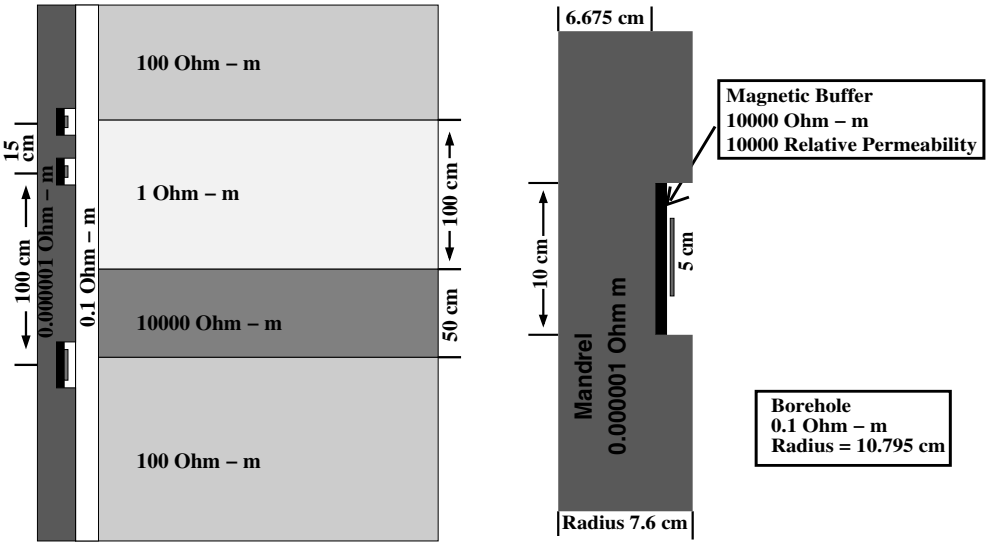


FIG. 2. *2D cross section of the geometry of an induction LWD problem composed of a metallic mandrel, one transmitter and two receiver coils equipped with magnetic buffers, a borehole, and four layers in the rock formation (with different resistivities). The right panel is an enlarged view of the geometry (left panel) in the vicinity of the transmitter antenna.*

differential equations (PDEs). Then, we describe boundary conditions needed for the simulation of our logging applications of interest. Finally, we derive a variational formulation in terms of either the electric or the magnetic field, and we reduce the dimension of the computational problem by considering axial symmetry.

3.1. Time-harmonic Maxwell equations. Assuming a time-harmonic dependence of the form $e^{j\omega t}$, where t denotes time and $\omega \neq 0$ is angular frequency, Maxwell's equations can be written as

$$
\left\{
\begin{array}{lll}
\nabla\times\mathbf{H} & = (\sigma + j\omega\epsilon)\mathbf{E} + \mathbf{J}^{imp} & \text{Ampere's law,} \\
\nabla\times\mathbf{E} & = -j\omega\mu\,\mathbf{H} - \mathbf{M}^{imp} & \text{Faraday's law,} \\
\nabla\cdot(\epsilon\mathbf{E}) & = \rho & \text{Gauss' law of electricity, and} \\
\nabla\cdot(\mu\mathbf{H}) & = 0 & \text{Gauss' law of magnetism.}
\end{array}
\right. \tag{3.1}
$$

Here $\mathbf{H}$ and $\mathbf{E}$ denote the magnetic and electric fields, respectively; $\mathbf{J}^{imp}$ is a prescribed, impressed electric current density; $\mathbf{M}^{imp}$ is a prescribed, impressed magnetic current density; ϵ, μ, and σ stand for dielectric permittivity, magnetic permeability, and electrical conductivity of the medium, respectively; and ρ denotes the electric charge distribution. We assume $\mu \neq 0$.

The equations described in (3.1) are to be understood in the distributional sense; i.e., they are satisfied in the classical sense in subdomains of regular material data, and they also imply appropriate interface conditions across material interfaces.

Energy considerations lead to the assumption that the absolute value of both electric field $\mathbf{E}$ and magnetic field $\mathbf{H}$ must be square integrable. According to $(3.1)_2$ and $(3.1)_4$, $\mathbf{M}^{imp}$ is divergence-free.

Maxwell's equations are not independent. Taking the divergence of Faraday's law

yields the Gauss' law of magnetism. By taking the divergence of Ampere's law, and by utilizing Gauss' electric law, we arrive at the so-called continuity equation,

$$\boldsymbol{\nabla} \cdot (\sigma \mathbf{E}) + j\omega\rho + \boldsymbol{\nabla} \cdot \mathbf{J}^{imp} = 0. \tag{3.2}$$

3.2. Boundary conditions (BCs). There exist a variety of BCs that can be incorporated into Maxwell's equations. In the following, we describe those BCs that are of interest for the logging applications discussed in this paper. At this point, we are considering general 3D domains. A discussion on boundary terms corresponding to the axisymmetry condition is postponed to section 3.4.

3.2.1. Perfect electric conductor (PEC). Maxwell's equations are to be satisfied in the whole space minus domains occupied by a PEC. A PEC is an idealization of a highly conductive media. Inside a region where $\sigma \to \infty$, the corresponding electric field converges to zero[3] by applying Ampere's law. Faraday's law implies that the tangential component of the electric field $\mathbf{E}$ must remain continuous across material interfaces in the absence of impressed magnetic surface currents. Consequently, the tangential component of the electric field must vanish along the PEC boundary, i.e.,

$$\boldsymbol{n} \times \mathbf{E} = \mathbf{0}, \tag{3.3}$$

where $\boldsymbol{n}$ is the unit normal (outward) vector.

Since the electric field vanishes inside a PEC, Faraday's law implies that the magnetic field should also vanish inside a PEC in the absence of magnetic currents. The same Faraday's law implies that the normal component of the magnetic field premultiplied by the permeability must remain continuous across material interfaces. Therefore, the normal component of the magnetic field must vanish along the PEC boundary, i.e.,

$$\boldsymbol{n} \cdot \mathbf{H} = 0. \tag{3.4}$$

The tangential component of magnetic field (surface current) and normal component of the electric field (surface charge density) need not be zero and may be determined a posteriori.

3.2.2. Source antennas. Antennas are modeled by prescribing an impressed volume current $\mathbf{J}^{imp}$. Using the equivalence principle (see, for example, [11]), we can replace the original impressed electric volume current $\mathbf{J}^{imp}$ with an equivalent electric surface current

$$\mathbf{J}_S^{imp} = [\mathbf{n} \times \mathbf{H}]_S, \tag{3.5}$$

defined on an arbitrary surface S enclosing the support of $\mathbf{J}^{imp}$, where $[\mathbf{n} \times \mathbf{H}]_S$ denotes

- the jump of $\mathbf{n} \times \mathbf{H}$ across S in the case of an interface condition, or
- simply $\mathbf{n} \times \mathbf{H}$ on S in the case of a boundary condition.

Similarly, an impressed magnetic volume current $\mathbf{M}^{imp}$ can be replaced by the equivalent magnetic surface current

$$\mathbf{M}_S^{imp} = -[\mathbf{n} \times \mathbf{E}]_{\mathbf{S}}, \tag{3.6}$$

defined on an arbitrary surface S enclosing the support of $\mathbf{M}^{imp}$.

[3] This result is true under the physical consideration that impressed volume current $\mathbf{J}^{imp}$ and $\sigma\mathbf{E}$ should remain finite, i.e., $\langle \mathbf{J}^{imp}, \psi \rangle, \langle \sigma\mathbf{E}, \psi \rangle < \infty$ for every test function ψ. See [21] for details.

3.2.3. Closure of the domain. We consider a bounded computational domain Ω. A variety of BCs can be imposed on the boundary $\partial\Omega$ such that the difference between solution of such a problem and solution of the original problem defined over $\mathbb{R}^3$ is small. For example, it is possible to use an infinite element technique (as described in [7]) or an absorption-type BC such as a perfect matched layer (PML) [6, 26, 13]. Also, since the EM fields and their derivatives decay exponentially in the presence of lossy media (nonzero conductivity), we may simply impose a homogeneous Dirichlet or Neumann BC on the boundary of a sufficiently large computational domain.

In the field of geophysical logging applications, it is customary to impose a homogeneous Dirichlet BC on the boundary of a large computational domain (for example, 2–20 meters in each direction from a 2 MHz source antenna in the presence of a resistive media). We will follow the same approach.

According to the BCs discussed above, we will divide boundary $\Gamma = \partial\Omega$ into the disjoint union of

- Γ_E, where $\mathbf{M}^{imp}_{\Gamma_E} = -[\mathbf{n}\times\mathbf{E}]_{\Gamma_\mathbf{E}}$ (with $\mathbf{M}^{imp}_{\Gamma_E}$ possibly zero), with
- Γ_H, where $\mathbf{J}^{imp}_{\Gamma_H} = [\mathbf{n}\times\mathbf{H}]_{\Gamma_H}$, (with $\mathbf{J}^{imp}_{\Gamma_H}$ possibly zero).

3.3. Variational formulation. From Maxwell's equations and the BCs described above, we derive the corresponding standard variational formulation in terms of the electric or magnetic field as follows.

First, we notice from Faraday's law that $\boldsymbol{\nabla}\times\mathbf{E} \in (L^2(\Omega))^3$ if and only if $\mathbf{M}^{imp} \in (L^2(\Omega))^3$. Since our objective is to find a solution $\mathbf{E} \in H(\mathbf{curl};\Omega) = \{\mathbf{F} \in (L^2(\Omega))^3 : \boldsymbol{\nabla}\times\mathbf{F} \in (L^2(\Omega))^3\}$, we shall assume in the case of the electric field formulation (E-formulation) derived below that $\mathbf{M}^{imp} \in (L^2(\Omega))^3$. If the prescribed $\mathbf{M}^{imp} \notin (L^2(\Omega))^3$, we may still solve Maxwell's equations with $H(\mathbf{curl})$-conforming finite elements for the magnetic field by using the H-formulation (3.3.2), or simply by prescribing an equivalent source $\tilde{\mathbf{M}}^{imp}$ such that $\mathbf{M}^{imp} - \tilde{\mathbf{M}}^{imp}$ does not radiate outside the antenna [27].

Similarly, for the H-formulation, we will assume that $\mathbf{J}^{imp} \in (L^2(\Omega))^3$.

3.3.1. E-formulation. By dividing Faraday's law by magnetic permeability μ, multiplying the resulting equation by $\boldsymbol{\nabla}\times\bar{\mathbf{F}}$, where $\mathbf{F} \in H_{\Gamma_E}(\mathbf{curl};\Omega) = \{\mathbf{F} \in H(\mathbf{curl};\Omega) \ : (\mathbf{n}\times\mathbf{F})|_{\Gamma_E} = 0\}$ is an arbitrary test function, and integrating over the domain Ω, we arrive at the identity

$$\int_\Omega \frac{1}{\mu}(\boldsymbol{\nabla}\times\mathbf{E})\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})dV = -j\omega\int_\Omega \mathbf{H}\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})dV - \int_\Omega \frac{1}{\mu}\mathbf{M}^{imp}\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})dV. \tag{3.7}$$

Integrating $\int_\Omega \mathbf{H}\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})\,dV$ by parts, and applying Ampere's law, we obtain

$$\begin{aligned}\int_\Omega \mathbf{H}\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})\,dV &= \int_\Omega(\boldsymbol{\nabla}\times\mathbf{H})\cdot\bar{\mathbf{F}}\,dV - \int_{\Gamma_H}\mathbf{n}\times\mathbf{H}\cdot\bar{\mathbf{F}}_t\,dS\\ &= \int_\Omega(\sigma+j\omega\epsilon)\mathbf{E}\cdot\bar{\mathbf{F}}\,dV + \int_\Omega \mathbf{J}^{imp}\cdot\bar{\mathbf{F}}\,dV - \int_{\Gamma_H}\mathbf{n}\times\mathbf{H}\cdot\bar{\mathbf{F}}_t\,dS.\end{aligned} \tag{3.8}$$

$\mathbf{F}_t = \mathbf{F} - (\mathbf{F}\cdot\mathbf{n})\cdot\mathbf{n}$ is the tangential component of vector $\mathbf{F}$ on Γ_H, and $\mathbf{n}$ is the unit normal outward (with respect to Ω if $\Gamma_H \subset \partial\Omega$) vector. Substitution of (3.8) into

(3.7) and use of (3.5) yields the following variational formulation:

$$
\left\{
\begin{array}{l}
\text{Find } \mathbf{E} \in \mathbf{E}_{\Gamma_E} + H_{\Gamma_E}(\mathbf{curl};\Omega) \text{ such that} \\
\displaystyle\int_\Omega \frac{1}{\mu}(\boldsymbol{\nabla}\times\mathbf{E})\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})\,dV - \int_\Omega k^2\mathbf{E}\cdot\bar{\mathbf{F}}\,dV = -j\omega\int_\Omega \mathbf{J}^{imp}\cdot\bar{\mathbf{F}}\,dV \\
\quad + j\omega\displaystyle\int_{\Gamma_H}\mathbf{J}^{imp}_{\Gamma_H}\cdot\bar{\mathbf{F}}_t\,dS - \int_\Omega\frac{1}{\mu}\mathbf{M}^{imp}\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})\,dV \;\forall\,\mathbf{F}\in H_{\Gamma_E}(\mathbf{curl};\Omega),
\end{array}
\right. \tag{3.9}
$$

where $k^2=\omega^2\epsilon - j\omega\sigma$ is the wave number and $\mathbf{E}_{\Gamma_E}$ is a lift (typically $\mathbf{E}_{\Gamma_E}=0$) of the essential BC data $\mathbf{E}_{\Gamma_E}=-M^{imp}_{\Gamma_E}$ (denoted with the same symbol).

Conversely, we can derive (3.1), (3.3), and (3.5) from variational problem (3.9).

3.3.2. H-formulation. By dividing Ampere's law by $\sigma+j\omega\epsilon$, multiplying the resulting equation by $\boldsymbol{\nabla}\times\bar{\mathbf{F}}$, where $\mathbf{F}\in H_{\Gamma_H}(\mathbf{curl};\Omega)=\{\mathbf{F}\in H(\mathbf{curl};\Omega)\ :(\mathbf{n}\times\mathbf{F})|_{\Gamma_H}=0\,\}$ is an arbitrary test function, and integrating over the domain Ω, we arrive at the identity

$$
\begin{aligned}
-j\omega\int_\Omega\frac{1}{k^2}(\boldsymbol{\nabla}\times\mathbf{H})\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})dV &= \int_\Omega\mathbf{E}\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})\,dV \\
&\quad - j\omega\int_\Omega\frac{1}{k^2}\mathbf{J}^{imp}\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})\,dV.
\end{aligned} \tag{3.10}
$$

Integrating $\int_\Omega\mathbf{E}\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})\,dV$ by parts and applying Faraday's law, we obtain

$$
\begin{aligned}
\int_\Omega\mathbf{E}\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})\,dV &= \int_\Omega(\boldsymbol{\nabla}\times\mathbf{E})\cdot\bar{\mathbf{F}}\,dV - \int_{\Gamma_E}\mathbf{n}\times\mathbf{E}\cdot\bar{\mathbf{F}}_t\,dS \\
&= -j\omega\int_\Omega\mu\mathbf{H}\cdot\bar{\mathbf{F}}\,dV - \int_\Omega\mathbf{M}^{imp}\cdot\bar{\mathbf{F}}\,dV - \int_{\Gamma_E}\mathbf{n}\times\mathbf{E}\cdot\bar{\mathbf{F}}_t\,dS.
\end{aligned} \tag{3.11}
$$

Substitution of (3.11) into (3.10) and use of (3.6) yields the following variational formulation:

(3.12)

$$
\left\{
\begin{array}{l}
\text{Find } \mathbf{H} \in \mathbf{H}_{\Gamma_H} + H_{\Gamma_H}(\mathbf{curl};\Omega) \text{ such that} \\
\displaystyle\int_\Omega\frac{1}{\sigma+j\omega\epsilon}(\boldsymbol{\nabla}\times\mathbf{H})\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})dV + j\omega\int_\Omega\mu\mathbf{H}\cdot\bar{\mathbf{F}}dV = -\int_\Omega\mathbf{M}^{imp}\cdot\bar{\mathbf{F}}dV \\
\quad + \displaystyle\int_{\Gamma_E}\mathbf{M}^{imp}_{\Gamma_E}\cdot\bar{\mathbf{F}}_t\,dS + \int_\Omega\frac{1}{\sigma+j\omega\epsilon}\mathbf{J}^{imp}\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})dV\;\forall\,\mathbf{F}\in H_{\Gamma_H}(\mathbf{curl};\Omega),
\end{array}
\right.
$$

where $\mathbf{H}_{\Gamma_H}$ is a lift (typically $\mathbf{H}_{\Gamma_H}=0$) of the essential BC data $\mathbf{H}_{\Gamma_H}=J^{imp}_{\Gamma_H}$ (denoted with the same symbol).

3.4. Cylindrical coordinates and axisymmetric problems. We consider cylindrical coordinates (ρ,ϕ,z). For the geophysical logging applications considered in this article, we assume that both the logging instrument and the rock formation properties are axisymmetric (invariant with respect to the azimuthal coordinate ϕ) around the axis of the borehole. Under this assumption, we obtain that for any vector field $\mathbf{A}=\hat{\rho}A_\rho+\hat{\phi}A_\phi+\hat{z}A_z$,

$$
\boldsymbol{\nabla}\times\mathbf{A} = -\hat{\rho}\frac{\partial A_\phi}{\partial z} + \hat{\phi}\left(\frac{\partial A_\rho}{\partial z}-\frac{\partial A_z}{\partial\rho}\right) + \hat{z}\frac{1}{\rho}\frac{\partial(\rho A_\phi)}{\partial\rho}\ . \tag{3.13}
$$

3.4.1. E-formulation. Next, we consider the space of all test functions $\mathbf{F} \in H_D(\mathbf{curl};\Omega)$ such that $\mathbf{F} = (0, F_\phi, 0)$. According to (3.13),

$$\boldsymbol{\nabla}\times\mathbf{F} = -\hat{\rho}\frac{\partial F_\phi}{\partial z} + \hat{z}\frac{1}{\rho}\frac{\partial(\rho F_\phi)}{\partial \rho}\ . \tag{3.14}$$

Variational formulation (3.9) reduces to a formulation in terms of the scalar field E_ϕ only, namely,

$$\left\{\begin{array}{l} \text{Find } E_\phi \in E_{\phi,D} + \tilde{H}^1_D(\Omega) \text{ such that} \\ \displaystyle\int_\Omega \frac{1}{\mu}\left(\frac{\partial E_\phi}{\partial z}\frac{\partial \bar{F}_\phi}{\partial z} + \frac{1}{\rho^2}\frac{\partial(\rho E_\phi)}{\partial \rho}\frac{\partial(\rho \bar{F}_\phi)}{\partial \rho}\right) dV - \int_\Omega k^2 E_\phi\, \bar{F}_\phi\, dV \\ \qquad = -\,j\omega \displaystyle\int_\Omega J^{imp}_\phi\, \bar{F}_\phi\, dV + j\omega \int_{\Gamma_N} J^{imp}_{\phi,\Gamma_N}\, \bar{F}_\phi\, dS \\ \qquad - \displaystyle\int_\Omega \frac{1}{\mu}\left[-M^{imp}_\rho \frac{\partial \bar{F}_\phi}{\partial z} + M^{imp}_z \frac{1}{\rho}\frac{\partial(\rho \bar{F}_\phi)}{\partial \rho}\right] dV \quad \forall\, F_\phi \in \tilde{H}^1_D(\Omega), \end{array}\right. \tag{3.15}$$

where $\tilde{H}^1_D(\Omega) = \{E_\phi : (0, E_\phi, 0) \in H_D(\mathbf{curl};\Omega)\} = \{E_\phi \in L^2(\Omega) \ :\ \frac{1}{\rho}E_\phi + \frac{\partial E_\phi}{\partial \rho} \in L^2(\Omega),\ \frac{\partial E_\phi}{\partial z} \in L^2(\Omega),\ E_\phi|_{\Gamma_D} = 0\}$. Similarly, for a test function $\mathbf{F} = (F_\rho, 0, F_z)$, variational problem (3.9) simplifies to

$$\left\{\begin{array}{l} \text{Find } \mathbf{E} = (E_\rho, 0, E_z) \in \mathbf{E}_D + \tilde{H}_D(\mathbf{curl};\Omega) \text{ such that} \\ \displaystyle\int_\Omega \frac{1}{\mu}\left(\frac{\partial E_\rho}{\partial z} - \frac{\partial E_z}{\partial \rho}\right)\left(\frac{\partial \bar{F}_\rho}{\partial z} - \frac{\partial \bar{F}_z}{\partial \rho}\right) dV - \int_\Omega k^2(E_\rho \bar{F}_\rho + E_z \bar{F}_z)\, dV \\ = -\,j\omega \displaystyle\int_\Omega J^{imp}_\rho \bar{F}_\rho + J^{imp}_z \bar{F}_z\, dV + j\omega \int_{\Gamma_N} J^{imp}_{\rho,\Gamma_N}\bar{F}_\rho + J^{imp}_{z,\Gamma_N}\bar{F}_z\, dS \\ - \displaystyle\int_\Omega \frac{1}{\mu} M^{imp}_\phi \left[\frac{\partial \bar{F}_\rho}{\partial z} - \frac{\partial \bar{F}_z}{\partial \rho}\right] dV \quad \forall\, \mathbf{F} = (F_\rho, 0, F_z) \in \tilde{H}_D(\mathbf{curl};\Omega), \end{array}\right. \tag{3.16}$$

where $\tilde{H}_D(\mathbf{curl};\Omega) = \{(E_\rho, E_z) : \mathbf{E} = (E_\rho, 0, E_z) \in L^2(\Omega),\ (\boldsymbol{\nabla}\times\mathbf{E})|_\phi = \frac{\partial E_\rho}{\partial z} - \frac{\partial E_z}{\partial \rho} \in L^2(\Omega),\ (\mathbf{n}\times\mathbf{E})|_{\Gamma_D} = 0\}$.

In summary, problem (3.9) decouples into a system of two simpler problems described by (3.15) and (3.16).

Remark 1. It has been shown in [4, Lemma 4.9] that space $\tilde{H}^1_D(\Omega)$ can also be expressed as $\tilde{H}^1_D(\Omega) = \{E_\phi \in L^2(\Omega) \ :\ \frac{1}{\rho}E_\phi \in L^2(\Omega),\ \nabla_{(\rho,z)}E_\phi \in L^2(\Omega)\}$.

3.4.2. H-formulation. Using the same decomposition of test functions (i.e., $\mathbf{F} = (0, F_\phi, 0)$, and $\mathbf{F} = (F_\rho, 0, F_z)$) for variational problem (3.12), we arrive at the following two decoupled variational problems in terms of $(0, H_\phi, 0)$ (3.17) and $(H_\rho, 0, H_z)$ (3.18), respectively:

$$\left\{\begin{array}{l} \text{Find } H_\phi \in H_{\phi,D} + \tilde{H}^1_D(\Omega) \text{ such that} \\ \displaystyle\int_\Omega \frac{1}{\sigma + j\omega\epsilon}\left(\frac{\partial H_\phi}{\partial z}\frac{\partial \bar{F}_\phi}{\partial z} + \frac{1}{\rho^2}\frac{\partial(\rho H_\phi)}{\partial \rho}\frac{\partial(\rho \bar{F}_\phi)}{\partial \rho}\right) dV \\ \quad + j\omega \displaystyle\int_\Omega \mu H_\phi\, \bar{F}_\phi dV = -\int_\Omega M^{imp}_\phi\, \bar{F}_\phi\, dV + \int_{\Gamma_N} M^{imp}_{\phi,\Gamma_N}\, \bar{F}_\phi\, dS \\ \quad + \displaystyle\int_\Omega \frac{1}{\sigma + j\omega\epsilon}\left[-J^{imp}_\rho \frac{\partial \bar{F}_\phi}{\partial z} + J^{imp}_z \frac{1}{\rho}\frac{\partial(\rho \bar{F}_\phi)}{\partial \rho}\right] dV \quad \forall\, F_\phi \in \tilde{H}^1_D(\Omega)\ . \end{array}\right. \tag{3.17}$$

$$
\left\{
\begin{array}{l}
\text{Find } \mathbf{H} = (H_\rho, 0, H_z) \in \mathbf{H}_D + \tilde{H}_D(\mathbf{curl}; \Omega) \text{ such that} \\
\displaystyle\int_\Omega \frac{1}{\sigma + j\omega\epsilon} \left(\frac{\partial H_\rho}{\partial z} - \frac{\partial H_z}{\partial \rho} \right) \left(\frac{\partial \bar{F}_\rho}{\partial z} - \frac{\partial \bar{F}_z}{\partial \rho} \right) dV \\
\displaystyle + j\omega \int_\Omega \mu (H_\rho \bar{F}_\rho + H_z \bar{F}_z)\, dV = - \int_\Omega M_\rho^{imp} \bar{F}_\rho + M_z^{imp} \bar{F}_z \, dV \\
\displaystyle + \int_{\Gamma_N} M_{\rho,\Gamma_N}^{imp} \bar{F}_\rho + M_{z,\Gamma_N}^{imp} \bar{F}_z \, dS + \int_\Omega \frac{1}{\sigma + j\omega\epsilon} J_\phi^{imp} \left[\frac{\partial \bar{F}_\rho}{\partial z} - \frac{\partial \bar{F}_z}{\partial \rho} \right] dV \\
\qquad\qquad \forall\, \mathbf{F} = (F_\rho, 0, F_z) \in \tilde{H}_D(\mathbf{curl}; \Omega) \, .
\end{array}
\right. \tag{3.18}
$$

From the formulation of problems (3.15) through (3.18), we remark the following:

- Physically, solutions of problems (3.16) and (3.17) correspond to the TE_ϕ-mode (i.e., $E_\phi = 0$), and solutions of problems (3.15) and (3.18) correspond to the TM_ϕ-mode (i.e., $H_\phi = 0$).
- The axis of symmetry is not a boundary of the original 3D problem, and therefore, a BC should not be needed to solve this problem. Nevertheless, formulations of problems (3.15) through (3.18) require the use of spaces $\tilde{H}_D^1(\Omega)$ and $\tilde{H}_D(\mathbf{curl}; \Omega)$ described above. The former space involves the singular weight $\frac{1}{\rho}$, which implicitly requires a homogeneous Dirichlet BC along the axis of symmetry. The latter space can be considered as it is (by using 2D edge elements), and no BC is necessary[4] to solve the problem.

4. Self-adaptive goal-oriented *hp*-FEM. We are interested in solving variational problems (3.9) and (3.12) (or alternatively, (3.15), (3.16), (3.17), and (3.18)), which we state here in terms of sesquilinear form b and antilinear form f:

$$
\left\{
\begin{array}{l}
\text{Find } \mathbf{E} \in \mathbf{E}_D + \mathbf{V}, \\
b(\mathbf{E}, \mathbf{F}) = f(\mathbf{F}) \quad \forall \mathbf{F} \in \mathbf{V} \, ,
\end{array}
\right. \tag{4.1}
$$

where

- $\mathbf{E}_D$ is a lift of the essential (Dirichlet) BC.
- $\mathbf{V}$ is a Hilbert space.
- $f \in \mathbf{V}'$ is an antilinear and continuous functional on $\mathbf{V}$.
- b is a sesquilinear form. We have

$$
b(\mathbf{E}, \mathbf{F}) = \left\{
\begin{array}{ll}
\underbrace{\displaystyle\int_\Omega \frac{1}{\mu} (\boldsymbol{\nabla} \times \mathbf{E}) \cdot (\boldsymbol{\nabla} \times \bar{\mathbf{F}})\, dV}_{a^E(\mathbf{E},\mathbf{F})} - \underbrace{\displaystyle\int_\Omega k^2 \mathbf{E} \cdot \bar{\mathbf{F}}\, dV}_{c^E(\mathbf{E},\mathbf{F})} & \text{E-Form}, \\
\underbrace{\displaystyle\int_\Omega \frac{1}{k^2} (\boldsymbol{\nabla} \times \mathbf{E}) \cdot (\boldsymbol{\nabla} \times \bar{\mathbf{F}})\, dV}_{a^H(\mathbf{E},\mathbf{F})} - \underbrace{\displaystyle\int_\Omega \mu \mathbf{E} \cdot \bar{\mathbf{F}}\, dV}_{c^H(\mathbf{E},\mathbf{F})} & \text{H-Form},
\end{array}
\right. \tag{4.2}
$$

where sesquilinear forms a^E, a^H, c^E, and c^H are Hermitian, continuous, and

[4] From the computational point of view, this effect can be achieved by artificially *adding* a homogeneous natural (Neumann) BC.

semipositive definite. We define an "energy" inner product on $\mathbf{V}$ as

(4.3)
$$(\mathbf{E},\mathbf{F}) := \begin{cases} \underbrace{\int_\Omega \frac{1}{\mu}(\boldsymbol{\nabla}\times\mathbf{E})\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})\,dV}_{a^E(\mathbf{E},\mathbf{F})} + \underbrace{\int_\Omega |k^2|\mathbf{E}\cdot\bar{\mathbf{F}}\,dV}_{c^E(\mathbf{E},\mathbf{F})} & \text{E-Form,} \\ \underbrace{\int_\Omega \frac{1}{|k^2|}(\boldsymbol{\nabla}\times\mathbf{E})\cdot(\boldsymbol{\nabla}\times\bar{\mathbf{F}})\,dV}_{a^H(\mathbf{E},\mathbf{F})} + \underbrace{\int_\Omega \mu\mathbf{E}\cdot\bar{\mathbf{F}}\,dV}_{c^H(\mathbf{E},\mathbf{F})} & \text{H-Form,} \end{cases}$$

with the corresponding (energy) norm denoted by $\|\mathbf{E}\|$. Notice the inclusion of the material properties in the definition of the norm.

4.1. Representation of the error in the quantity of interest. Given an *hp*-FE subspace $\mathbf{V}_{hp} \subset \mathbf{V}$, we discretize (4.1) as follows:

$$\begin{cases} \text{Find } \mathbf{E}_{hp} \in \mathbf{E}_D + \mathbf{V}_{hp}, \\ b(\mathbf{E}_{hp},\mathbf{F}_{hp}) = f(\mathbf{F}_{hp}) \quad \forall \mathbf{F}_{hp} \in \mathbf{V}_{hp}\ . \end{cases} \tag{4.4}$$

The objective of goal-oriented adaptivity is to construct an optimal *hp*-grid, in the sense that it minimizes the problem size needed to achieve a given tolerance error for a given *quantity of interest* L, with L denoting a linear and continuous functional. By recalling the linearity of L, we have

$$\text{Error of interest} = L(\mathbf{E}) - L(\mathbf{E}_{hp}) = L(\mathbf{E}-\mathbf{E}_{hp}) = L(\mathbf{e}), \tag{4.5}$$

where $\mathbf{e} = \mathbf{E} - \mathbf{E}_{hp}$ denotes the error function. By defining the residual $\mathbf{r}_{hp} \in \mathbf{V}'$ as $\mathbf{r}_{hp}(\mathbf{F}) = f(\mathbf{F}) - b(\mathbf{E}_{hp},\mathbf{F}) = b(\mathbf{E}-\mathbf{E}_{hp},\mathbf{F}) = b(\mathbf{e},\mathbf{F})$, we look for the solution of the *dual problem*:

$$\begin{cases} \text{Find } \bar{\mathbf{W}} \in \mathbf{V}, \\ b(\mathbf{F},\mathbf{W}) = L(\mathbf{F}) \quad \forall \mathbf{F} \in \mathbf{V}. \end{cases} \tag{4.6}$$

Problem (4.6) has a unique solution in $\mathbf{V}$. The solution $\mathbf{W}$ is usually referred to as the *influence function*.

By discretizing (4.6) via, for example, $\mathbf{V}_{hp} \subset \mathbf{V}$, we obtain

$$\begin{cases} \text{Find } \bar{\mathbf{W}}_{hp} \in \mathbf{V}_{hp}, \\ b(\mathbf{F}_{hp},\mathbf{W}_{hp}) = L(\mathbf{F}_{hp}) \quad \forall \mathbf{F}_{hp} \in \mathbf{V}_{hp}\ . \end{cases} \tag{4.7}$$

Definition of the dual problem plus the Galerkin orthogonality for the original problem imply the final representation formula for the error in the quantity of interest, namely,

$$L(\mathbf{e}) = b(\mathbf{e},\mathbf{W}) = b(\mathbf{e},\underbrace{\mathbf{W}-\mathbf{F}_{hp}}_{\epsilon}) = \tilde{b}(\mathbf{e},\epsilon)\ .$$

At this point, $\mathbf{F}_{hp} \in \mathbf{V}_{hp}$ is arbitrary, and $\tilde{b}(\mathbf{e},\epsilon) = b(\mathbf{e},\bar{\epsilon})$ denotes the bilinear form corresponding to the original sesquilinear form.

Notice that, in practice, the dual problem is solved not for $\mathbf{W}$ but for its complex conjugate $\bar{\mathbf{W}}$, utilizing the bilinear form and *not* the sesquilinear form. The linear

system of equations is factorized only once, and the extra cost of solving (4.7) reduces to only one backward and one forward substitution (if a direct solver is used).

Once the error in the quantity of interest has been determined in terms of bilinear form $\tilde{b}$, we wish to obtain a sharp upper bound for $|L(\mathbf{e})|$ that depends upon the mesh parameters (element size h and order of approximation p) *only locally*. Then, a self-adaptive algorithm intended to minimize this bound will be defined.

First, using a procedure similar to the one described in [10], we approximate $\mathbf{E}$ and $\mathbf{W}$ with *fine grid* functions $\mathbf{E}_{\frac{h}{2},\, p+1}$, $\mathbf{W}_{\frac{h}{2},\, p+1}$, which have been obtained by solving the corresponding linear system of equations associated with the finite element subspace $\mathbf{V}_{\frac{h}{2},\, p+1}$. In the remainder of this article, $\mathbf{E}$ and $\mathbf{W}$ will denote the fine grid solutions of the direct and dual problems ($\mathbf{E} = \mathbf{E}_{\frac{h}{2},\, p+1}$, and $\mathbf{W} = \mathbf{W}_{\frac{h}{2},\, p+1}$, respectively), and we will restrict ourselves to discrete finite element spaces only.

Next, we bound the error in the quantity of interest by a sum of element contributions. Let b_K denote a contribution from element K to sesquilinear form b. It then follows that

$$|L(\mathbf{e})| = |b(\mathbf{e}, \boldsymbol{\epsilon})| \leq \sum_K |b_K(\mathbf{e}, \boldsymbol{\epsilon})| \,, \tag{4.8}$$

where summation over K indicates summation over elements.

4.2. Projection-based interpolation operator. Once we have a representation formula for the error in the quantity of interest in terms of the sum of element contributions given by (4.8), we wish to express this upper bound in terms of local quantities, i.e. in terms of quantities that *do not* vary globally when we modify the grid locally. For this purpose, we introduce the idea of *projection-based interpolation* operators.

First, in order to simplify the notation, we define the following three spaces of admissible solutions:

- $\mathbf{V} = H_D(\mathbf{curl}; \Omega)$,
- $\mathbf{V}^{2D} = \tilde{H}_D(\mathbf{curl}; \Omega)$, and
- $V^{1D} = \tilde{H}^1_D(\Omega)$.

The corresponding hp-finite element spaces will be denoted by $\mathbf{V}_{hp}$, $\mathbf{V}^{2D}_{hp}$, and V^{1D}_{hp}, respectively.

At this point, we introduce three *projection-based interpolation* operators that have been defined in [9, 8], and used in [10, 23] for the construction of the fully automatic energy-norm-based hp-adaptive algorithm:

- $\Pi^{curl,3D}_{hp} : \mathbf{V} \longrightarrow \mathbf{V}_{hp}$,
- $\Pi^{curl,2D}_{hp} : \mathbf{V}^{2D} \longrightarrow \mathbf{V}^{2D}_{hp}$, and
- $\Pi^{1D}_{hp} : V^{1D} \longrightarrow V^{1D}_{hp}$.

We shall also consider three Galerkin projection operators:

- $\mathbf{P}^{curl,3D}_{hp} : \mathbf{V} \longrightarrow \mathbf{V}_{hp}$,
- $\mathbf{P}^{curl,2D}_{hp} : \mathbf{V}^{2D} \longrightarrow \mathbf{V}^{2D}_{hp}$, and
- $P^{1D}_{hp} : V^{1D} \longrightarrow V^{1D}_{hp}$.

To further simplify the notation, we will utilize the unique symbol Π^{curl}_{hp} to denote all projection-based interpolation operators mentioned above. Depending upon the problem formulation (and corresponding space of admissible solutions), Π^{curl}_{hp} should be understood as $\Pi^{curl,3D}_{hp}$ for problems (3.9) and (3.12), $\Pi^{curl,2D}_{hp}$ for problems (3.16) and (3.18), or Π^{1D}_{hp} for problems (3.15) and (3.17). Similarly, we will use the unique

symbol $\mathbf{P}_{hp}^{curl}$ to denote either $\mathbf{P}_{hp}^{curl,3D}$, $\mathbf{P}_{hp}^{curl,2D}$, or P_{hp}^{1D}.

We define $\mathbf{E}_{hp} = \mathbf{P}_{hp}^{curl}\mathbf{E}$. Equation (4.8) then becomes

$$|L(\mathbf{e})| \leq \sum_K |b_K(\mathbf{E}, \boldsymbol{\epsilon})| = \sum_K |b_K(\mathbf{E} - \Pi_{hp}^{curl}\mathbf{E}, \boldsymbol{\epsilon}) + b_K(\Pi_{hp}^{curl}\mathbf{E} - \mathbf{P}_{hp}^{curl}\mathbf{E}, \boldsymbol{\epsilon})| \,. \tag{4.9}$$

Given an element K, we conjecture that $|b_K(\Pi_{hp}^{curl}\mathbf{E} - \mathbf{P}_{hp}\mathbf{E}, \boldsymbol{\epsilon})|$ will be negligible compared to $|b_K(\mathbf{E} - \Pi_{hp}^{curl}\mathbf{E}, \boldsymbol{\epsilon})|$. Under this assumption, we conclude that

$$|L(\mathbf{e})| \preceq \sum_K |b_K(\mathbf{E} - \Pi_{hp}^{curl}\mathbf{E}, \boldsymbol{\epsilon})| \,. \tag{4.10}$$

In particular, for $\boldsymbol{\epsilon} = \mathbf{W} - \Pi_{hp}^{curl}\mathbf{W}$, we have

$$|L(\mathbf{e})| \preceq \sum_K |b_K(\mathbf{E} - \Pi_{hp}^{curl}\mathbf{E}, \mathbf{W} - \Pi_{hp}^{curl}\mathbf{W})| \,. \tag{4.11}$$

By applying the Cauchy–Schwarz inequality, we obtain the next upper bound for $|L(\mathbf{e})|$:

$$|L(\mathbf{e})| \preceq \sum_K \|\tilde{\mathbf{e}}\|_K \|\tilde{\boldsymbol{\epsilon}}\|_K, \tag{4.12}$$

where $\tilde{\mathbf{e}} = \mathbf{E} - \Pi_{hp}^{curl}\mathbf{E}$, $\tilde{\boldsymbol{\epsilon}} = \mathbf{W} - \Pi_{hp}^{curl}\mathbf{W}$, and $\|\cdot\|_K$ denotes energy-norm $\|\cdot\|$ restricted to element K.

4.3. Fully automatic goal-oriented *hp*-refinement algorithm. We describe an *hp* self-adaptive algorithm that utilizes the main ideas of the fully automatic (energy-norm-based) *hp*-adaptive algorithm described in [10, 23]. We start by recalling the main objective of the self-adaptive (energy-norm-based) *hp*-refinement strategy, which consists of solving the following maximization problem:

$$\begin{cases} \text{Find an optimal } \tilde{hp}\text{-grid in the following sense:} \\ \tilde{hp} = \arg\max_{\widehat{hp}} \sum_K \dfrac{\|\mathbf{E} - \Pi_{hp}^{curl}\mathbf{E}\|_K^2 - \|\mathbf{E} - \Pi_{\widehat{hp}}^{curl}\mathbf{E}\|_K^2}{\Delta N} \,, \end{cases} \tag{4.13}$$

where

- $\mathbf{E} = \mathbf{E}_{\frac{h}{2},\, p+1}$ is the *fine grid* solution, and
- $\Delta N > 0$ is the increment in the number of unknowns from grid hp to grid $\widehat{hp}$.

Similarly, for goal-oriented *hp*-adaptivity, we propose the following algorithm based on estimate (4.12):

$$\begin{cases} \text{Find an optimal } \tilde{hp}\text{-grid in the following sense:} \\ \tilde{hp} = \arg\max_{\widehat{hp}} \sum_K \left[\dfrac{\|\mathbf{E} - \Pi_{hp}^{curl}\mathbf{E}\|_K \cdot \|\mathbf{W} - \Pi_{hp}^{curl}\mathbf{W}\|_K}{\Delta N} \right. \\ \qquad\qquad \left. - \dfrac{\|\mathbf{E} - \Pi_{\widehat{hp}}^{curl}\mathbf{E}\|_K \cdot \|\mathbf{W} - \Pi_{\widehat{hp}}^{curl}\mathbf{W}\|_K}{\Delta N} \right] \,, \end{cases} \tag{4.14}$$

where

- $\mathbf{E} = \mathbf{E}_{\frac{h}{2},\, p+1}$ and $\mathbf{W} = \mathbf{W}_{\frac{h}{2},\, p+1}$ are the *fine grid* solutions corresponding to the direct and dual problems, and
- $\Delta N > 0$ is the increment in the number of unknowns from grid hp to grid $\widehat{hp}$.

Implementation of the goal-oriented hp-adaptive algorithm is based on the optimization procedure used for energy-norm hp-adaptivity [10, 23], which utilizes a multistep approach (first optimization of edges, and then optimization of interior degrees of freedom). The subspace associated to an optimal finite element grid is always contained in the subspace associated with the finite element *fine* grid computed during the previous step.

4.4. Implementation details. In what follows, we discuss the main implementation details needed to extend the fully automatic (energy-norm-based) hp-adaptive algorithm [10, 23] to a fully automatic goal-oriented hp-adaptive algorithm.

1. First, the solution $\mathbf{W}$ of the dual problem on the fine grid is necessary. This goal can be attained either by using a direct (frontal) solver or an iterative (two-grid) solver (see [18]).
2. Subsequently, we should treat both solutions as satisfying two different PDEs. We select functions $\mathbf{E}$ and $\mathbf{W}$ as the solutions of the system of two PDEs.
3. We proceed to redefine the evaluation of the error. The energy-norm error evaluation of a 2D function is replaced by the product $\| \mathbf{E} - \Pi_{hp}^{curl}\mathbf{E} \| \cdot \| \mathbf{W} - \Pi_{hp}^{curl}\mathbf{W} \|$.
4. After these simple modifications, the energy-norm-based self-adaptive algorithm may now be utilized as a self-adaptive goal-oriented hp algorithm.

5. Numerical results. In this section, we apply the goal-oriented hp self-adaptive strategy described in section 4 to simulate the response of the induction LWD instrument operating at 2 MHz considered in section 2.3, using formulation (3.15) for solenoidal coils and (3.17) for toroidal coils. Exactly the same results are obtained with formulations (3.18) and (3.16), respectively, as predicted by the theory. Thus, formulations (3.18) and (3.16) have been used as an extra verification of the simulations, and the corresponding results have been omitted in this article to avoid duplicity.

Figure 3 displays the first vertical difference of the electric field (divided by the distance between the two receivers) for the described LWD instrument equipped with solenoidal coils (left and center panel). The right panel corresponds to the computation of the normalized second vertical difference of the electric field when considering an extra receiving antenna 15 cm above the second receiving antenna. The three curves (two for the second vertical difference of the electric field) correspond to

1. the rock formation with no mud-filtrate invasion,
2. the rock formation with a 2 Ω·m 40 cm horizontal mud layer invading the 1 Ω·m rock formation layer, and a 5Ω·m 90cm horizontal mud layer invading the 10000 Ω·m rock formation layer, and
3. the previous (mud-invaded) rock formation, using a mandrel with relative magnetic permeability of 100.

For toroidal antennas, we display in Figure 4 the first vertical difference of the magnetic field (divided by the distance between the two receivers). The three displayed curves correspond to the three situations discussed above.

These results illustrate the strong dependence of the LWD response on the rock formation resistivity. We observe that solenoidal antennas are more sensitive to highly conductive formations as well as to the electrical permeability of the mandrel, while

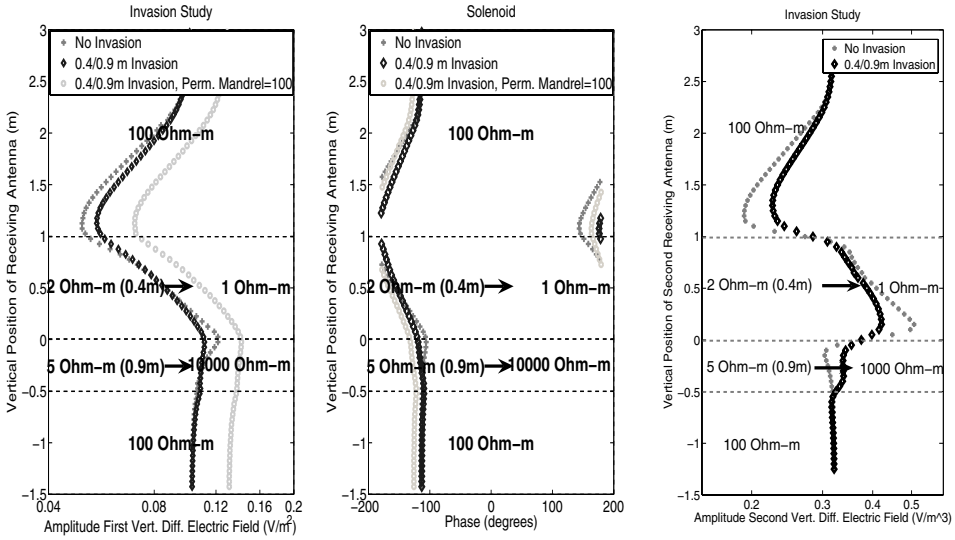


FIG. 3. *LWD problem equipped with a solenoidal source. Amplitude (left panel) and phase (center panel) of the first vertical difference of the electric field (divided by the distance between receivers) at the receiving coils. The normalized amplitude of the second vertical difference of the electric field is displayed in the right panel. Results obtained with the self-adaptive goal-oriented hp-FEM. The spatial distribution of electrical resistivity is also displayed to facilitate the physical interpretation of results.*

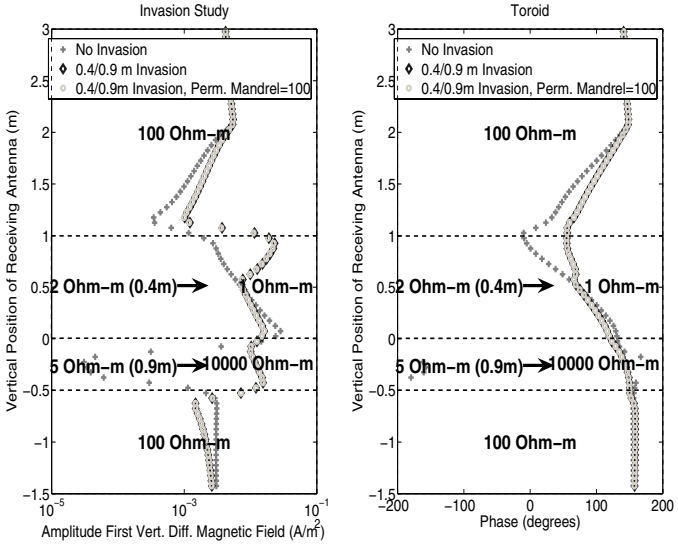


FIG. 4. *LWD problem equipped with a toroidal source. Amplitude (left panel) and phase (right panel) of the first vertical difference of the magnetic field (divided by the distance between receivers) at the receiving coils. Results obtained with the self-adaptive goal-oriented hp-FEM. The spatial distribution of electrical resistivity is also displayed to facilitate the physical interpretation of results.*

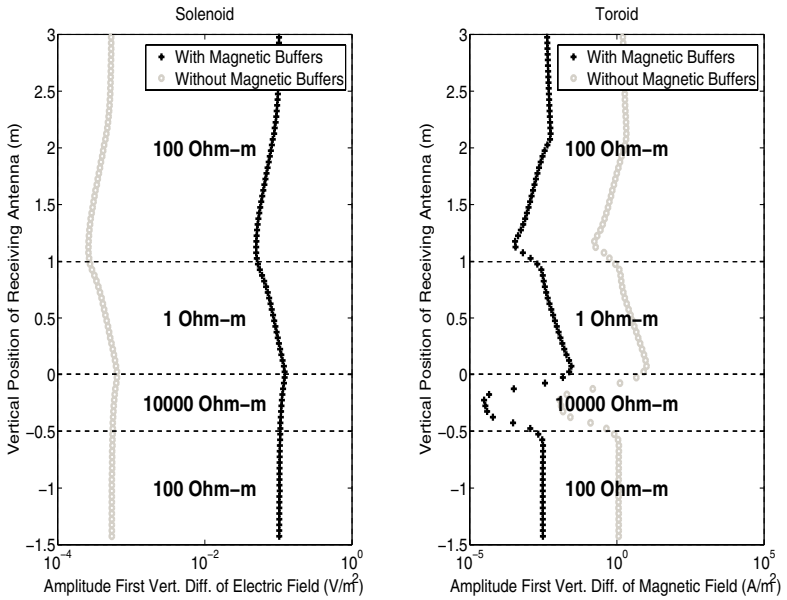


FIG. 5. *LWD problem equipped with a solenoidal source. Results obtained with the self-adaptive goal-oriented hp-FEM correspond to the use of solenoidal antennas (left panel), and toroidal antennas (right panel). The spatial distribution of electrical resistivity is also displayed to facilitate the physical interpretation of results.*

toroidal antennas are more sensitive to highly resistive formations. We also observe that the second vertical difference of the electric field is more sensitive to water invasion than the first vertical difference of the electric field (in both conductive and resistive formations).

Figure 5 illustrates the effect of the magnetic buffers. By removing the magnetic buffers from the logging instrument's design, the amplitude of the received signal decreases by a factor of up to 200 in the case of a solenoidal source. For practical applications, a strong signal on the receivers is desired to minimize the noise-to-signal ratio. Thus, it is appropriate to use magnetic buffers in combination with solenoidal antennas. In contrast, the use of magnetic buffers with toroidal antennas is not advisable since they weaken the received signal. In both cases, the phase and shape of the solution is not sensitive to the presence (or not) of magnetic buffers, and the corresponding results have been omitted.

The solver of linear equations utilized for these simulations is a multifrontal massively parallel sparse direct solver (MUMPS) [2, 1] running in a single-processor machine equipped with a Pentium IV 3.0 GHz processor. The total amount of time utilized by our FEM depends upon the choice of initial grid and the quantity of interest to be computed. Twelve minutes were needed to compute each curve—*log*—of Figure 5, composed of 80 points.

The exponential convergence obtained using the self-adaptive goal-oriented *hp*-FEM is shown in Figure 6 (left panel), by considering an arbitrary fixed position of the logging instrument for a solenoid antenna. The final grid delivers a relative error in the quantity of interest below 0.00001%; i.e., the first 7 significant digits of the quantity of interest are exact. In Figure 6 (right panel), we display the exponential

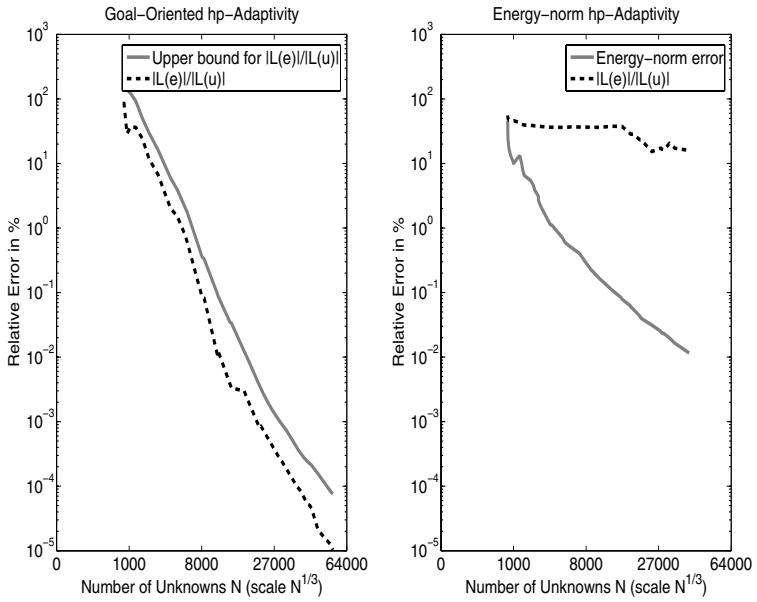


FIG. 6. *LWD problem equipped with a solenoidal source. Left panel: convergence behavior obtained with the self-adaptive goal-oriented hp-FEM shows exponential convergence rates for estimate* (4.8) *(solid curve) used for optimization. The dashed curve describes the relative error in the quantity of interest. Right panel: convergence behavior obtained with the self-adaptive energy-norm hp-FEM shows exponential convergence rates for the energy-norm. The dashed curve describes the relative error in the quantity of interest.*

convergence of the energy-norm-based *hp*-FEM. The final *hp*-grid delivers an energy-norm error below 0.01%. Nevertheless, the quantity of interest still contains a relative error above 15%.

A final goal-oriented *hp*-grid delivering a relative error in the quantity of interest of 0.1% is displayed in Figure 7.

6. Summary and conclusions. We have successfully applied a self-adaptive goal-oriented *hp*-FEM algorithm to simulate the axisymmetric response of an induction LWD instrument in a borehole environment. These simulations would not be possible with energy-norm adaptive algorithms. Also, the use of *hp*-FEM provides the flexibility needed to accurately approximate the solution within the formation (using the p method) as well as the strong singularities caused by the abrupt geometry of the mandrel (using the h method).

Numerical results illustrate the exponential convergence of the method (allowing for high accuracy simulations), the suitability of the presented formulations for axisymmetric electrodynamic problems, and the main physical characteristics of the presented induction LWD instrument. These results suggest the use of solenoidal antennas for the assessment of highly conductive rock formation materials, and toroidal antennas for the assessment of highly resistive materials. Solenoidal antennas should be used in combination with magnetic buffers to strengthen the measured EM signal, while the use of magnetic buffers with toroidal antennas should be avoided. Both types of antennas can be used to study mud-filtrate invasion. Second vertical differences of electromagnetic fields are more sensitive to mud-filtrate invasion than first

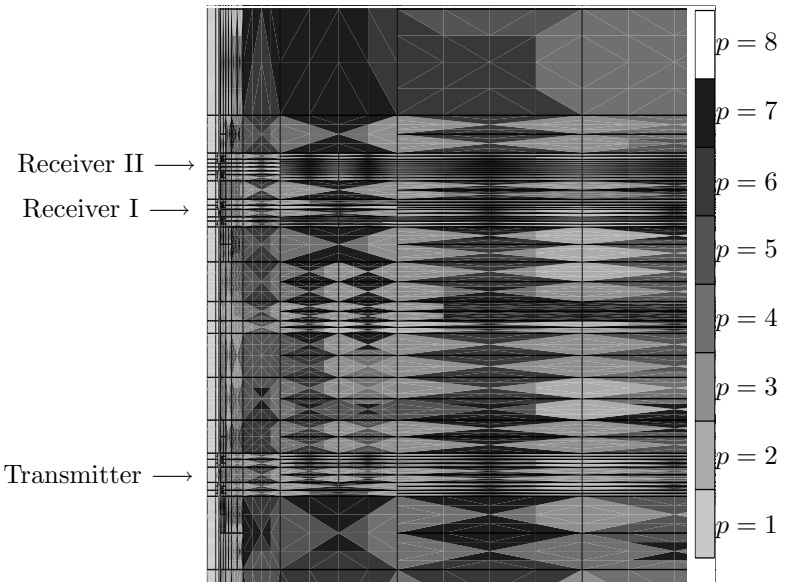


FIG. 7. *LWD instrument equipped with a solenoidal source. Portion (*120 *cm* × 200 *cm) of the final hp-grid. Different shades indicate different polynomial orders of approximation, ranging from* 1 *(light grey) to* 8 *(white).*

vertical differences.

Since the influence function used by the self-adaptive goal-oriented *hp*-adaptive algorithm is approximated via finite elements, the numerical method presented in this article is *problem independent*, and it can be applied to 1D, 2D, and 3D finite element discretizations of H^1-, $H(\mathbf{curl})$-, and $H(div)$-spaces.

Appendix. A loop-antenna radiating in a homogeneous lossy medium in the presence of a highly conductive metallic mandrel. In this appendix, we consider a problem with a known analytical solution. We use this problem as an additional mechanism to verify the code, as well as to provide comparative results between analytical and numerical solutions.

We consider a solenoid (or a toroid) of radius a radiating at a frequency of 2 MHz in a homogeneous lossy medium (with resistivity equal to $1\,\Omega\cdot\mathrm{m}$), in the presence of an infinitely large cylindrical mandrel (with resistivity equal to $10^{-6}\,\Omega\cdot\mathrm{m}$) of radius $b < a$. The coil and the mandrel exhibit axial symmetry (see Figure 8).

For a solenoidal coil located at $z = 0$, the resulting solution for $a \le \rho \le b$ is given by [15, 20]

$$E_\phi(\rho,z) = \frac{-\omega\mu}{4\pi a}\int_{-\infty}^{\infty}[J_1(k_\rho a) + \Gamma H_1^{(1)}(k_\rho a)]H_1^{(1)}(k_\rho\rho)e^{ik_z z}dk_\rho, \tag{A.1}$$

where $\Gamma = -J_1(k_\rho b)/H_1^{(1)}(k_\rho b)$, J_p and $H_p^{(1)}$ are the Bessel and Hankel functions, respectively, of the first type of order p, and $k_z = \sqrt{k^2 - k_\rho^2}$.

For a toroidal coil located at $z = 0$, the resulting solution for $a \le \rho \le b$ is given by

$$H_\phi(\rho,z) = \frac{-i}{4\pi a}\int_{-\infty}^{\infty}[J_1(k_\rho a) + \Gamma H_1^{(1)}(k_\rho a)]H_1^{(1)}(k_\rho\rho)e^{ik_z z}dk_\rho, \tag{A.2}$$

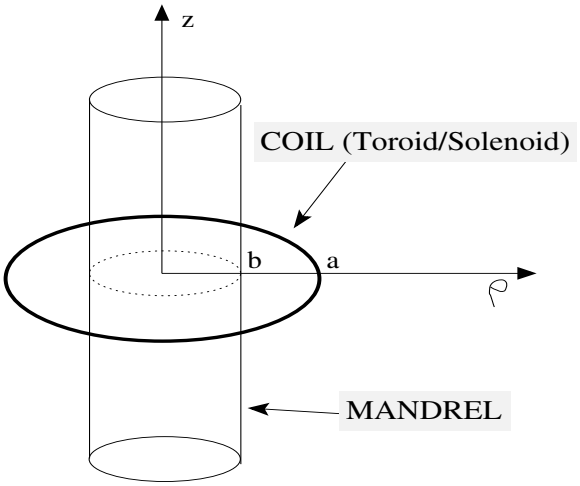


FIG. 8. *Geometry of a loop-antenna radiating in a homogeneous lossy medium in the presence of a highly conductive metallic mandrel.*

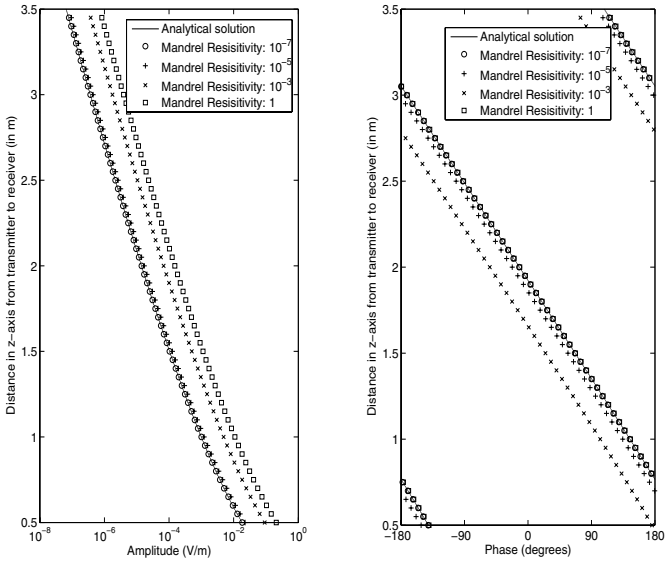


FIG. 9. *Solution (electric field) along the vertical axis passing through a solenoid radiating in a homogeneous medium in the presence of a metallic mandrel. Analytical solution (mandrel is a PEC) against the numerical solution for different mandrel resistivities (10^{-7}, 10^{-5}, 10^{-3}, and $1\ \Omega \cdot m$) obtained with the self-adaptive goal-oriented hp-FEM.*

where $\Gamma = -J_0(k_\rho b)/H_0^{(1)}(k_\rho b)$.

In Figures 9 and 10, we display a comparison between analytical and numerical results (obtained using the self-adaptive *hp* goal-oriented algorithm) for the solenoidal and toroidal coils, respectively. We selected $b = 0.0254$ cm and $a = 0.03048$ cm. The numerical results accurately reproduce the analytical ones, in terms of both amplitude

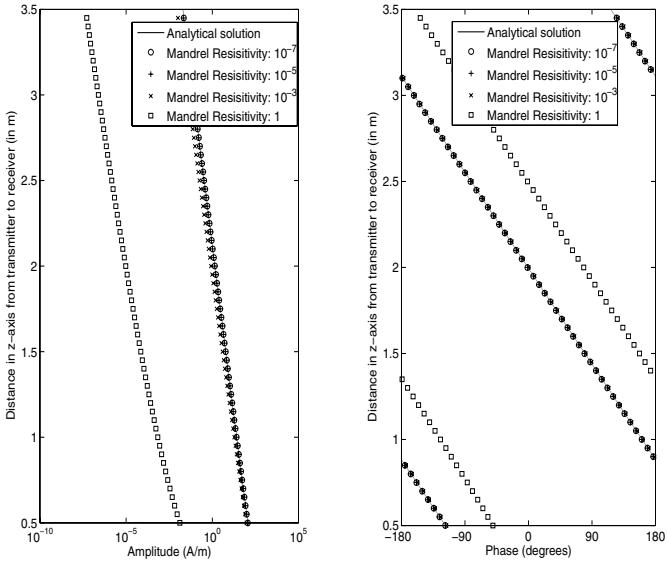


FIG. 10. *Solution (magnetic field) along a vertical axis passing through a toroid radiating in a homogeneous medium in the presence of a metallic mandrel. Analytical solution (mandrel is a PEC) against the numerical solution for different mandrel resistivities (10^{-7}, 10^{-5}, 10^{-3}, and $1\ \Omega \cdot m$) obtained with the self-adaptive goal-oriented hp-FEM.*

and phase.

When considering a solenoid, the logging instrument response using a mandrel of resistivity $10^{-5}\ \Omega \cdot \mathrm{m}$ or a PEC mandrel are indistinguishable in terms of amplitude. A similar situation occurs for a toroid. In terms of phase, induction instruments equipped with solenoidal coils appear to be more sensitive to the mandrel resistivity than those equipped with toroidal coils.

Acknowledgments. We would like to acknowledge the expertise and technical advice received from L. Tabarovsky, A. Bespalov, T. Wang, and other members of the Science Department of Baker-Atlas.

REFERENCES

[1] P. R. AMESTOY, I. S. DUFF, J.-Y. L'EXCELLENT, AND J. KOSTER, *A fully asynchronous multifrontal solver using distributed dynamic scheduling*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 15–41.

[2] P. R. AMESTOY, I. S. DUFF, AND J.-Y. L'EXCELLENT, *Multifrontal parallel distributed symmetric and unsymmetric solvers*, Comput. Methods Appl. Mech. Engrg., 184 (2000), pp. 501–520.

[3] B. I. ANDERSON, *Modeling and Inversion Methods for the Interpretation of Resistivity Logging Tool Response*, Ph.D. thesis, Delft University of Technology, Delft, The Netherlands, 2001; available online at http://repository.tudelft.nl/file/80838/161972.

[4] F. ASSOUS, C. CIARLET, JR., AND S. LABRUNIE, *Theoretical tools to solve the axisymmetric Maxwell equations*, Math. Methods Appl. Sci., 25 (2002), pp. 49–78.

[5] R. BECKER AND R. RANNACHER, *Weighted a posteriori error control in FE methods*, in ENUMATH 97, Proceedings of the 2nd European Conference on Numerical Mathematics and Advanced Applications, Heidelberg, Germany, 1997, H. G. Bock, F. Brezzi, R.

Glowinski, G. Kanschat, Y. A. Kuznetov, J. Periaux, and R. Rannacher, eds., World Scientific, Singapore, 1998, pp. 621–637.

[6] J. Berenger, *Three-dimensional perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 127 (1996), pp. 363–379.

[7] W. Cecot, W. Rachowicz, and L. Demkowicz, *An hp-adaptive finite element method for electromagnetics.* III: *A three-dimensional infinite element for Maxwell's equations*, Internat. J. Numer. Methods Engrg., 57 (2003), pp. 899–921.

[8] L. Demkowicz, *Finite element methods for Maxwell equations*, in Encyclopedia of Computational Mechanics 1, E. Stein, R. de Borst, and T. J. R. Hughes, eds., Wiley and Sons, New York, 2004, Chapter 26.

[9] L. Demkowicz and A. Buffa, H^1, $H(\mathbf{curl})$, *and* $H(div)$ *conforming projection-based interpolation in three dimensions: Quasi optimal p-interpolation estimates*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 267–296.

[10] L. Demkowicz, W. Rachowicz, and Ph. Devloo, *A fully automatic hp-adaptivity*, J. Sci. Comput., 17 (2002), pp. 117–142.

[11] R. F. Harrington, *Time-Harmonic Electromagnetic Fields*, McGraw-Hill, New York, 1961.

[12] V. Heuveline and R. Rannacher, *Duality-based adaptivity in the hp-finite element method*, J. Numer. Math., 11 (2003), pp. 95–113.

[13] Y. K. Hue and F. L. Teixeira, *Three-dimensional simulation of eccentric LWD tool response in boreholes through dipping formations*, IEEE Trans. Geosci. Remote Sensing, 43 (2005), pp. 257–268.

[14] J. Li and C. Liu, *A three-dimensional transmission line matrix method (TLM) for simulation of logging tools*, IEEE Trans. Geosci. Remote Sensing, 38 (2000), pp. 1522–1529.

[15] J. R. Lovell, *Finite Element Methods in Resistivity Logging*, Ph.D. thesis, Delft University of Technology, Delft, The Netherlands, 1993; available online at http://repository.tudelft.nl/file/264367/201280.

[16] J. T. Oden and S. Prudhomme, *Goal-oriented error estimation and adaptivity for the finite element method*, Comput. Math. Appl., 41 (2001), pp. 735–756.

[17] M. Paraschivoiu and A. T. Patera, *A hierarchical duality approach to bounds for the outputs of partial differential equations*, Comput. Methods Appl. Mech. Engrg., 158 (1998), pp. 389–407.

[18] D. Pardo, *Integration of hp-Adaptivity with a Two Grid Solver: Applications to Electromagnetics*, Ph.D. thesis, Department of Computational and Applied Mathematics, The University of Texas at Austin, Austin, TX, 2004.

[19] D. Pardo, L. Demkowicz, C. Torres-Verdin, and L. Tabarovsky, *A goal-oriented hp-adaptive finite element method with electromagnetic applications. Part* I: *Electrostatics*, Internat. J. Numer. Methods Engrg., 65 (2006), pp. 1269–1309.

[20] M. Paszynski, L. Demkowicz, and D. Pardo, *Verification of goal-oriented hp-adaptivity*, Comput. Math. Appl., 50 (2005), pp. 1395–1404.

[21] C. R. Paul and S. A. Nasar, *Introduction to Electromagnetic Fields*, McGraw-Hill, New York, 1982.

[22] S. Prudhomme and J. T. Oden, *On goal-oriented error estimation for elliptic problems: Application to the control of pointwise errors*, Comput. Methods Appl. Mech. Engrg., 176 (1999), pp. 313–331.

[23] W. Rachowicz, D. Pardo, and L. Demkowicz, *Fully automatic hp-adaptivity in three dimensions*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 4816–4842.

[24] R Rannacher and F. Suttmeier, *A posteriori error control in finite element methods via duality techniques: Application to perfect plasticity*, Comput. Mech., 21 (1998), pp. 123–133.

[25] P. Solin and L. Demkowicz, *Goal-oriented hp-adaptivity for elliptic problems*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 449–468.

[26] F. L. Teixeira and W. C. Chew, *Finite-difference computation of transient electromagnetic waves for cylindrical geometries in complex media*, IEEE Trans. Geosci. Remote Sensing, 38 (2000), pp. 1530–1543.

[27] J. Van Bladel, *Singular Electromagnetic Fields and Sources*, Oxford University Press, New York, 1991.

[28] T. Wang and J. Signorelli, *Finite-difference modeling of electromagnetic tool response for logging while drilling*, Geophys., 69 (2004), pp. 152–160.

[29] Z. Q. Zhang and Q. H. Liu, *Applications of the BCGS-FFT method to* 3*-D induction well logging problems*, IEEE Trans. Geosci. Remote Sensing, 41 (2003), pp. 998–1004.

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 2107–2126

ANALYSIS OF NONLOCAL ELECTROSTATIC EFFECTS IN CHIRAL SMECTIC C LIQUID CRYSTALS*

JINHAE PARK† AND M. CARME CALDERER†

Abstract. We present modeling and analysis of smectic C phases of liquid crystals capable of sustaining spontaneous polarization. The layered liquid crystals are also assumed to be chiral. We study minimization of the total energy subject to electrostatic constraints. In order to determine mathematically and physically relevant boundary conditions, we appeal to the analogy between the current problem and the vorticity in fluids. We place a special emphasis on the nonlocal and self-energy effects arising from spontaneous polarization. We discuss examples pertaining to the electric field created by the liquid crystal in dielectric medium, and also to the possible role of a domain shape as an energy reduction mechanism.

Key words. chiral, energy minimizer, electrostatic self-interaction, ferroelectricity, helical filaments, liquid crystals, nonlocal effects, smectic C phase, smectic layers, vortex tubes

AMS subject classifications. 35Q35, 47J30, 49J40, 49J45, 74A35, 76A15, 82B21, 82B26, 82D45

DOI. 10.1137/050641120

1. Introduction.

This article analyzes nonlocal electrostatic effects associated with polarized states of liquid crystals. We assume that the liquid crystals are of smectic type, possess spontaneous polarization, and may also be chiral. We study minimization of the total energy in $\mathbf{R}^3$, subject to electrostatic constraints.

In smectic liquid crystals, centers of mass of molecules are arranged locally in one-dimensional layers described by a complex field $\psi = \rho e^{i\omega}$; level sets of the phase function ω denote layer locations, with $\nabla\omega$ being parallel to the layer normal. Nonparallel unit vector fields $\mathbf{n}$ and $\mathbf{p}$ describe the orientational ordering of biaxial molecules. Another feature of smectic C phases is that the director $\mathbf{n}$ makes a preferred angle α with the layer normal vector. The angle α is a temperature- and material dependent quantity ranging typically from 0 to $\frac{\pi}{4}$. We visualize smectic C phases in terms of cones with axis along the layer normal and semiangle α. The director $\mathbf{n}$ is then parallel to a generating straight line of the cone. Since the systems that we consider are ferroelectric, that is, they have spontaneous polarization, we take $\mathbf{p}$ to be parallel to the polarization field $\mathbf{P}$; $\mathbf{n}$ corresponds to the uniaxial director measuring the average alignment of molecular long axes of either rod-like or bent-core molecules [33, 34]. The electrostatic potential φ is also a variable of the problem.

Many different types of liquid crystals form smectic C phases (i.e., one-dimensional layer structures). The earlier low molecular weight liquid crystals labeled as smectic C [25] owe ferroelectricity to the molecular dipoles associated with side chains. In such cases, the direction of $\mathbf{P}$ is determined and tends to be perpendicular to $\mathbf{n}$ and $\nabla\omega$. (This type of ferroelectricity, known as *improper* [21], is absent in the smectic A phase due to $\mathbf{n}$ being parallel to $\nabla\omega$.) Since the magnitude of the polarization is usually small, studies of such liquid crystals normally neglect nonlocal electric field

*Received by the editors September 24, 2005; accepted for publication (in revised form) June 15, 2006; published electronically October 20, 2006. This work is partially supported by the National Science Foundation, grant number DMS-0128832.

http://www.siam.org/journals/siap/66-6/64112.html

†School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (jinhae@math.umn.edu, park196@math.purdue.edu; mcc@math.umn.edu).

effects and assign constant values to applied fields. Many of the low molecular weight smectic C liquid crystals are cholesteric. On the other hand, B2 phases of bent-core molecule liquid crystals are mostly nonchiral and have large polarization values. In the B2 phase, the layers have a locally polar C_{2v} symmetry group in the Schoenflies symbols [40, p. 489]. Although $\mathbf{P}$ tends to be perpendicular to $\mathbf{n}$, the plane determined by these vectors is free to rotate about $\mathbf{n}$. Unlike the case of small molecule smectics, the B2 phases are not subject to the constraint of $\mathbf{P}$, that is, parallel to $\nabla\omega \times \mathbf{n}$. However, analysis of the latter requires accounting for nonlocal energy.

Throughout this work, we will assume that the liquid crystals are chiral and use the conventional notation C* to denote such chiral smectic phases. An important bulk configuration of chiral smectic C phases is that the variables $\mathbf{n}$ and $\mathbf{p}$ rotate spatially around the axis of the previously described cone, with $\mathbf{n}$ being parallel to a generating straight line of the cone. This accordingly results in zero net polarization over a helical pitch. In suppressing the helix by applying an external electric field or boundary conditions, ferroelectric states with opposite polarization emerge. The transition between the chiral state and the ferroelectric ones is not regarded as a typical phase transition; however, it is at the core of device applications of ferroelectric liquid crystals. Mathematically, we addressed in previous work [27] some stability properties with respect to boundary conditions and electric fields, in the case that nonlocal effects are neglected.

The total energy we analyze consists of nematic, smectic, Ginzburg–Landau, electrostatic, and surface contributions. The nematic and smectic free energy densities follow the forms of Oseen–Frank, de Gennes, and Chen–Lubensky, which penalize departure from preferred molecular alignment and orientation with respect to the layers. Other macroscopic theories of smectic C phases can be found in the literature [23, 33, 37]. Our choice is motivated by the covariant structure of the Lubensky form; this is quadratic in second order gradients of ψ and especially amenable to treatment by calculus of variations. We relax the constraint $\mathbf{p} \perp \{\mathbf{n}, \nabla\omega\}$ and instead incorporate a penalty energy into the model. The Ginzburg–Landau energy tends to select a preferred magnitude of the polarization according to temperature and material parameters. The current energy is appropriate to modeling B2 ferroelectric phases, provided that we omit the previously mentioned penalty term and chirality.

The electrostatic energy comprises a dielectric and a ferroelectric contribution. The latter accounts for the energy of self-interaction between the polarization and its own electric field, as well as the electrostatic energy outside the liquid crystal domain. So far, this situation is analogous to that of a ferroelectric solid. However, there are some fundamental differences between these two behaviors. In the solid, the directions of polarizations are determined by lattice directions; the latter are difficult to alter using an external field. The "softness" of a liquid crystal allows for changes in molecular alignments so as to reduce the energy. For instance, in the case of a liquid crystal located between conducting plates, the distribution of polarization is such that it gets compensated by free charges in the conductor, inducing a zero electric field outside. As a result, the nonlocal energy vanishes. Suppressing the helix in the smectic C* phase is a mechanism of reducing nonlocal energy. In a related work [19], Khachaturyan showed that a polarized homogeneous state of the nematic phase is unstable with respect to a slight dipole-dipole interaction, resulting in a symmetry breaking.

If a B2 liquid crystal is embedded in another liquid, such as its own isotropic phase, it may actually change its shape so that the nonlocal energy is zero. This gives a good explanation for the telephone-cord shape observed in many ferroelectric liquid

crystal filaments [5, 8, 16, 17]. Modeling of static liquid crystal helical filaments can be found in [2, 36]. Chevron structures with alternating domains of opposite polarization are also found in some materials [5]. The phenomenon of changing shape to reduce electrostatic energy has also been observed in thin polarized piezoelectric films (for example, ZnO nanobelt [20]).

In the current analysis, defects are not included. In particular, we assume that $\nabla\omega$ is defined everywhere on $\partial\Omega$ with possible exceptions on the edges of Ω. This, in turn, determines the type of the domain occupied by the liquid crystal, and the nature of boundary conditions on the phase function ω. We make use of analogies with the geometry of vortex tubes and sheets in fluid mechanics. Indeed, we take Ω to be a cylinder-like domain analogous to a vortex tube, which has the lateral surface Σ corresponding to a vortex sheet, and is contained between surfaces $\mathcal{S}_1$ and $\mathcal{S}_2$. We deal with two types of boundary conditions on ω: one corresponding to the layer structure reaching the boundary in a tangential fashion, and the other with layers being perpendicular to the boundary. The latter correspond to the geometry of the Clark–Lagerwall effect in ferroelectric displays [4, 14, 39]. We also prescribe the electrostatic potential on a part of the boundary. These boundary conditions together with assumptions for the constitutive parameters (see section 3) allow us to prove existence of minimizers of the total energy by using direct methods of calculus of variations. One important issue is whether the minimizers thus obtained correspond to chiral structures or ferroelectric ones. We apply asymptotic analysis to obtain a classification of minimizers [1].

In [18], Joo and Phillips studied the phase transitions between chiral nematic, smectic A*, and C* liquid crystals, and carried out extensive stability analyses. Their work gives a rigorous classification of the energy minimizing phase regimes. Another important merit of the article is establishing the coercivity of the smectic C* energy for the first time. For a mathematical analysis of the phase transitions between the chiral and smectic A* liquid crystals with focus on the analogies of the phase transition between conductor and superconductor, the reader is referred to [1]. Studies of periodic ferroelectric and antiferroelectric phases and analysis of time dependent problems arising in switching have also been carried out by the authors [28]. Experimental treatments and studies of smectic C* liquid crystals including the influence of an electric field are found in [10, 11, 12, 29, 30, 31]. For structural understanding and modeling of ferroelectricity, we refer to the books by Goody et al. [13], Lagerwall [21], Pikin [32], and by Muševič, Blinc, and Žekš [26].

This article is organized as follows. In section 2, we present free energy functions of smectic C* materials with concentration on the polarization and electrostatic energies. We discuss constraint relaxation and approximations of the electrostatic energies. In section 3, we prove existence of minimizers and study examples regarding the relationship between domain shapes and polarizations. We also present two different versions of the variational problem, with one of them corresponding to a liquid crystal placed between metallic plates. In the other formulation, the liquid crystal is placed in a dielectric medium, subject to the electric field generated by the material polarization. In section 4, we carry out asymptotic studies of the minimizers obtained in section 3 to determine whether they correspond to chiral or ferroelectric structures. We outline some conclusions in section 5.

2. Free energy functions. We present the energy functional for the smectic materials to be analyzed. This includes nematic, smectic, Ginzburg–Landau, surface, and electric contributions. We will also show how they give rise to simple forms found

in the literature.

Equilibrium configurations of smectic C* liquid crystals occupying a smooth domain Ω in $\mathbf{R}^3$ are given by quadruples $(\psi, \mathbf{n}, \mathbf{P}, \mathbf{E})$ of fields, $\psi : \Omega \to \mathbb{C}$, $\mathbf{n} : \Omega \to \mathbf{S}^2$, $\mathbf{P} : \Omega \to \mathbf{R}^3$, and $\mathbf{E} : \mathbf{R}^3 \to \mathbf{R}^3$, that are critical points of the total energy functional

$$\begin{aligned}\mathcal{E}(\mathbf{n}, \mathbf{P}, \psi, \mathbf{E}) = \int_\Omega \{F_N(\mathbf{n}, \mathbf{P}, \nabla \mathbf{n}) + F_{Sm}(\nabla \mathbf{n}, \nabla \psi, \nabla^2 \psi) \\ + F_P(\mathbf{n}, \mathbf{P}, \nabla \mathbf{P}, \nabla \psi)\}\, d\mathbf{x} + \int_{\partial\Omega} F_S(\mathbf{n}, \nu)\, dS \\ + \int_{\mathbf{R}^3} F_E(\mathbf{n}\chi_\Omega, \mathbf{P}\chi_\Omega, \mathbf{E}) d\mathbf{x},\end{aligned} \tag{2.1}$$

subject to Maxwell's equations

$$-\nabla \cdot \mathbf{D} = 0, \quad \nabla \times \mathbf{E} = \mathbf{0} \quad \text{in } \Omega, \tag{2.2}$$

where $\mathbf{D}$ is the electric displacement vector, $\mathbf{E}$ is the electric field, and the functions F_N, F_{Sm}, F_E, F_S, F_P represent the Oseen–Frank, the smectic, the electrostatic, the surface anchoring [6, p. 99], and the Ginzburg–Landau energy densities, respectively. χ_Ω denotes the characteristic function.

2.1. Dielectric, nonlocal, and self-interaction terms. The electrostatic energy density in $\mathbf{R}^3$ [7, 22] is given by[1]

$$F_E = -\frac{1}{2}\big((\varepsilon \mathbf{E} \cdot \mathbf{E})\chi_\Omega + (\mathbf{E} \cdot \mathbf{E})\chi_{\Omega^c}\big) - (\mathbf{P} \cdot \mathbf{E})\chi_\Omega, \tag{2.3}$$

$$\mathbf{D} = \varepsilon \mathbf{E}\chi_\Omega + \mathbf{E}\chi_{\Omega^c} + \mathbf{P}\chi_\Omega, \tag{2.4}$$

$$\varepsilon = \varepsilon_\perp \mathbf{I} + \varepsilon_a \mathbf{n} \otimes \mathbf{n}, \tag{2.5}$$

where ε, $\varepsilon_\perp$, and ε_a represent the susceptibility tensor, dielectric permittivity, and dielectric anisotropy, respectively.

Note that (2.3) can be written as

$$F_E = -\frac{1}{2}\Big[\varepsilon_\perp |\mathbf{E}|^2 + \varepsilon_a (\mathbf{n} \cdot \mathbf{E})^2 + |\mathbf{E}|^2 \chi_{\Omega^c}\Big] - (\mathbf{P} \cdot \mathbf{E})\chi_\Omega. \tag{2.6}$$

So, the electrostatic energy is given by

$$\begin{aligned}\int_{\mathbf{R}^3} F_E\, d\mathbf{x} = -\frac{1}{2}\int_\Omega \{\varepsilon_\perp |\mathbf{E}|^2 + \varepsilon_a (\mathbf{E} \cdot \mathbf{n})^2\}\, d\mathbf{x} \\ - \frac{1}{2}\int_{\mathbf{R}^3 - \Omega} |\mathbf{E}|^2\, d\mathbf{x} - \int_\Omega \mathbf{P} \cdot \mathbf{E}\, d\mathbf{x}.\end{aligned} \tag{2.7}$$

The two terms in the last row in (2.7) correspond to the self-interaction and nonlocal electrostatic energies, respectively. The term in the first row gives the dielectric contribution.

[1] We note that the electric displacement $\mathbf{D}$ is usually written as $\mathbf{D} = \varepsilon \mathbf{E}\chi_\Omega + \mathbf{E}\chi_{\Omega^c} + 4\pi \mathbf{P}$. For simplicity, we replace $4\pi \mathbf{P}$ by $\mathbf{P}$.

2.2. Nematic and smectic free energies. The Oseen–Frank free energy density is given by

$$F_N = K_1(\nabla\cdot\mathbf{n})^2 + K_2(\mathbf{n}\cdot\nabla\times\mathbf{n}+\tau)^2 + K_3|\mathbf{n}\times(\nabla\times\mathbf{n})+\gamma\mathbf{P}|^2 + (K_2+K_4)(tr(\nabla\mathbf{n})^2-(\nabla\cdot\mathbf{n})^2), \tag{2.8}$$

where K_i, $i=1,2,3,4$, denote elasticity constants. The scalar τ represents the chiral pitch of the helical structure of the cholesteric phases [9], and $K_3\gamma^2|\mathbf{P}|^2$ is an intrinsic bending stress [7, p. 384]. Here, γ is a parameter included for the purpose of dimensionalization; hereafter, we will take γ to be 1. Such a term appears only in connection with the modeling of the smectic C* since nematics with intrinsic bending have not been observed. Both quantities result from the loss of mirror symmetry of the smectic C* phases. The fourth term in $F_{\rm N}$ is a null-Lagrangian; its integral is determined by $\mathbf{n}$ on $\partial\Omega$. The classical Oseen–Frank energy corresponds to the case that $\mathbf{P}$ is zero. Existence and regularity of minimizers for the classical Oseen–Frank energy were studied by Hardt, Kinderlehrer, and Lin [15].

The free energy density associated with the smectic layering follows the covariant form presented in [3]:

$$F_{Sm} = D(\mathbb{D}^2\psi)(\mathbb{D}^2\psi)^* + [C_{||}n_in_j + C_\perp(\delta_{ij}-n_in_j)](\mathbb{D}_i\psi)(\mathbb{D}_j\psi)^* + r|\psi|^2 + \frac{g}{2}|\psi|^4, \tag{2.9}$$

with $\mathbb{D}\equiv\nabla - iq\mathbf{n}$, q the modulation wave number of the smectic layer, and $r = a(T-T^*)$, $a>0$; here T denotes the (constant) temperature of the material and T^* is the transition temperature from nematic to smectic. Model (2.9) yields the de Gennes model for smectic A* when $C_{||}-C_\perp = 0$ and $D=0$. The smectic C phase is characterized by $C_\perp<0$. Moreover, $C_\perp\geq 0$ in the smectic A*, and $C_\perp = 0$ characterizes the transition to smectic C. Equivalently, the energy (2.9) can also be written as follows:

$$F_{Sm} = D|\mathbb{D}^2\psi|^2 + C_\perp|\mathbb{D}\psi|^2 + C_a|\mathbf{n}\cdot\mathbb{D}\psi|^2 + r|\psi|^2 + \frac{g}{2}|\psi|^4. \tag{2.10}$$

Remark. The first term in (2.9) is obtained from [24] and is a modification of $D_\perp(\delta_{ij}-n_in_j)(\delta_{kl}-n_kn_l)(\mathbb{D}_i\mathbb{D}_j)(\mathbb{D}_k\mathbb{D}_l)^*$ in the original Chen–Lubensky model. The purpose of introducing the new term is to obtain coercivity of the energy. This fact was first observed in [18].

2.3. Anchoring energy. The anchoring energy is the Rapini–Papoular surface energy [7, 21] given by

$$F_S = \omega_n\big(1-\alpha_0(\mathbf{n}\cdot\nu)^2\big), \tag{2.11}$$

where ω_n and $|\alpha_0|<1$ are material constants, and ν denotes the unit normal to the surface. Note that the surface energy due to the polarization is not explicitly included in (2.11). In fact, the role of such a surface energy is an approximation to the nonlocal energy in (2.7), which we explicitly include in the problem.

2.4. Ginzburg–Landau energy with relaxation. The energy associated with the phase transition to the ferroelectric phases is given by the Ginzburg–Landau expression [22], $a_0|\mathbf{P}|^2 + b_0|\mathbf{P}|^4$, where $b_0>0$ and $a_0 = \alpha(T-T_c)$. Ferroelectric phases correspond to the case $T<T_c$.

In contrast to solids, the direction of polarization in many liquid crystals is determined by the direction $\mathbf{n}$ and the layer normal $\nabla\omega$. In fact, a symmetry argument shows that $\mathbf{P}$ is perpendicular to both $\mathbf{n}$ and $\nabla\omega$. For this reason, we express $\mathbf{P}$ as follows [7, p. 384]:

$$\mathbf{P} = \begin{cases} |\mathbf{P}| \frac{\nabla\omega\times\mathbf{n}}{|\nabla\omega\times\mathbf{n}|} & \text{if } \nabla\omega\times\mathbf{n} \neq \mathbf{0}, \\ \mathbf{0} & \text{if } \nabla\omega\times\mathbf{n} = \mathbf{0}. \end{cases} \tag{2.12}$$

Therefore, the Ginzburg–Landau energy with relaxation is given by

$$F_P = B|\nabla\mathbf{P}|^2 + a_0|\mathbf{P}|^2 + b_0|\mathbf{P}|^4 + \frac{1}{\epsilon^2}|(|\nabla\omega\times\mathbf{n}|)\mathbf{P} - |\mathbf{P}|(\nabla\omega\times\mathbf{n})|^2, \tag{2.13}$$

where $B > 0$, $a_0 < 0$, $b_0 > 0$, and $\epsilon > 0$. Here, we note that the last term in (2.13), $\frac{1}{\epsilon^2}|(|\nabla\omega\times\mathbf{n}|)\mathbf{P} - |\mathbf{P}|(\nabla\omega\times\mathbf{n})|^2$, is a penalty term for (2.12), and $|\nabla\mathbf{P}|^2$ is a regularizing term.

2.5. Electrostatic approximations. The presence of polarization in the sample causes a point charge density $\rho_p = -\nabla\cdot\mathbf{P}$ in the bulk, and $\sigma = \mathbf{P}\cdot\nu$, where σ is the surface density of charges [38]. This will help us interpret the energies $\int_{\mathbf{R}^3-\Omega}|\mathbf{E}|^2\,d\mathbf{x}$ and $\int_\Omega \mathbf{P}\cdot\mathbf{E}\,d\mathbf{x}$. For this, let us consider the special case that Ω is a ball of radius r_0 centered at $\mathbf{0}$, with the constant surface charge density σ. We calculate the electric potential φ [38] as

$$\varphi(\mathbf{x}) = \begin{cases} \frac{\sigma r_0^2}{|\mathbf{x}|} & \text{if } |\mathbf{x}| > r_0, \\ \sigma r_0 & \text{if } |\mathbf{x}| \le r_0. \end{cases}$$

Then for $|\mathbf{x}| > r_0$ we get

$$|\mathbf{E}(r)| = \frac{\sigma r_0^2}{r^2},$$

and hence

$$\begin{aligned} \int_{\mathbb{R}^3-\Omega}|\mathbf{E}|^2\,d\mathbf{x} &= \int_0^{2\pi}\int_0^{\pi}\int_{r_0}^{\infty}\frac{\sigma^2 r_0^4}{r^2}\sin\phi\,dr\,d\phi\,d\theta \\ &= r_0\int_{\partial\Omega}(\mathbf{P}\cdot\nu)^2\,dS. \end{aligned}$$

This explains why the term $\int_{\partial\Omega}(\mathbf{P}\cdot\nu)^2\,dS$ appears in liquid crystal models [21].

Now, let us consider the self-interaction energy. For our illustration, we consider the case that $\varepsilon_a \ll \varepsilon_\perp$, and we neglect ε_a in the model. It then follows that the self-interaction energy is related with the Coulomb energy by

$$\int_\Omega\int_\Omega \frac{\nabla\cdot\mathbf{P}(\mathbf{x})\nabla\cdot\mathbf{P}(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}\,d\mathbf{x}\,d\mathbf{y} = -4\pi\varepsilon_\perp\int_\Omega \mathbf{P}\cdot\mathbf{E}\,d\mathbf{x}.$$

We point out that the Coulomb energy is often approximated by the polar energy $\int_\Omega(\nabla\cdot\mathbf{P})^2\,d\mathbf{x}$ found in the literature [7, 21]. We summarize the total energy involved

in the electrostatic approximation without an external field as follows:

$$\mathcal{E}(\mathbf{n},\mathbf{P},\psi)=\int_\Omega\{F_N+F_{Sm}+G_P\}d\mathbf{x}+\int_{\partial\Omega}G_S\,dS, \tag{2.14}$$

$$\begin{aligned}G_S=\omega_p\left(1-\frac{\mathbf{P}\cdot\nu}{|\mathbf{P}|}\right)+\omega_r\left(1-\frac{(\mathbf{P}\cdot\nu)^2}{|\mathbf{P}|^2}\right)\\+\omega_n\Big(1-(\mathbf{n}\cdot\nu)^2\Big),\end{aligned} \tag{2.15}$$

$$G_P=F_P+B_1(\nabla\cdot\mathbf{P})^2. \tag{2.16}$$

Minimization of the energy (2.14) is studied in previous work [27].

3. Existence of minimizers. In this section, we study the boundary conditions for smectic C* layer configurations and prove existence of minimizers. We also discuss applications and provide examples that explain the relationship between domain shape and ferroelectricity.

Throughout this paper, we assume that the constitutive parameters satisfy

$$g>0,\quad q\geq 0,\quad \tau\geq 0,\quad r<0,\quad b_0>0,\quad \omega_n>0, \tag{3.1}$$

$$D>0,\quad C_\perp<0,\quad C_a>0,\quad c_1\geq K_2+K_4\geq c_0, \tag{3.2}$$

$$K_1\geq K_2+K_4,\quad K_3\geq K_2+K_4,\quad 0\geq K_4, \tag{3.3}$$

where c_0 and c_1 are positive constants. The latter inequalities are necessary conditions to ensure coercivity of the energy [1].

3.1. Boundary conditions. Let $\Omega\subset\mathbf{R}^3$ be a bounded, cylinder-like domain, with boundary $\partial\Omega=\Sigma\cup\mathcal{S}_1\cup\mathcal{S}_2$. We assume that the lateral surface Σ is of class C^2, and that $\mathcal{S}_1$ and $\mathcal{S}_2$ are plane cross sections with unit normal ν_1 and ν_2, respectively. Letting $\psi=\rho e^{i\omega}$, we rewrite F_{Sm} as

$$\begin{aligned}F_{Sm}&=D|\mathbb{D}^2\psi|^2+C_\perp|\mathbb{D}\psi|^2+C_a|\mathbf{n}\cdot\mathbb{D}\psi|^2+r|\psi|^2+\frac{g}{2}|\psi|^4\\&=D[(\Delta\rho-\rho|\nabla\omega-q\mathbf{n}|^2)^2+(\rho\nabla\cdot(\nabla\omega-q\mathbf{n})+2\nabla\rho\cdot(\nabla\omega-q\mathbf{n}))^2]\\&\quad+C_a[(\nabla\rho\cdot\mathbf{n})^2+\rho^2(\nabla\omega\cdot\mathbf{n}-q)^2]+r\rho^2+\frac{g}{2}\rho^4\\&\quad+C_\perp\left(|\nabla\rho|^2+\rho^2|\nabla\omega-q\mathbf{n}|^2\right).\end{aligned} \tag{3.4}$$

The following lemma based on Gauss' theorem motivates the boundary conditions taken into account.

LEMMA 3.1. *Let $\Omega\subset\mathbf{R}^3$ be as previously defined. Let f be a smooth scalar function defined in Ω. Then f satisfies the following identity:*

$$\begin{aligned}\int_\Omega|\triangle f|^2\,d\mathbf{x}&=\int_{\partial\Omega}\left[\nabla\cdot(\nabla f)(\nu\cdot\nabla f)-\frac{1}{2}\nabla(|\nabla f|^2)\cdot\nu\right]dS\\&\quad+\sum_{i,j=1,2,3}\int_\Omega(\partial_i\partial_j f)^2\,d\mathbf{x}.\end{aligned} \tag{3.5}$$

Let $k>q$ be a given constant. We assume that the boundary $\partial\Omega$ satisfies either

$$\begin{aligned}&\nabla\omega\cdot\nu=0\ \text{ on }\ \Sigma,\quad \nabla\omega\cdot\nu_i=k\ \text{ on }\ \mathcal{S}_i,\ i=1,2,\\&|\nabla\omega|^2=k^2\ \text{ on }\ \Sigma\cup\mathcal{S}_1\cup\mathcal{S}_2,\end{aligned} \tag{3.6}$$

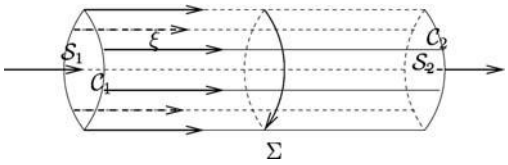


FIG. 3.1. *A vortex tube: $\mathcal{C}_1$ and $\mathcal{C}_2$ are the boundary curves of $\mathcal{S}_1$ and $\mathcal{S}_2$.*

or

$$\begin{aligned} &\nabla\omega\cdot\nu = \pm k \ \text{ on } \Sigma, \quad \nabla\omega\cdot\nu_i = 0 \ \text{ on } \ \mathcal{S}_i,\ i=1,2,\\ &|\nabla\omega|^2 = k^2 \ \text{ on } \ \Sigma\cup\mathcal{S}_1\cup\mathcal{S}_2. \end{aligned} \tag{3.7}$$

Such relations correspond to smectic layers reaching the boundary in a perpendicular and tangential fashion, respectively, with a prescribed wave number k. In case of (3.7), the surface integration in (3.5) becomes $\pm 2k^2\int_\Sigma H\,dS$, where H is the mean curvature. In contrast, with boundary conditions (3.6), the surface integration in (3.5) is zero.

Remark. The choice of domain and boundary conditions of the problem is motivated by vorticity geometry. Indeed, Ω and Σ play the roles of vortex tube and vortex sheet, respectively. Moreover, $\nabla\omega$ is analogous to the fluid vorticity ξ, as in Figure 3.1.

3.2. Existence of minimizers. For simplicity, we restrict ourselves to the case that ρ is constant (say $\rho = 1$), that is, no nematic defects are present in the sample, and rewrite the smectic energy as follows:

$$\begin{aligned} F_{Sm} = {} & D(\Delta\omega - q\nabla\cdot\mathbf{n})^2 + D\left(|\nabla\omega - q\mathbf{n}|^2 + \frac{C_\perp}{2D}\right)^2\\ & + C_a(\nabla\omega\cdot\mathbf{n} - q)^2 + \left(r + \frac{g}{2} - \frac{C_\perp^2}{4D}\right). \end{aligned} \tag{3.8}$$

We also assume that $\mathbf{D}$ and $\mathbf{E}$ satisfy Maxwell's equations (2.2). We use an electric potential φ satisfying $\mathbf{E} = \nabla\varphi$ and impose boundary conditions for φ so that (2.2) reads

$$\begin{cases} -\nabla\cdot((\varepsilon_\perp\mathbf{I} + \varepsilon_a\mathbf{n}\otimes\mathbf{n})\nabla\varphi) = \nabla\cdot\mathbf{P} \text{ in } \Omega,\\ \varphi = \varphi_0 \text{ on } \Sigma,\\ -\big((\varepsilon_\perp\mathbf{I} + \varepsilon_a\mathbf{n}\otimes\mathbf{n})\nabla\varphi\big)\cdot\nu = \mathbf{P}\cdot\nu \text{ on } \mathcal{S}_1\cup\mathcal{S}_2, \end{cases} \tag{3.9}$$

where $\varphi_0 \in H^{\frac{1}{2}}(\Sigma)$ is prescribed.

Define

$$\begin{aligned} &\mathcal{X} = \Big\{(\mathbf{n},\mathbf{P}) \in W^{1,2}(\Omega,\mathbf{S}^2)\times W^{1,2}(\Omega,\mathbf{R}^3) : ||P||_\infty \le P_0\Big\},\\ &\mathcal{H} = \{\omega \in W^{2,2}(\Omega) \mid \omega \text{ satisfies (3.6) or (3.7) on } \partial\Omega\},\\ &H^1_{\varphi_0} = \{\varphi \in H^1(\Omega) : \varphi = \varphi_0 \text{ on } \Sigma\},\\ &\mathcal{A}^* = \mathcal{H}\times\mathcal{X}, \quad \text{and}\\ &\mathcal{A} = \mathcal{A}^*\times H^1_{\varphi_0}, \end{aligned}$$

where P_0 is the given polarization saturation constant depending on the material and temperature. For constant potential φ_0, the boundary condition $\varphi = \varphi_0$ on Σ can be

considered as $\varphi = \varphi_0$ in $\mathbf{R}^3 - \Omega$. In this case, the nonlocal energy $\int_{\mathbf{R}^3-\Omega} |\mathbf{E}|^2 \, d\mathbf{x}$ is zero. For simplicity, we will drop the nonlocal energy in $\mathcal{E}$. We then rewrite the total energy functional $\mathcal{E}$ as a sum:

$$\mathcal{E} = \mathcal{W} - \frac{1}{2}\int_\Omega \{\varepsilon_\perp |\nabla\varphi|^2 + \varepsilon_a(\mathbf{n}\cdot\nabla\varphi)^2 + 2\mathbf{P}\cdot\nabla\varphi\}\, d\mathbf{x}, \tag{3.10}$$

where

$$\mathcal{W} = \int_\Omega \{F_N + F_{Sm} + F_P\}\, d\mathbf{x} + \int_{\partial\Omega} \omega_n((1-\alpha_0(\mathbf{n}\cdot\nu)^2)\, dS. \tag{3.11}$$

Since we are interested in the case that K_2 and K_3 are large, we assume that $\min\{K_2, K_3\} \geq 8c_1$ [1]. We note that for all $\mathbf{n} \in \mathbf{W}^{1,2}(\Omega, \mathbf{S}^2)$ the following identities hold:

$$\begin{aligned} &tr(\nabla\mathbf{n})^2 = |\nabla\mathbf{n}|^2 - |\nabla\times\mathbf{n}|^2, \\ &|\nabla\times\mathbf{n}|^2 = |\mathbf{n}\cdot\nabla\times\mathbf{n}|^2 + |\mathbf{n}\times\nabla\times\mathbf{n}|^2. \end{aligned}$$

Using these identities, we get

$$\begin{aligned} F_N = {} & (K_1 - K_2 - K_4)(\nabla\cdot\mathbf{n})^2 + (K_2+K_4)|\nabla\mathbf{n}|^2 - (K_2+K_4)|\nabla\times\mathbf{n}|^2 \\ & + K_2(\mathbf{n}\cdot\nabla\times\mathbf{n} + \tau)^2 + K_3|\mathbf{n}\times\nabla\times\mathbf{n} + \mathbf{P}|^2. \end{aligned} \tag{3.12}$$

Now, the following inequalities hold:

$$\begin{aligned} & K_2(\mathbf{n}\cdot\nabla\times\mathbf{n}+\tau)^2 + K_3|\mathbf{n}\times\nabla\times\mathbf{n} - \mathbf{P}|^2 - (K_2+K_4)|\nabla\times\mathbf{n}|^2 \\ & \geq 4c_1\left(\frac{1}{2}|\mathbf{n}\times\nabla\times\mathbf{n}|^2 - 2|\mathbf{P}|^2\right) + 2c_1(\mathbf{n}\cdot\nabla\times\mathbf{n}+\tau)^2 - (K_2+K_4)|\nabla\times\mathbf{n}|^2 \\ & \geq 2c_1|\nabla\times\mathbf{n} + \tau\mathbf{n}|^2 - 8c_1|\mathbf{P}|^2 - (K_2+K_4)|\nabla\times\mathbf{n}|^2 \\ & \geq 2c_1\left(\frac{1}{2}|\nabla\times\mathbf{n}|^2 - 2\tau^2\right) - 8c_1|\mathbf{P}|^2 - (K_2+K_4)|\nabla\times\mathbf{n}|^2 \\ & \geq (c_1 - K_2 - K_4)|\nabla\times\mathbf{n}|^2 - 4c_1(\tau^2 + 2|\mathbf{P}|^2) \\ & \geq -4c_1(\tau^2 + 2|\mathbf{P}|^2). \end{aligned} \tag{3.13}$$

It follows from (3.12), (3.13), and Lemma 3.1 that $\mathcal{W}$ is bounded below in $\mathcal{A}^*$. Therefore, we have

$$M_1 \leq \inf_{(\mathbf{n},\mathbf{P},\omega)\in\mathcal{A}^*} \mathcal{W}(\mathbf{n},\mathbf{P},\omega) < \infty$$

for some $M_1 \in \mathbf{R}$.

Now, we rewrite the Oseen–Frank energy in (3.12) as follows:

$$\begin{aligned} F_N = {} & (K_1 - K_2 - K_4)(\nabla\cdot\mathbf{n})^2 + (K_2+K_4)|\nabla\mathbf{n}|^2 - K_4\left(\mathbf{n}\cdot\nabla\times\mathbf{n} - \frac{\tau K_2}{K_4}\right)^2 \\ & + K\left|\mathbf{n}\times\nabla\times\mathbf{n} + \frac{K_3}{K}\mathbf{P}\right|^2 + \left(K_3 - \frac{K_3^2}{K}\right)|\mathbf{P}|^2 + \tau^2\left(K_2 - \frac{K_2^2}{|K_4|}\right), \end{aligned}$$

where $K = K_3 - K_2 - K_4$.

Let $\{(\mathbf{n}^j, \mathbf{P}^j, \omega^j)\}$ be a minimizing sequence for $\mathcal{W}$. Since $|\mathbf{n}^j| = 1$, we get

$$\begin{aligned}&\mathbf{n}^j \rightharpoonup \mathbf{n}^\infty \text{ in } W^{1,2}(\Omega),\\ &\mathbf{n}^j \to \mathbf{n}^\infty \text{ almost everywhere in } \Omega, \quad \text{and}\\ &\mathbf{P}^j \rightharpoonup \mathbf{P}^\infty \text{ in } W^{1,2}(\Omega)\end{aligned}$$

as $j \to \infty$. Furthermore, we have

$$\begin{aligned}&\mathbf{n}^j \times \nabla \times \mathbf{n}^j \rightharpoonup \mathbf{n}^\infty \times \nabla \times \mathbf{n}^\infty \text{ in } L^2(\Omega),\\ &\mathbf{n}^j \cdot \nabla \times \mathbf{n}^j \rightharpoonup \mathbf{n}^\infty \cdot \nabla \times \mathbf{n}^\infty \text{ in } L^2(\Omega)\end{aligned}$$

as $j \to \infty$. Note that for all j,

$$\int_\Omega |\nabla \omega^j|^2 \, d\mathbf{x} \le 2 \int_\Omega (|\nabla \omega^j - q\mathbf{n}^j|^2 + q^2) \, d\mathbf{x}, \quad \text{and} \tag{3.14}$$

$$\begin{aligned}\int_\Omega |\Delta \omega^j|^2 \, d\mathbf{x} &= \int_\Omega |\Delta \omega^j - q\nabla \cdot \mathbf{n}^j + q\nabla \cdot \mathbf{n}^j|^2 \, d\mathbf{x}\\ &\le 2 \int_\Omega \left[|\Delta \omega^j - q\nabla \cdot \mathbf{n}^j|^2 + q^2(\nabla \cdot \mathbf{n}^j)^2\right] \, d\mathbf{x}\end{aligned} \tag{3.15}$$

hold. From (3.14), (3.15), and Lemma 3.1, we get

$$||\omega^j||_{W^{2,2}(\Omega)} \le R$$

for some $R > 0$. Hence, we obtain that

$$\begin{aligned}&\omega^j \rightharpoonup \omega^\infty \text{ in } W^{2,2}(\Omega), \quad \text{and}\\ &\nabla \omega^j \times \mathbf{n}^j \to \nabla \omega^\infty \times \mathbf{n}^\infty \text{ in } L^2(\Omega)\end{aligned}$$

as $j \to \infty$. Using $||a| - |b|| \le |a - b|$ for a and b in $\mathbf{R}$, we show that

$$\left|(|\nabla \omega^j \times \mathbf{n}^j|)\mathbf{P}^j - |\mathbf{P}^j|(\nabla \omega^j \times \mathbf{n}^j)\right|^2 \to \left|(|\nabla \omega^\infty \times \mathbf{n}^\infty|)\mathbf{P}^\infty - |\mathbf{P}^\infty|(\nabla \omega^\infty \times \mathbf{n}^\infty)\right|^2$$

in L^1 as $j \to \infty$.

Since $\mathbf{n} \cdot \nu \in H^{\frac{1}{2}}(\partial\Omega) \subset L^2(\partial\Omega)$ with strong topology, $\int_{\partial\Omega} \omega_n(1 - \alpha_0(\mathbf{n} \cdot \nu)^2) \, dS$ is lower semicontinuous. Therefore, $\mathcal{W}$ is coercive and weakly lower semicontinuous; that is,

$$\mathcal{W}(\mathbf{n}^\infty, \mathbf{P}^\infty, \omega^\infty) \le \lim_{j \to \infty} \mathcal{W}(\mathbf{n}^j, \mathbf{P}^j, \omega^j).$$

Therefore we have the following lemma.

LEMMA 3.2. *Assuming that* $\min\{K_2, K_3\} \ge 8c_1$*, there exists a minimizing triple* $(\mathbf{n}, \mathbf{P}, \omega)$ *of the energy functional* $\mathcal{W}$ *in* $\mathcal{A}^*$.

Now, we prove existence of minimizers for $\mathcal{E}$ in $\mathcal{A}$. For any $(\mathbf{n}, \mathbf{P}, \omega) \in \mathcal{A}^*$, by the fundamental theory of elliptic PDEs (3.9) has a unique solution, which we denote by $\Phi_{\varphi_0}(\mathbf{n}, \mathbf{P})$, and thus $\Phi_{\varphi_0}(\mathbf{n}, \mathbf{P})$ is the unique minimizer of the functional $-\int_\Omega F_E \, d\mathbf{x}$ in $H^1_{\varphi_0}$. Substituting $\Phi_{\varphi_0}(\mathbf{n}, \mathbf{P})$ for φ in $\mathcal{E}$, we define $\mathcal{E}^*$ by

$$\mathcal{E}^*(\mathbf{n}, \mathbf{P}, \omega) = \mathcal{E}(\mathbf{n}, \mathbf{P}, \omega, \Phi_{\varphi_0}(\mathbf{n}, \mathbf{P})).$$

Let

$$D(\nabla\varphi, \mathbf{n}) = -(\varepsilon_\perp \mathbf{I} + \varepsilon_a \mathbf{n}\otimes\mathbf{n})\nabla\varphi, \tag{3.16}$$

$$A(\nabla\varphi, \mathbf{n}) = [(\varepsilon_\perp \mathbf{I} + \varepsilon_a \mathbf{n}\otimes\mathbf{n})\nabla\varphi]\cdot\nabla\varphi. \tag{3.17}$$

After modifying Theorem 4.1 in [15], we have the following theorem.

THEOREM 3.3. *A quadruple* $(\mathbf{n}, \mathbf{P}, \omega, \varphi)$ *is a critical point of* $\mathcal{E}$ *in* $\mathcal{A}$ *subject to* (3.9) *if and only if*

$$\varphi = \Phi_{\varphi_0}(\mathbf{n}, \mathbf{P}) \quad \text{and} \quad \delta\mathcal{E}^*(\mathbf{n}, \mathbf{P}, \omega) = 0 \text{ in } \mathcal{A}^*. \tag{3.18}$$

THEOREM 3.4. *For* φ_0 *as above and* $\min\{K_2, K_3\} \geq 8c_1$, *there exists a triple* $(\mathbf{n}, \mathbf{P}, \omega)$ *which minimizes* $\mathcal{E}^*$ *in* $\mathcal{A}^*$, *and therefore* $\mathcal{E}$ *achieves its minimum in* $\mathcal{A}$.

Proof. Let $(\tilde{\mathbf{n}}, \tilde{\mathbf{P}}, \tilde{\omega})$ be a minimizer of $\mathcal{W}$ in $\mathcal{A}^*$. Then

$$\inf_{(\mathbf{n},\mathbf{P},\omega)\in\mathcal{A}^*} \mathcal{E}^*(\mathbf{n}, \mathbf{P}, \omega) \leq \mathcal{E}^*(\tilde{\mathbf{n}}, \tilde{\mathbf{P}}, \tilde{\omega}) < \infty.$$

If $\tilde{\varphi}$ is some fixed $W^{1,2}$ extension of φ_0 to Ω, then for any $\eta > 0$,

$$\begin{aligned} -2\int_\Omega F_E\, d\mathbf{x} &= \int_\Omega [A(\nabla\Phi_{\varphi_0}(\mathbf{n}, \mathbf{P}), \mathbf{n}) + \mathbf{P}\cdot\nabla\Phi_{\varphi_0}(\mathbf{n}, \mathbf{P})]\, d\mathbf{x} \\ &\leq \int_\Omega [A(\nabla\tilde{\varphi}, \mathbf{n}) + \mathbf{P}\cdot\nabla\tilde{\varphi}]\, d\mathbf{x} \\ &\leq C(\Omega, \varphi_0), \end{aligned}$$

for some $C(\Omega, \varphi_0)$ depending on Ω and φ_0. Since $\mathcal{W}$ is bounded from below, it follows from the above that $\mathcal{E}^*$ is also bounded below.

Now, choose a minimizing sequence $(\mathbf{n}^i, \mathbf{P}^i, \omega^i)$ in $\mathcal{A}^*$ and set $\varphi^i = \Phi_{\varphi_0}(\mathbf{n}^i, \mathbf{P}^i)$. Using the same computation as in the proof of the previous lemma, we obtain that

$$\begin{aligned} &\mathbf{n}^j \rightharpoonup \mathbf{n}^\infty \text{ in } W^{1,2}, \\ &\mathbf{n}^j \to \mathbf{n}^\infty \text{ almost everywhere in } \Omega, \\ &\mathbf{P}^j \rightharpoonup \mathbf{P}^\infty \text{ in } W^{1,2}, \quad \text{and} \\ &\omega^j \rightharpoonup \omega^\infty \text{ in } W^{2,2} \end{aligned}$$

as $j \to \infty$. Since $(\mathbf{P}^j)$ converges strongly to $\mathbf{P}^\infty$ in L^2,

$$\int_\Omega \mathbf{P}^j\cdot\nabla\xi\, d\mathbf{x} \to \int_\Omega \mathbf{P}^\infty\cdot\nabla\xi\, d\mathbf{x} \quad \text{as } j\to\infty \tag{3.19}$$

for any $\xi \in H^1(\Omega)$. It follows from (3.9) and (3.19) that (φ^j) converges strongly to φ^∞ in $H^1_{\varphi_0}$.

Since φ^i is the minimizer of $-\int_\Omega F_E(\mathbf{n}^i, \mathbf{P}^i, \nabla\varphi)\, d\mathbf{x}$ for each fixed $\mathbf{n}^i$ and $\mathbf{P}^i$, we have

$$-\int_\Omega F_E(\mathbf{n}^i, \mathbf{P}^i, \nabla\varphi^i)\, d\mathbf{x} \leq -\int_\Omega F_E(\mathbf{n}^i, \mathbf{P}^i, \nabla\varphi^\infty)\, d\mathbf{x}.$$

By Lebesgue's theorem, we obtain

$$\lim_{i\to\infty}\left\{-\int_\Omega F_E(\mathbf{n}^i, \mathbf{P}^i, \nabla\varphi^i)\, d\mathbf{x}\right\} = -\int_\Omega F_E(\mathbf{n}^\infty, \mathbf{P}^\infty, \nabla\varphi^\infty)\, d\mathbf{x},$$

so that

$$\limsup_{i\to\infty}\left\{-\int_\Omega F_E(\mathbf{n}^i,\mathbf{P}^i,\nabla\varphi^i)\,d\mathbf{x}\right\}\le-\int_\Omega F_E(\mathbf{n}^\infty,\mathbf{P}^\infty,\nabla\varphi^\infty)\,d\mathbf{x}.$$

This implies that

$$\liminf_{i\to\infty}\int_\Omega F_E(\mathbf{n}^i,\mathbf{P}^i,\nabla\varphi^i)\,d\mathbf{x}\ge\int_\Omega F_E(\mathbf{n}^\infty,\mathbf{P}^\infty,\nabla\varphi^\infty)\,d\mathbf{x}.$$

Since $\mathcal{W}$ is lower semicontinuous, we finally conclude that

$$\mathcal{E}^*(\mathbf{n}^\infty,\mathbf{P}^\infty,\omega^\infty)=\inf_{(\mathbf{n},\mathbf{P},\omega)\in\mathcal{A}^*}\mathcal{E}^*(\mathbf{n},\mathbf{P},\omega).\qquad\square$$

3.3. Applications. Problems analogous to the model problem of the previous subsection are often found in applications. Let us discuss two examples. First we consider the case of a liquid crystal contained in a dielectric liquid medium with free ions that form a charged layer of density σ on the interface. Letting Ω be a bounded domain occupied by liquid crystals, we are interested in a minimization problem of the energy functional (3.10) with an electric potential $\varphi\in W^{1,2}(\mathbf{R}^3)$, satisfying

$$\text{(3.20)}\qquad\begin{cases}-\nabla\cdot((\varepsilon_\perp\mathbf{I}+\varepsilon_a\mathbf{n}\otimes\mathbf{n})\nabla\varphi)=\nabla\cdot\mathbf{P}\text{ in }\Omega,\\ -\Delta\varphi=0\text{ in }\mathbf{R}^3-\bar\Omega,\\ -[(\varepsilon_\perp\mathbf{I}+\varepsilon_a\mathbf{n}\otimes\mathbf{n})\nabla\varphi-\varepsilon_0\nabla\varphi]\cdot\nu=\mathbf{P}\cdot\nu+\sigma\text{ on }\partial\Omega,\end{cases}$$

where ε_0 is the dielectric coefficient of the medium.

If φ is a solution of (3.20), then $\varphi+C$ is also a solution of (3.20) for any constant C. Define

$$\mathcal{V}=\left\{v\in W^{1,2}(\mathbf{R}^3);\int_{\mathbf{R}^3}v\,d\mathbf{x}=0\right\}.$$

By the standard theory of elliptic PDEs and calculus of variations, for given $\mathbf{n},\mathbf{P}$, the solution of (3.20) is corresponding to the minimizer of the energy functional $\mathcal{F}$ on $\mathcal{V}$

$$\begin{aligned}\mathcal{F}=&\frac12\int_\Omega[(\varepsilon_\perp\mathbf{I}+\varepsilon_a\mathbf{n}\otimes\mathbf{n})\nabla\varphi]\cdot\nabla\varphi\,d\mathbf{x}+\frac12\int_{\mathbf{R}^3-\bar\Omega}\varepsilon_0\nabla\varphi\cdot\nabla\varphi\,d\mathbf{x}\\&+\int_{\mathbf{R}^3}\mathbf{P}\chi_\Omega\cdot\nabla\varphi\,d\mathbf{x}+\int_{\partial\Omega}\sigma\varphi\,dS.\end{aligned}$$

Since $\mathcal{F}$ is $\mathcal{V}$-elliptic, there exists a unique minimizer φ, denoted by $\Phi(\mathbf{n},\mathbf{P})$, of $\mathcal{F}$ on $\mathcal{V}$. We substitute $\nabla\Phi(\mathbf{n},\mathbf{P})$ for $\mathbf{E}$ in the electrostatic energy $\int_{\mathbf{R}^3}F_E\,d\mathbf{x}$. Using similar arguments as in Theorems 3.3, and 3.4, we obtain the following corollary.

COROLLARY 3.5. *Assume that* $\min\{K_2,K_3\}\ge 8c_1$. *Then there exists a minimizing quadruple* $(\mathbf{n},\mathbf{P},\omega,\varphi)$ *of* $\mathcal{E}$ *on* $\mathcal{A}^*\times\mathcal{V}$ *satisfying* (3.20).

The second case comes up in device applications. Now the liquid crystal domain Ω is confined between two conducting plates Ω_1 and Ω_2. We assume that there is no free charge in Ω, and we neglect end-effects. In this case, the boundary value problem is stated as follows: For a given potential $\tilde\varphi$, find φ satisfying the following conditions:

$$\text{(3.21)}\qquad\begin{cases}\nabla\cdot(D(\nabla\varphi,\mathbf{n}))=\nabla\cdot\mathbf{P}\text{ in }\Omega,\\ -\Delta\varphi=0\text{ in }\mathbf{R}^3-\overline{\Omega\cup\Omega_1\cup\Omega_2},\\ \Big(D(\nabla\varphi,\mathbf{n})+\nabla\varphi\Big)\cdot\nu=\mathbf{P}\cdot\nu\text{ on }\partial\Omega\cap\partial(\mathbf{R}^3-\overline{\Omega\cup\Omega_1\cup\Omega_2}),\\ \varphi=\varphi_i\text{ in }\Omega_i\text{ for }i=1,2,\\ \varphi\to0\text{ as }|\mathbf{x}|\to\infty,\end{cases}$$

where $D(\nabla\varphi, \mathbf{n})$ is defined in (3.16) and φ_i, $i = 1, 2$, are constant potential functions. If φ_i is set to zero, $\nabla\varphi$ gives the electric field created by the polar distribution. Analogous boundary value problems for ferroelectric solids are considered in [35]. Since φ_i, $i = 1, 2$, are constant, we can consider $\varphi = \varphi_i$ as Dirichlet boundary conditions on Ω_i.

Let

$$\tilde{\Omega} = \mathbf{R}^3 - \overline{\Omega \cup \Omega_1 \cup \Omega_2}.$$

Then the solution of the problem (3.21) can be sought as the solution $\varphi \in W^{1,2}(\mathbf{R}^3 - \Omega_1 \cup \Omega_2)$ satisfying

$$\text{(3.22)} \qquad \begin{cases} \nabla \cdot (D(\nabla\varphi, \mathbf{n})) = \nabla \cdot \mathbf{P} \text{ in } \Omega, \\ -\Delta\varphi = 0 \text{ in } \tilde{\Omega}, \\ \varphi = \varphi_i \text{ on } \partial\Omega_i, \quad i = 1, 2, \\ \Big(D(\nabla\varphi, \mathbf{n}) + \nabla\varphi\Big) \cdot \nu = \mathbf{P} \cdot \nu \text{ on } \partial\Omega \cap \partial\tilde{\Omega}. \end{cases}$$

Define

$$\tilde{\mathcal{V}} = \{v \in W^{1,2}(\mathbf{R}^3 - \Omega_1 \cup \Omega_2); v = \varphi_i \text{ on } \partial\Omega_i, i = 1, 2\},$$

and

$$\tilde{\mathcal{F}} = \int_\Omega \{[(\varepsilon_\perp \mathbf{I} + \varepsilon_a \mathbf{n} \otimes \mathbf{n})\nabla\varphi] \cdot \nabla\varphi + \mathbf{P} \cdot \nabla\varphi\}\, d\mathbf{x} + \int_{\tilde{\Omega}} |\nabla\varphi|^2\, d\mathbf{x}.$$

We see that $\tilde{\mathcal{F}}$ is $\tilde{\mathcal{V}}$-elliptic. There exists a unique minimizer φ, denoted by $\Phi(\mathbf{n}, \mathbf{P})$, of $\tilde{\mathcal{F}}$ on $\tilde{\mathcal{V}}$. Repeating the same arguments as in previous problem, we conclude the corollary as below.

COROLLARY 3.6. *Assume that* $\min\{K_2, K_3\} \geq 8c_1$. *Then there exists a minimizing quadruple* $(\mathbf{n}, \mathbf{P}, \omega, \varphi)$ *of* $\mathcal{E}$ *on* $\mathcal{A}^* \times \tilde{\mathcal{V}}$ *satisfying* (3.22).

3.4. Shapes and polarization. We close this section with examples to illustrate the relationship between the nonlocal energy and the shape of the domain.

Example 1. *Uniformly polarized rectilinear cylinder.* Let Ω be a cylinder in $\mathbf{R}^3$, $x^2 + y^2 \leq r_1^2$, occupied by a smectic C material such that

$$\nabla\omega = \mathbf{e}_r, \quad \mathbf{n} = f(r)\mathbf{e}_r + g(r)\mathbf{e}_\theta, \quad \mathbf{P} = P_0\mathbf{k}.$$

Let $\tilde{\Omega}$ denote a second cylinder, $x^2 + y^2 \leq r_2^2$, $r_1 < r_2$. Suppose that $\mathbf{E}$ is the electric field on the cylindrical surface $\partial\tilde{\Omega}$. Applying Gauss' theorem to $\tilde{\Omega}$, we get

$$\int_{\partial\tilde{\Omega}} \mathbf{E} \cdot \nu\, dS = 0$$

since the net charge inside $\tilde{\Omega}$ is zero. By symmetry, we observe that $|\mathbf{E}|$ is constant on $\partial\tilde{\Omega}$, $\mathbf{E} \cdot \nu = |\mathbf{E}|$, and thus

$$\int_{\mathbf{R}^3 - \Omega} |\mathbf{E}|^2\, d\mathbf{x} = 0.$$

We note that the electric field $\mathbf{E}$ due to the polarization may not be zero outside Ω if the shape is nonsymmetric. For instance, in the case of a bent cylindrical domain,

FIG. 3.2. *A polarized helical filament.*

the electric field created by the polarization is not rotationally symmetric. Therefore, we cannot conclude that $|\mathbf{E}|$ is constant on $\partial\tilde{\Omega}$, and so $\mathbf{E}$ is, in general, nonzero on $\partial\tilde{\Omega}$. In this case, the self energy $\int_{\mathbf{R}^3-\Omega} |\mathbf{E}|^2 \, d\mathbf{x}$ is also nonzero.

Example 2. *Polarized helical filament.* In Example 1, we replace Ω by a thin filament [2],

$$\Omega = \{\mathbf{x} \in \mathbf{R}^3 : \mathbf{x} = \mathcal{C}(s) + \xi\mathbf{e}, s \in [0, l], \, \xi \in [0, r], \, \mathbf{e} \cdot \mathbf{T} = 0, \mathbf{e} \cdot \mathbf{e} = 1\},$$

where $\mathcal{C} : [0, l] \to \mathbf{R}^3$ is a smooth curve and $\mathbf{T}$ is the unit tangent vector. The domain Ω is a thin filament and not necessarily a right cylinder, as in Figure 3.2. Let $\mathbf{N}$ and $\mathbf{B}$ denote the normal and binormal vectors, respectively, to the curve.

The curve $\mathcal{C}$ represents the center curve of the curvilinear cylindrical domain Ω. We assume that for each $s \in [0, r]$ the smectic layer normal is parallel to $\mathbf{T}$, and the director field $\mathbf{n}$ is parallel to the plane determined by $\mathbf{T}$ and $\mathbf{B}$, making a constant tilt angle with $\mathbf{T}$. Accordingly, we set

$$\nabla\omega = \mathbf{T}, \quad \mathbf{n} = \alpha\mathbf{T} + \beta\mathbf{B}, \alpha^2 + \beta^2 = 1, \quad \mathbf{P} = -P_0\mathbf{N}.$$

Define a coordinate system (s, ξ, θ) so that

$$\mathbf{e}_\xi = \cos\theta\mathbf{N} + \sin\theta\mathbf{B}, \qquad \mathbf{e}_\theta = -\sin\theta\mathbf{N} + \cos\theta\mathbf{B}.$$

Since the net charge in $\tilde{\Omega}$ is zero, by Gauss' theorem

$$\int_{\partial\tilde{\Omega}} \mathbf{E} \cdot \nu \, dS = 0.$$

In general, though, we cannot conclude that $\mathbf{E}$ is symmetric around the curve $\mathcal{C}$, and so $\mathbf{E}$ may not be zero outside the filament region. Now the question is whether or not there is a shape such that $\mathbf{E} = \mathbf{0}$ in $\mathbf{R}^3 - \Omega$. Heuristically, we can view such a shape as the limiting case of a helical filament as the pitch approaches zero. Note that this would allow us to recover the symmetry property of the domain and of the electric field, and conclude that $\int_{\mathbf{R}^3-\Omega} |\mathbf{E}|^2 \, d\mathbf{x}$ is negligible.

4. Classification of energy minimizers. We apply asymptotic arguments to determine whether energy minimizers correspond to either helical configurations or ferroelectric ones. In this section, we consider the energy as in (2.14). We wish to identify the smectic layer geometry and find parameter conditions leading to helical director configurations in the bulk with zero average polarization, as well as those giving homogeneous ferroelectric states. For this, we will consider a rectangular domain between two parallel plates:

$$\Omega = \{\mathbf{x} = (x, y, z) : \, 0 < y, z < L, \; 0 < x < d\},$$

for fixed $0 < L$, $0 < d$. Let $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ denote the corresponding orthonormal system of vectors.

4.1. Helical energy minimizers. We determine the structure of the energy minimizers $(\mathbf{n}, \omega, \mathbf{P})$ when K_2 and K_3 as well as the smectic coefficients dominate over the Ginzburg–Landau energy and surface energy parameters, and $C_\perp < 0$. Such a situation arises at temperatures below the threshold of the smectic A to smectic C transition, yielding helical configurations of $\mathbf{n}$ and $\mathbf{P}$. It is well known that in the higher temperature transition from nematic to smectic A, K_2 becomes unbounded and the smectic coefficient $C_\perp \geq 0$.

We take the admissible set so that

$$k = q\sqrt{\frac{|C_\perp|}{2Dq^2} + 1}. \tag{4.1}$$

We consider admissible fields such that $\mathbf{n}$ makes a constant angle α with the layer normal vector $\nabla\omega$. We also choose α such that

$$\tan\alpha = \sqrt{\frac{|C_\perp|}{2Dq^2}}. \tag{4.2}$$

Specifically, we let

$$\begin{cases} \mathbf{n}_0 = (a\cos\frac{\tau z}{a^2}, a\sin\frac{\tau z}{a^2}, c), \\ a = \sin\alpha \neq 0, \qquad c = \cos\alpha \neq 0, \qquad \frac{a^2}{c^2} = \frac{|C_\perp|}{2Dq^2}, \\ \mathbf{P}_0 = \frac{c\tau}{a}(-\sin\frac{\tau z}{a^2}\mathbf{i} + \cos\frac{\tau z}{a^2}\mathbf{j}), \\ \omega_0 = kz, \qquad k = \frac{q}{c}, \qquad \nu = \mathbf{i}. \end{cases} \tag{4.3}$$

A simple calculation gives

$$\begin{aligned} &\nabla\omega_0 \cdot \mathbf{n}_0 = q, \qquad \triangle\omega_0 = 0, \qquad |\nabla\omega_0 - q\mathbf{n}_0|^2 = \frac{|C_\perp|}{2D}, \\ &\nabla\cdot\mathbf{n}_0 = 0, \qquad \mathbf{n}_0\cdot\nabla\times\mathbf{n}_0 + \tau = 0, \qquad |\mathbf{n}_0\times(\nabla\times\mathbf{n}_0) + \mathbf{P}_0| = 0, \\ &\nabla\cdot\mathbf{P}_0 = 0, \qquad |\nabla\mathbf{P}_0| = \frac{c\tau^2}{a^3}, \qquad \mathbf{P}_0 = \nabla\omega_0\times\mathbf{n}_0. \end{aligned}$$

We observe that the quantity $\tan\alpha = \sqrt{\frac{|C_\perp|}{2Dq^2}}$ is of the order of $\tan\frac{\pi}{6}$ according to experimental measurements of the director tilt angle. This, together with available information on the wave number q in the smectic A phase, allows us to determine the relative value of the smectic parameters $|C_\perp|$ and D.

The total energy corresponding to the fields in (4.3) is given by

$$\begin{aligned} \mathcal{E}_0 = L^2 d\left[r + \frac{g}{2} - \frac{C_\perp^2}{4D} + (K_2+K_4)\frac{\tau^2}{a^2} + \frac{c^2\tau^2}{a^2}\left(B\frac{\tau^2}{a^4} + a_0 + b_0\frac{c^2}{a^2}\tau^2\right)\right] \\ + (2L^2 + 4dL)(C_p\omega_p + C_r\omega_r + C_n\omega_n), \end{aligned}$$

where C_r, C_p, C_n are expressions involving a, c, q, and τ.

Letting $K_0 = \frac{1}{2}\min\{K_2, K_3\}$, we write the Oseen–Frank energy as follows:

$$\begin{aligned} F_N = &(K_2 - K_0)(\mathbf{n}\cdot\nabla\times\mathbf{n} + \tau)^2 + (K_3 - K_0)|n\times\nabla\times\mathbf{n} + \mathbf{P}|^2 \\ &+ (K_1 - K_2 - K_4)(\nabla\cdot\mathbf{n})^2 + (K_2+K_4)|\nabla\mathbf{n}|^2 \\ &+ K_0|\mathbf{n}\cdot\nabla\times\mathbf{n} + \tau|^2 + K_0|\mathbf{n}\times\nabla\times\mathbf{n} + \mathbf{P}|^2 - (K_2+K_4)|\nabla\times\mathbf{n}|^2. \end{aligned} \tag{4.4}$$

We invoke the following estimate of the last three terms:

$$K_0(|\mathbf{n}\cdot\nabla\times\mathbf{n}+\tau|^2+|\mathbf{n}\times\nabla\times\mathbf{n}+\mathbf{P}|^2)-(K_2+K_4)|\nabla\times\mathbf{n}|^2 \geq -4c_1(\tau^2+2|\mathbf{P}|^2).$$

We then obtain the following estimates:

$$\int_{\partial\Omega} G_S\,dS \geq -\left[2|\omega_p|+2|\omega_r|+|\omega_n|(1+|\alpha|)\right](2L^2+4dL),$$
$$\int_{\Omega}(F_{Sm}+F_N+G_P)\,d\mathbf{x} \geq \left[r+\frac{g}{2}-\left(\frac{C_\perp^2}{4D}+\frac{(a_0-8c_1)^2}{4b_0}+4c_1\tau^2\right)\right]dL^2.$$

Letting $(\mathbf{n},\mathbf{P},\omega)$ denote an energy minimizer, we get

$$\begin{aligned}0 &\leq \mathcal{E}(\mathbf{n},\mathbf{P},\omega)+[2|\omega_p|+2|\omega_r|+|\omega_n|(1+|\alpha|)](2L^2+4dL)\\ &\quad+\left[\frac{C_\perp^2}{4D}+\frac{(a_0-8c_1)^2}{4b_0}+4c_1\tau^2-r-\frac{g}{2}\right]dL^2\\ &\leq \mathcal{E}_0+[2|\omega_p|+2|\omega_r|+|\omega_n|(1+|\alpha|)](2L^2+4dL)\\ &\quad+\left[\frac{C_\perp^2}{4D}+\frac{(a_0-4c_1)^2}{4b_0}+4c_1\tau^2-r-\frac{g}{2}\right]dL^2\equiv\bar{\mathcal{E}}_0.\end{aligned}\tag{4.5}$$

Since $\frac{|C_\perp|}{2Dq^2}$ is bounded, we note that the quantity on the right-hand side of the inequality is independent of D, $C_\perp$, K_1, K_2, and K_3, with the only K_i constants appearing as the sum K_2+K_4. From (4.5) together with (3.12) and (3.13) we get the following theorem.

THEOREM 4.1. *Let $q>0$, $\tau>0$ be fixed. Suppose that the constitutive parameters satisfy assumptions* (3.1)–(3.3). *Suppose that $K_2,K_3\geq 8c_1$ and $0<\frac{|C_\perp|}{2Dq^2}\leq 1$. If $(\mathbf{n},\mathbf{P},\omega)$ is a minimizer of $\mathcal{E}$, then the following estimates hold:*

$$||(\nabla\omega\times\mathbf{n})|\mathbf{P}|-\mathbf{P}(|\nabla\omega\times\mathbf{n}|)||_{2,\Omega}^2\leq\epsilon^2\bar{\mathcal{E}}_0,\tag{4.6}$$

$$||\nabla\mathbf{n}||_{2,\Omega}\leq\frac{\bar{\mathcal{E}}_0}{K_2+K_4},\tag{4.7}$$

$$||\mathbf{n}\times\nabla\times\mathbf{n}+\mathbf{P}||_{2,\Omega}^2\leq\frac{\bar{\mathcal{E}}_0}{\min\{K_2,K_3\}},\tag{4.8}$$

$$||\mathbf{n}\cdot\nabla\times\mathbf{n}+\tau||_{2,\Omega}^2\leq\frac{\bar{\mathcal{E}}_0}{\min\{K_2,K_3\}},\tag{4.9}$$

$$\left\|\left|\frac{1}{q}\nabla\omega-\mathbf{n}\right|^2-\frac{|C_\perp|}{2Dq^2}\right\|_{2,\Omega}^2\leq\frac{\bar{\mathcal{E}}_0}{Dq^2},\tag{4.10}$$

$$\left\|\frac{1}{q}\nabla\omega\cdot\mathbf{n}-1\right\|_{2,\Omega}^2\leq\frac{\bar{\mathcal{E}}_0}{C_a},\tag{4.11}$$

$$||\nabla\mathbf{P}||_{2,\Omega}\leq\frac{\bar{\mathcal{E}}_0}{B}.\tag{4.12}$$

Next, we proceed to take limits in (4.6)–(4.11). We use the following representation for $\mathbf{n}$:

$$\mathbf{n}=\sin\theta\cos\phi\mathbf{i}+\sin\theta\sin\phi\mathbf{j}+\cos\theta\mathbf{k},$$

where $\phi = \phi(x, y, z)$ and $\theta = \theta(x, y, z)$ are functions resulting from energy minimization.

THEOREM 4.2. *Suppose that the hypotheses of the previous theorem hold. Then the energy minimizing fields* $(\mathbf{n}, \mathbf{P}, \omega)$ *satisfy the following limiting relations:*

$$\lim_{C_a \to \infty} \nabla\omega \cdot \mathbf{n} = q, \tag{4.13}$$

$$\lim_{|C_\perp| \to \infty} |\nabla\omega| = q\sqrt{\frac{|C_\perp|}{2Dq^2} + 1}, \tag{4.14}$$

$$\lim_{\epsilon \to 0} \mathbf{P} = \cot\alpha\tau \frac{\mathbf{k} \times \mathbf{n}}{|\mathbf{k} \times \mathbf{n}|}, \quad \cot\alpha = \sqrt{\frac{2Dq^2}{|C_\perp|}}, \tag{4.15}$$

$$\lim_{K \to \infty} (\mathbf{n} \times \nabla \times \mathbf{n} + \mathbf{P}) = 0, \tag{4.16}$$

$$\lim_{K \to \infty} (\mathbf{n} \cdot \nabla \times \mathbf{n} + \tau) = 0, \tag{4.17}$$

where $K = \min\{K_2, K_3\}$. *Furthermore, we get*

$$\lim_{|C_\perp| \to \infty} \omega = \left(q\sqrt{\frac{|C_\perp|}{2Dq^2} + 1} \right) z, \tag{4.18}$$

$$\lim_{K \to \infty} \mathbf{n} = \sin\alpha \cos\frac{\tau}{a^2} z\mathbf{i} + \sin\alpha \sin\frac{\tau}{a^2} z\mathbf{j} + \cos\alpha\mathbf{k}. \tag{4.19}$$

Proof. From the geometry of the domain and the boundary conditions, it follows that $\nabla\omega = |\nabla\omega|\mathbf{k}$, which together with (4.10) and (4.11) yields (4.13), (4.14), and (4.18). It now follows from (4.13) and (4.14) that $\theta = \alpha$ is the constant given by (4.2). These together with (4.6) yield $\mathbf{P} = |\mathbf{P}|\frac{\mathbf{k}\times\mathbf{n}}{|\mathbf{k}\times\mathbf{n}|}$. Combining this equation with (4.16) and (4.17) gives $\phi = \frac{\tau}{a^2} z$ in (4.19). This also yields (4.15). ∎

Note that from the property $\lim_{D\to\infty}(\triangle\omega - q\nabla\cdot\mathbf{n}) = 0$ it follows that the limiting director field has zero divergence, in agreement with (4.19).

4.2. Ferroelectric energy minimizers. In the previous theorem, the elasticity constants K_2 and K_3 become unbounded with respect to the parameters of the polarization contribution to the energy. We will show that the ferroelectric configurations,

$$\mathbf{n} = \pm\sin\alpha\mathbf{j} + \cos\alpha\mathbf{k}, \quad \mathbf{P} = \pm P_0\mathbf{i}, \quad P_0 = \sqrt{\frac{|a_0|}{b_0}}, \tag{4.20}$$

with α the constant in (4.15), are limits of minimizers at the limit of K_1 large, and when the polar coefficients ω_p and ω_r dominate over the twist and bending elasticity constants K_2 and K_3. This situation occurs at temperatures lower than those of the helical regime. The role of the surface energy is also relevant in such a case.

Next, we take the following set of admissible fields to determine the ferroelectric limits:

$$\mathbf{n} = \pm\sin\alpha\mathbf{j} + \cos\alpha\mathbf{k}, \quad \mathbf{P} = \pm\sqrt{\frac{|a_0|}{b_0}}\mathbf{i}, \tag{4.21}$$

with $0 < \alpha < \frac{\pi}{2}$, and ω as in (4.15) and (4.18), respectively. We find that the energy $\mathcal{E}_1$ corresponding to such fields is

(4.22)
$$\mathcal{E}_1 = L^2 \left[\left(K_2 \tau^2 + \frac{|a_0|}{b_0} K_3 \right) d + 2 \left(\omega_p + \omega_r + \omega_n - \left(\frac{|a_0|}{b_0} \right)^2 \omega_r - \alpha_0 \omega_n \sin^2 \alpha \right) \right].$$

Replacing $\mathcal{E}_0$ with $\mathcal{E}_1$, the estimates of Theorem 4.1 hold. These allow us to establish the following asymptotic limits of minimizers:

$$\nabla \cdot \mathbf{n} = 0 \quad \text{as} \quad K_1 \to \infty, \tag{4.23}$$

$$|\nabla \omega| = k \quad \text{as} \quad |C_\perp| \to \infty. \tag{4.24}$$

Letting $D \to \infty$ and taking (4.23) into account, it follows that $\triangle \omega = 0$. This together with the boundary conditions on $\partial\Omega$ gives $\nabla\omega = (0, 0, k)$, with k as in (4.1). Moreover, letting $C_a \to \infty$ gives $\cos\alpha = \frac{q}{k}$, and $\mathbf{P} = \sqrt{\frac{|a_0|}{b_0}} \frac{\mathbf{k}\times\mathbf{n}}{|\mathbf{k}\times\mathbf{n}|}$ results by letting $\epsilon \to 0$ and using the expression for $\nabla\omega$. By letting $\omega_r \to \infty$, we get $\phi = \pm\frac{\pi}{2}$.

We finally make the following remarks:

1. The limiting fields $(\mathbf{n}, \mathbf{P}, \omega)$ given by (4.18) and (4.20) satisfy the Euler–Lagrange equations with the prescribed boundary conditions.

2. Likewise, $(\mathbf{n}, \mathbf{P}, \omega)$ as in (4.18) and (4.19) solve the Euler–Lagrange equations at the limit $|C_\perp| \to \infty$.

5. Conclusions. We studied modeling of ferroelectric smectic C* liquid crystals and investigated nonlocal electrostatic effects. We discussed how the proposed model is consistent with well-known approaches found in the physics literature. We proved existence of minimizers for the total energy by means of direct methods of calculus of variations, within the class of fields satisfying physically relevant boundary conditions, with respect to the layering configuration. We presented examples to illustrate the relationship between domain shape and reduction of the nonlocal energy. For instance, we argued that a thin filament may become helical in order to lower the nonlocal energy. We also studied the asymptotic properties of the energy minimizers as the parameters of the energy become unbounded upon the temperature reaching transition values from smectic C* to lower temperature ferroelectric limits.

Acknowledgments. The authors would like to thank professor Marta Lewicka and professor Baisheng Yan for many interesting comments and suggestions, and the referees for carefully reading the manuscript.

REFERENCES

[1] P. Bauman, M. C. Calderer, C. Liu, and D. Phillips, *The phase transition between chiral and smectic A* liquid crystals*, Arch. Ration. Mech. Anal., 165 (2002), pp. 161–186.

[2] P. Biscari and M. C. Calderer, *Telephone-cord instabilities in thin smectic capillaries*, Phys. Rev. E, 71 (2005), paper 051701.

[3] J. Chen and T. C. Lubensky, *Landau-Ginzburg mean-field theory for the nematic to smectic C and nematic to smectic A liquid crystal transitions*, Phys. Rev. A, 14 (1976), pp. 1202–1297.

[4] N. A. Clark and S. T. Lagerwall, *Submicrosecond bistable electro-optic switching in liquid crystals*, Appl. Phys. Lett., 36 (1980), pp. 899–901.

[5] D. A. Coleman, J. Fernsler, N. Chattham, M. Nakata, Y. Takanishi, E. Köblova, D. R. Link, R. F. Shao, W. G. Jang, J. E. Maclennan, O. Mondainn-Monval, C. Boyer, W. Weissflog, G. Petzel, L. C. Chien, J. Zasadzinski, J. Watanabe, D. M. Walba,

H. Takezoe, and N. A. Clark, *Polarization-modulated liquid crystal phases*, Science, 301 (2003), pp. 1204–1211.
[6] P. J. Collings and J. S. Patel, eds., *Handbook of Liquid Crystal Research*, Oxford University Press, New York, 1997.
[7] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, Clarendon Press, Oxford, UK, 1993.
[8] Z. Dogic and S. Fraden, *Development of model colloidal liquid isotropic-smectic transition*, Philos. Trans. Roy. Soc. London Ser. A, 359 (2001), pp. 997–1015.
[9] F. C. Frank, *On the theory of liquid crystals*, Discuss. Faraday Soc., 25 (1958), pp. 19–28.
[10] M. Glogarova, J. Fousek, L. Lejcek, and J. Pavel, *The structure of ferroelectric liquid crystals in planar and its response to electric fields*, Ferroelectrics, 58 (1984), pp. 161–178.
[11] M. Glogarova, L. Lejcek, and J. Pavel, *The influence of an external electric field on the structure of chiral Sm C* liquid crystals*, Mol. Cryst. Liq. Cryst., 91 (1983), pp. 309–325.
[12] M. Glogarova and J. Pavel, *The structure of chiral Sm C* liquid crystals in planar samples and its change in an electric field*, J. Physique, 45 (1984), pp. 143–149.
[13] J. W. Goody, R. Blinc, N. A. Clark, S. T. Lagerwall, M. A. Osipov, S. A. Pikin, T. Sakurai, K. Yoshino, and B. Zeks, *Ferroelectric Liquid Crystals, Principles, Properties and Applications*, Gordon and Breach, New York, 1991.
[14] M. A. Handschy and N. A. Clark, *Structures and responses of ferroelectric liquid crystals in the surface-stabilized geometry*, Ferroelectrics, 59 (1984), pp. 69–116.
[15] R. Hardt, D. Kinderlehrer, and F. H. Lin, *Existence and partial regularity of static liquid crystal configurations*, Comm. Math. Phys., 105 (1986), pp. 547–570.
[16] A. Jákli, Ch. Lischka, W. Weissflog, G. Pelzl, and A. Saupe, *Helical filamentary growth in liquid crystals consisting of banana-shaped molecules*, Liq. Cryst., 27 (2000), pp. 1405–1409.
[17] A. Jákli, D. Krüerke, and G. G. Nair, *Liquid crystal fibers of bent-core molecules*, Phys. Rev. E, 67 (2003), paper 051702.
[18] S. Joo and D. Phillips, *The phase transitions between chiral nematic, smectic A*, and C* liquid crystals*, Comm. Math. Phys., to appear.
[19] A. G. Khachaturyan, *Development of helical cholesteric structure in a nematic liquid crystal due to the dipole-dipole interaction*, J. Phys. Chem. Solids, 36 (1975), pp. 1055–1061.
[20] X. Y. Kong and Z. L. Wang, *Polar-surface dominated ZnO nanobelts and the electrostatic energy induced nanohelixes, nanosprings, and nanospirals*, Appl. Phys. Lett., 84 (2004), pp. 975–977.
[21] S. T. Lagerwall, *Ferroelectric and Antiferroelectric Liquid Crystals*, Wiley-VCH, New York, 1999.
[22] L. D. Landau, E. M. Lifshitz, and L. P. Pitaevskii, *Electrodynamics of Continuous Media*, Butterworth-Heinemann, Oxford, UK, 1998.
[23] L. Longa, D. Monselesan, and H. R. Trebin, *An extension of the Landau-Ginzburg-de Gennes theory for liquid crystals*, Liq. Cryst., 2 (1987), pp. 769–796.
[24] I. Lukyanchuk, *Phase transition between the cholesteric and twist grain boundary C phases*, Phys. Rev. E, 57 (1998), pp. 574–581.
[25] R. B. Meyer, L. Liebert, L. Strzelecki, and P. Keller, *Ferroelectric liquid crystals*, J. Physique Lettres, 36 (1975), pp. L69–L71.
[26] I. Muševič, R. Blinc, and B. Žekš, *The Physics of Ferroelectric and Antiferroelectric Liquid Crystals*, World-Scientific, Singapore, NJ, London, Hong Kong, 2000.
[27] J. Park and M. C. Calderer, *Variational problems and modeling of ferroelectricity in chiral smectic liquid crystals*, in Modeling of Soft Matter, IMA Vol. Math. Appl. 141, Springer, New York, 2005, pp. 169–188.
[28] J. Park and M. C. Calderer, *Phase Transitions between Ferroelectric and Antiferroelectric Liquid Crystal Phases: Static and Dynamical Problems*, preprint, 2004.
[29] J. Pavel, M. Glogarova, and S. S. Bawa, *Dielectric permittivity of ferroelectric liquid crystals influenced by a biasing electric field*, Ferroelectrics, 76 (1987), pp. 221–232.
[30] J. Pavel and M. Glogarova, *The effect of biasing electric field on relaxations in FLC investigated by the dielectric and optical methods*, Ferroelectrics, 121 (1991), pp. 45–53.
[31] J. Pavel, *Behavior of thin planar Sm C* samples in an electric field*, J. Physique, 45 (1984), pp. 137–141.
[32] S. A. Pikin, *Structural Transformations in Liquid Crystals*, Gordon and Breach, New York, 1991.
[33] H. R. Brand, P. E. Cladis, and H. Pleiner, *Fluid biaxial banana smectics: Symmetry at work*, Liquid Crystal Today, 9 (1999), pp. 1–10.
[34] H. R. Brand, P. E. Cladis, and H. Pleiner, *Symmetry and defects in C_M phase of polymeric*

liquid crystals, Macromolecules, 25 (1992), pp. 7223–7226.
[35] Y. Shu and K. Bhattacharya, *Domain patterns and macroscopic behavior of ferroelectric materials*, Phil. Mag. B, 81 (2001), pp. 2021–2051.
[36] R. A. Sones, R. Petschek, D. W. Cronin, and E. M. Terentjev, *Twisting transition in a fiber composed of chiral smectic-C liquid crystal polymer*, Phys. Rev. E, 53 (1996), pp. 3611–3617.
[37] I. Stewart, *The Static and Dynamic Continuum Theory of Liquid Crystals*, Taylor and Francis, London, New York, 2004.
[38] J. V. Stewart, *Intermediate Electromagnetic Theory*, World Scientific, Singapore, NJ, London, Hong Kong, 2001.
[39] T. P. Rieker, N. A. Clark, and S. Lagerwall, *Chevron local layer structure in surface stabilized ferroelectric smectic-C cells*, Phys. Rev. Lett., 59 (1987), pp. 2658–2672.
[40] D. M. Walba, *Ferroelectric liquid crystal conglomerate*, Materials-Chirality, 24 (2003), pp. 457–518.

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 2127–2148

A MUMFORD–SHAH LEVEL-SET APPROACH FOR GEOMETRIC IMAGE REGISTRATION*

MARC DROSKE† AND WOLFGANG RING‡

Abstract. A new method for nonrigid registration of multimodal images is presented. Due to the large interdependence of segmentation and registration, the approach is based on simultaneous segmentation and edge alignment. The two processes are directly coupled and thus benefit from using complementary information of the entire underlying data set. The approach is formulated as a bivariate, variational, free discontinuity problem in the Mumford–Shah framework. A geometric variable describing the contour set and a functional variable which represents the underlying deformation are simultaneously identified. The contour set is represented by a level-set function. We derive a regularized gradient flow and describe an efficient numerical implementation using finite element discretization and multigrid techniques. Finally, we illustrate the method in several applications, such as multimodal intrapatient registration and reconstruction by registration to a reference object.

Key words. image registration, active contours, level-set method, shape sensitivity analysis, Mumford–Shah functional

AMS subject classification. 49F22

DOI. 10.1137/050630209

1. Introduction.

We consider the problem of *image registration* for two given images which show different or complementary features of the same physical reality. Typical situations occur if the two images are obtained using different medical sensoring methods applied to the same patient or if the images show the same anatomical location but are obtained from different patients or from the same patient but at different times. Generally speaking, the aim of image registration is the assignment of complementary anatomical, physical, biological, functional, or other information obtained by different imaging devices to one geometric reference model. To do this, it is necessary to identify corresponding spatial points in the given image domains. The spatial equivalence of points in different images is expressed via a transformation map from one image domain into the other.

The need to register two data sets occurs in various applications, especially in medicine, geophysics, and computer vision. In the last two decades there has been a steep increase in the variety as well as the quality of modern (especially medical) imaging technology, thus making a large amount of information, either anatomical (e.g., computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, densitometry computed tomography (DXA)) or functional (e.g., functional MRI, positron

*Received by the editors April 28, 2005; accepted for publication (in revised form) April 5, 2006; published electronically October 24, 2006. This work was partially supported by the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "Mathematical methods for time series analysis and digital image processing" (SPP 1114). This collaboration was initiated during the participation of the authors at the program "Inverse Problems: Computational Methods and Emerging Application" at the Institute for Pure and Applied Mathematics, University of California, Los Angeles, in 2003.

http://www.siam.org/journals/siap/66-6/63020.html

†Institut für Numerische Simulation, Rheinische Friedrich-Wilhelms-Universität Bonn, Nußallee 15, D-53115 Bonn, Germany (marc.droske@ins.uni-bonn.de).

‡Institute of Mathematics, University of Graz, Heinrichstrasse 36, A-8010 Graz, Austria (wolfgang.ring@uni-graz.at).

emission tomography (PET, SPECT)), available for routine clinical use. The ability to combine complementary information provided by different imaging devices has proved to be of significant benefit to the clinician. For instance, in radiotherapy treatment planning, CT is mainly used, while MRI allows a detailed analysis of the tumor tissue. For all forms of medical treatment where two or more imaging techniques are applied to examine a patient, or where images of different patients are compared and where accurate knowledge about the location of the relevant objects is necessary, registration is an unavoidable task. One of the most important applications of registration techniques is in surgery planning guided by complementary imaging devices. In Figure 3 two different images of the same patient using different types of MRI scanners are shown. The obtained spatial correspondence (the transformation map) between the two images is shown below using a nonlinearly distorted mesh as visualization of the transformation. We mention already at this point that we shall present an approach where features that are present in both images and have a certain type of *geometrical similarity* are mapped onto each other. Thus, reference and template images which have very different intensity distributions but have certain geometrical features in common can be registered. An example using synthetic multimodal images (i.e., images which contain very different information but share certain geometric properties) is presented in Figure 1.

The transformation (the spatial correspondence of like points) itself can also provide important additional information. A transformation between subsequent images obtained at a very low temporal resolution can give insight into growth processes of characteristic objects found in the image. In this context, the variations between subsequent image acquisitions may be substantial, leading to difficulties for differential approaches such as *optical-flow* estimation. Furthermore, the resulting transformation of a registration to a healthy reference dataset may be useful in measuring the extent of the pathology of the individual patient. Scale-space and multiresolution methods have become a widely used methodology for variational registration approaches [21, 35, 36, 46], and a wide range of regularization techniques and similarity measures are known [26, 39, 43, 45, 47, 62].

Depending on the nature of the underlying input data, the *richness* of the space of possible deformations plays an important role. If it is a priori known that the given multichannel data can be registered by a rigid transformation, the unknowns within the representation are only the offset and the angle of the rigid transformation. In that case, regularity of the deformation is automatically guaranteed by the choice of the space of admissible transformations. In many applications, especially interpatient-registration, it is, however, of crucial importance to choose a space of transformations with larger flexibility in order to be able to resolve local variations in fine geometrical details, offering the possibility of a comprehensive analysis of the deformation field. In case of medical time series analysis of a single patient, growth processes of pathological objects such as tumors are of significant interest for diagnosis as well as surgery planning. In what follows, the computational resolution of the discretization of the deformation will be the same as the resolution of the input images, thus allowing a registration of details down to pixel accuracy. However, we want to point out that different (coarser) spaces of deformations may be incorporated in our approach in a straightforward manner.

The paper is organized as follows. In section 2 the approach of coupling registration to segmentation by the Mumford–Shah functional is formulated as a bivariate, variational, free discontinuity problem. Furthermore, different regularization techniques are discussed. In section 3 we present the necessary shape sensitivity analysis

using the conceptional framework of shape derivatives. This eventually leads to the formulation of a gradient flow equation for the given cost functional. To stabilize the shape gradient method, we propose regularizations for both the descent directions for the shape variable and the functional variable. This is done in section 4. In section 5 we will describe the actual algorithm used to compute stationary points of the variational formulation proposed in section 2. Composite finite elements (CFEs) as introduced by Hackbusch and Sauter [34] (see also [55, 61]) provide an elegant approach for the discretization of PDEs on complicated domains and, further, allow one to circumvent numerical difficulties for problems with discontinuous coefficients, especially in the context of multigrid solvers. Since we will treat the variational formulation as a shape optimization problem with contours evolving according to the shape analysis, and—as it turns out—certain elliptic PDEs have to be solved in every gradient step, we chose the CFE framework to incorporate efficient multigrid solvers. This is described briefly in section 6. Finally, computational results are presented in section 7, and a final conclusion is drawn in section 8.

2. Problem formulation. It is the aim of this paper to find a registration between two given images based on a matching of the edges in the images. In their pioneering paper, Mumford and Shah [49] introduced the following energy functional:

$$E_{\mathrm{MS}}(u,\Gamma) = \mu \int_\Omega (u-u_d)^2 \,\mathrm{d}\mathbf{x} + \int_{\Omega\setminus\Gamma} |\nabla u|^2 \,\mathrm{d}\mathbf{x} + \alpha \mathcal{H}^{n-1}(\Gamma), \tag{2.1}$$

where u_d is a given image defined on an open bounded set $\Omega \subset \mathbb{R}^n$ and u is aimed to be an approximation of u_d which should be smooth on $\Omega \setminus \Gamma$, where Γ is the set of potential edges (i.e., subsets of Hausdorff dimension $n-1$ located at singularities of the given image). Here $\mathcal{H}^{n-1}$ denotes the $(n-1)$-dimensional Hausdorff measure, and μ, α are positive weights, which control the balance between data fit, regularization of the reconstruction u on $\Omega\setminus\Gamma$, and the length of the contour Γ, respectively. Existence theory for (2.1) was established after De Giorgi, Carriero, and Leaci [27] proposed considering the minimization of an equivalent energy depending on u only. In this formulation, the energy given by an integral over the entire domain Ω and Γ is represented by S_u, the complement set of Lebesgue points of u, i.e., the measure theoretic discontinuity set of u. It can be proved (cf. Ambrosio, Fusco, Pallara [2, sect. 7.2, pp. 347–354], Braides [9, sect. 2.4, pp. 36–38]) using compactness in $SBV(\Omega)$ and lower-semicontinuity theorems, that—under mild conditions—there exists a solution $u \in SBV(\Omega)$ with $\mathcal{H}^{n-1}(S_u) < \infty$. Here $SBV(\Omega)$ denotes the space of special functions of bounded variation, i.e., functions for which S_u is a σ-finite $(n-1)$-dimensional Borel set. From the numerical point of view, discretizing the singularity set poses a serious problem. Various approximations of the Mumford–Shah functional have been introduced and Γ-convergence results have been proved (cf., e.g., [3, 4, 7, 54]). Ambrosio and Tortorelli [4], for example, proposed a phase-field type regularization and introduced an auxiliary variable which itself is regularized by an elliptic functional. Here, we refer to Feng and Prohl [32] for the numerical analysis of the phase-field approximation. Bourdin and Chambolle [8] proved Γ-convergence of the corresponding discrete finite element schemes.

The Mumford–Shah model has turned out to be very versatile and has been extended and applied in various ways [15, 23, 24, 25, 48, 58]. Esedoglu and Shen [30] suggested an inpainting method based on the Mumford–Shah idea. Further modifications have been made concerning the data-fit term in the Mumford–Shah functional, where the simple L^2 distance has been replaced by more elaborate data-fit criteria

[58]. Recently Unal and Slabaugh [59] have introduced a joint regularization and segmentation algorithm using a piecewise constant Mumford–Shah model in the level-set context which is similar to our approach.

2.1. A Mumford–Shah functional for simultaneous segmentation and registration. In this paper we shall use a Mumford–Shah idea for simultaneously finding the singularity sets in two given images and mapping the respective sets (and with them the two images) onto each other. We do not use a reformulation of the Mumford–Shah functional in the sense of De Giorgi, Carriero, and Leaci [27]. Instead, we will discretize the discontinuity set Γ directly by a level-set function. For the purpose of segmentation and registration we can confine to simple interface sets, which can be elegantly described and propagated via the level-set approach of Osher and Sethian [53]. See also the monographs [56, 51] and the collection [52] for comprehensive introductions to level-set techniques. Level-set methods have been successfully applied in various geometric segmentation models [18, 12, 13, 41, 58, 44, 51, 60]. In [37] Hintermüller and Ring have derived a Newton-type regularized optimization algorithm for minimizing the Mumford–Shah functional by representing Γ by a level-set function.

In our approach to the segmentation-registration problem, the edge sets in the images are found as minimizers of Mumford–Shah functionals and are mapped onto each other by the registration mapping Φ. To be more precise, the edge sets are found in such a way that a level-set encoded contour describes the edge set in the reference image, and, simultaneously, a transformation of the contour by a regular deformation matches the edge set of the template image. This is demonstrated in Figure 1, where only a small part of the edges actually overlap. We emphasize that this viewpoint is different from splitting this process into successively identifying the edge sets first and determining the corresponding deformation which maps these sets onto each other afterwards.

Naturally, the strategy described above would determine the registration map only on the edge set. Therefore, an energy term acting on Φ is added to the Mumford–Shah energy to ensure uniqueness for the registration mapping away from the edge set. As mentioned above, we choose the formulation of the Mumford–Shah energy which is defined for independent geometric and functional variables as described, e.g., in [5, section 4.2.1]. In this formulation, the problem of minimizing the Mumford–Shah functional can be treated as a shape optimization problem and solved numerically using level-set techniques [16, 18, 17, 37]. More precisely, we consider the functional

$$E_{\mathrm{MS}}(\Gamma,\Phi,R,T) = \frac{1}{2}\int_D |R-R_0|^2\,\mathrm{d}\mathbf{x} + \frac{\mu}{2}\int_{D\setminus\Gamma} |\nabla R|^2\,\mathrm{d}\mathbf{x} + \frac{1}{2}\int_D |T-T_0|^2\,\mathrm{d}\mathbf{x} + \frac{\mu}{2}\int_{D\setminus\Gamma^\Phi} |\nabla T|^2\,\mathrm{d}\mathbf{x} + \alpha\mathcal{H}^{N-1}(\Gamma) \tag{2.2}$$

(the additional regularization term on Φ is omitted for the moment). Here $D \subset \mathbb{R}^n$ is the domain of definition of the images with $n = 2, 3$, the data T_0 and R_0 are the given template and reference images, $\Gamma \subset D$ is (an approximation of) the edge set of the given image R_0, and $\Gamma^\Phi = \Phi(\Gamma)$ is the transformed edge set Γ under the transformation Φ. Strictly speaking, the term "edge sets of the data images" does not make sense, since the input images only have to be in L^2. When using this term we mean (approximations of) the measure theoretic discontinuity sets of the SBV functions R and T which *approximate* R_0 and T_0 in the Mumford–Shah sense. In the following we make the simplifying assumption that $\Gamma = \partial\Omega$ for an open set Ω with

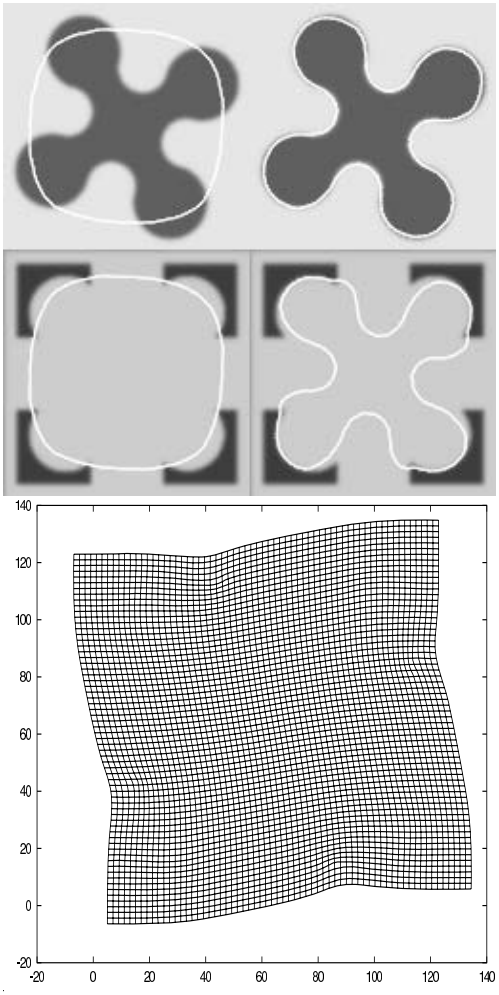


FIG. 1. *Multimodal complementary registration. The two images in the left column of the top figure show the initial contour* Γ *in the images* R *and* T*, where the initial deformation is the identity. In the right column the resulting contours, coupled by the deformation, are shown after* 75 *steps of the regularized gradient descent. The corresponding transformation is shown in the bottom figure using a distorted mesh for visualization.*

$\overline{\Omega} \subset D$. This assumption is justified if the edge sets in the data are related to object boundaries as is usually the case in medical data sets.

Let us point out here that different approaches can be used to drive the contour Γ towards the significant features of the images. A geodesic active contour model as proposed by Caselles, Kimmel, and Sapiro [14], would, e.g., lead to a coupled energy of the form

$$E_{\mathrm{ac}}(\Gamma, \Phi) = \int_{\Gamma} g_R d\mathcal{H}^{n-1} + \nu \int_{\Omega} g_R \, \mathrm{d}\mathbf{x} + \int_{\Gamma^{\Phi}} g_T d\mathcal{H}^{n-1} + \nu \int_{\Omega} g_T \, \mathrm{d}\mathbf{x}, \tag{2.3}$$

where g_R and g_T are suitable edge detectors for the images R and T. A common choice is, for example, $g_I(x) = \frac{1}{1+s|\nabla I|^2}$, $s > 0$, with $I = R$ or $I = T$. The idea of coupling segmentation with registration has also been proposed by Kapur, Yezzi,

and Zöllei [40]. So far the described methods depend on the matching of contours. Yet another approach for the registration of images is the matching of geometrical descriptors as described in [29]. The identification and subsequent matching of geometrical descriptors other than edges within a Mumford–Shah framework, however, have not yet been considered.

As already pointed out, the transformation Φ is not uniquely determined by the functional (2.2). Thus, an additional regularization term E_{reg} is necessary. Writing the transformation $\Phi = \mathbf{id} + \mathbf{d}$ with a displacement vector field $\mathbf{d} : D \to \mathbb{R}^2$, we use $\mathbf{d}$ as the optimization variable instead of Φ. We set

$$E(\Gamma, \mathbf{d}, R, T) = E_{\mathrm{MS}}(\Gamma, \mathbf{d}, R, T) + \nu E_{\mathrm{reg}}(\mathbf{d}). \tag{2.4}$$

In the following we will use Φ and $\mathbf{d}$ synonymously to denote the transformation.

There are a wide range of different choices of regularization energies in the literature. Apart from adding a Dirichlet integral, which corresponds to a regularization as proposed by Horn and Schunk in [38], and anisotropic inhomogeneous regularizations introduced by Nagel and Enkelmann in [50]—both originally appearing in the *optical-flow* context—linearized elastic regularizations are widely used. In [29] Droske and Rumpf proposed a nonlinear elastic polyconvex regularization energy [19] of the form

$$E_{\mathrm{reg}}(\mathbf{d}) = \int_\Omega \alpha \|\nabla \mathbf{d}\|^p + \beta \|\mathbf{Cof}\, \nabla \mathbf{d}\|^q + \gamma(\det \mathbf{d})\, \mathrm{d}\mathbf{x}, \tag{2.5}$$

where $\gamma(s) \to \infty$ for $s \to 0, \infty$. Here $\mathbf{Cof}\, A \in \mathbb{R}^{n\times n}$ denotes the cofactor matrix of a matrix $A \in \mathbb{R}^{n\times n}$. This approach allows one to utilize injectivity techniques for elasticity introduced by Ball [6] in order to ensure that the resulting deformation is a homeomorphism. In the context of aligning feature sets this is particularly important, since we want the transformed contour Γ^Φ to have the same topology as Γ. An extensive discussion of appropriation regularization terms penalizing the departure from rigidity in the context of image registration can be found in Keeling and Ring [42].

At this point of the investigation, the study of different regularization strategies is not our major objective. For the sake of simplicity we will use the Dirichlet integral

$$E_{\mathrm{reg}}(\mathbf{d}) = \|\mathbf{d}\|^2_{\mathbf{H}^1_0(D)} = \int_D \|\nabla \mathbf{d}\|^2\, \mathrm{d}\mathbf{x} \tag{2.6}$$

as regularization term for the remainder of this paper.

2.2. The reduced functional. The functional (2.4) is quadratic in the variables R and T. It is therefore possible to minimize E with respect to R and T for fixed Γ and $\mathbf{d}$ by solving a linear optimality system. With this, we can consider the reduced functional

$$\hat{E}(\Gamma, \mathbf{d}) = E(\Gamma, \mathbf{d}, R(\Gamma), T(\Gamma, \mathbf{d})), \tag{2.7}$$

where $R(\Gamma)$ and $T(\Gamma, \mathbf{d})$ denote the minimizers of (2.4) for fixed Γ and $\mathbf{d}$ with respect to R and T. It is obvious that $R(\Gamma)$ depends only on Γ, whereas $T(\Gamma, \mathbf{d})$ depends also on $\mathbf{d}$ via the domain of integration $D \setminus \Gamma^{\mathbf{d}} = D \setminus \Gamma^{\Phi(\mathbf{d})}$ in the last term in (2.2). If we specify the functional spaces for the variables R and T as $R \in H^1(D \setminus \Gamma)$ and

$T \in H^1(D \setminus \Gamma^{\mathbf{d}})$, we find $R(\Gamma)$ and $T(\Gamma, \mathbf{d})$ as solutions to the system of optimality conditions

$$\begin{aligned} \left\langle \frac{\partial E}{\partial R}, \varphi \right\rangle_{(H^1(D\setminus\Gamma))^*, H^1(D\setminus\Gamma)} &= 0 \quad \text{for all } \varphi \in H^1(D \setminus \Gamma), \\ \left\langle \frac{\partial E}{\partial T}, \psi \right\rangle_{(H^1(D\setminus\Gamma^{\mathbf{d}}))^*, H^1(D\setminus\Gamma^{\mathbf{d}})} &= 0 \quad \text{for all } \psi \in H^1(D \setminus \Gamma^{\mathbf{d}}). \end{aligned} \tag{2.8}$$

This yields

$$\mu \int_{D\setminus\Gamma} \langle \nabla R(\Gamma), \nabla \varphi \rangle \, \mathrm{d}\mathbf{x} + \int_{D\setminus\Gamma} R(\Gamma)\, \varphi \, \mathrm{d}\mathbf{x} = \int_{D\setminus\Gamma} R_0 \, \varphi \, \mathrm{d}\mathbf{x} \tag{2.9}$$

for all $\varphi \in H^1(D \setminus \Gamma)$ and

$$\mu \int_{D\setminus\Gamma^{\mathbf{d}}} \langle \nabla T(\Gamma, \mathbf{d}), \nabla \psi \rangle \, \mathrm{d}\mathbf{x} + \int_{D\setminus\Gamma^{\mathbf{d}}} T(\Gamma, \mathbf{d})\, \psi \, \mathrm{d}\mathbf{x} = \int_{D\setminus\Gamma^{\mathbf{d}}} T_0 \, \psi \, \mathrm{d}\mathbf{x} \tag{2.10}$$

for all $\psi \in H^1(D \setminus \Gamma^{\mathbf{d}})$.

3. Sensitivity analysis. In this section we derive the expressions for the derivatives $\langle \frac{\partial \hat{E}}{\partial \mathbf{d}}, \boldsymbol{\delta} \rangle$ and $d\hat{E}\,((\Gamma, \mathbf{d}), F)$. The latter expression denotes the Eulerian derivative of the functional $\hat{E}$ in the direction of a perturbation vector field of the form $F\,\mathbf{n}_\Gamma$, where $\mathbf{n}_\Gamma$ is the exterior unit normal vector field to Γ. We assume that $\Gamma = \partial\Omega \subset D$, and we specify $\mathbf{n}_\Gamma$ as being the *exterior* unit normal vector field with respect to Ω. See [57, 28] for the concepts of classical shape sensitivity analysis and [37, Appendix A.1] for a more level-set based derivation of the classical results. We refer also to [11], where a framework is presented which includes the concept of topological derivative in the level-set methodology.

3.1. Basic shape derivative formulas. Let us give a brief overview of the calculus of variations for energies which depend on a geometric variable such as a subdomain Ω of a fixed domain D or a submanifold Γ of D. For a smooth vector field $\vec{V} : D \to \mathbb{R}^n$ with $\overline{\Omega} \subset D$ let us first consider the initial value problem

$$\begin{aligned} X'(t) &= \vec{V}(X(t)), \\ X(0) &= X_0 \end{aligned} \tag{3.1}$$

for $X_0 \in D$. The flow $T_t : \Omega \to \mathbb{R}^n$ (with respect to $\vec{V}$) is then defined as $T_t(\mathbf{x}) = X(t)$, where $X(t)$ is the solution of (3.1) with $X_0 = \mathbf{x}$. For a functional $E : \mathcal{E} \to \mathbb{R}$ and a fixed perturbation vectorfield $\vec{V}$, the *Eulerian derivative* is defined by

$$dE(\Gamma; \vec{V}) = \lim_{t \searrow 0} \frac{E(T_t(\Gamma)) - E(\Gamma)}{t} \tag{3.2}$$

provided that the limit exists. Here $\mathcal{E} \subset 2^D$ denotes a suitable set of geometrical variables. The functional E is said to be shape-differentiable at Γ if the limit exists for all $\vec{V} \in B$ and if $dE(\Gamma) \in B'$, i.e., $dE(\Gamma)$ is a bounded linear functional on B, where B is a Banach space of perturbation vector fields. The analogous definitions apply to functions $E(\Omega)$ depending on open sets, not on submanifolds. We will need the following result [57, sect. 2.33, p. 115].

LEMMA 1. *Let Γ be a $\mathcal{C}^2$-hypersurface, and let $f \in H^2_{\mathrm{loc}}(\mathbb{R}^n)$. Then the functional*

$$E(\Gamma) = \int_\Gamma f \,\mathrm{d}\mathcal{H}^{N-1}$$

is shape-differentiable for any perturbation $\vec{V} \in \mathcal{C}^1_0(\mathbb{R}^n)$, and the shape derivative is given by

$$dE(\Gamma;\vec{V}) = \int_\Gamma \left(\nabla f \cdot \vec{V} + f \,\mathrm{div}_\Gamma \vec{V}\right) \mathrm{d}\mathcal{H}^{N-1} \tag{3.3}$$

$$= \int_\Gamma \left(\frac{\partial f}{\partial \mathbf{n}_\Gamma} + f\,\kappa\right) \vec{V} \cdot n_\Gamma \,\mathrm{d}\mathcal{H}^{N-1}, \tag{3.4}$$

where $\mathbf{n}_\Gamma$ denotes the normal to the interface Γ and κ is the additive curvature of Γ.

DEFINITION 1 (material derivative). *We consider a family of (sufficiently smooth) open sets $\mathcal{F}$ and suppose that we are given $f(\Omega) \in B(\Omega)$ for each $\Omega \in \mathcal{F}$, where $B(\Omega)$ is some Banach space of functions on Ω. Let us fix $\Omega_0 \in \mathcal{F}$ and suppose that $\vec{V} \in \mathcal{C}^1_0(\mathbb{R}^n, \mathbb{R}^n)$ is given. We set $\Omega_t = T_t(\Omega_0)$ and assume that $f(\Omega_t) \in B(\Omega_t)$. The limit*

$$\dot{f}(\Omega;\vec{V}) = \lim_{t \searrow 0} \frac{f(\Omega_t) \circ T_t - f(\Omega_0)}{t}$$

is called the (weak) material derivative if it exists in the strong (weak) topology on $B(\Omega_0)$.

DEFINITION 2 (shape derivative). *If the weak material derivative and the expression $\nabla f(\Omega) \cdot \vec{V}$ exist in $B(\Omega)$, then we set*

$$f'(\Omega;\vec{V}) = \dot{f}(\Omega;\vec{V}) - \langle \nabla f(\Omega), \vec{V}\rangle$$

and call it the shape derivative *of f at Ω in direction V.*

In the next section we will also need the following result [57, sect. 2.31, p. 112].

PROPOSITION 1. *Let $f(\Omega)$ be given such that the weak L^1-material derivative $\dot{f}(\Omega;\vec{V})$ and the shape derivative $f'(\Omega;\vec{V}) \in L^1(\Omega)$ exist. Then, the functional*

$$E(\Omega) = \int_\Omega f(\Omega, \mathbf{x}) \,\mathrm{d}\mathbf{x}$$

is shape-differentiable, and the derivative is given by

$$dE(\Omega;\vec{V}) = \int_\Omega f'(\Omega;\vec{V}) \,\mathrm{d}\mathbf{x} + \int_\Gamma f \,\langle \vec{V}, \mathbf{n}_\Gamma\rangle \,\mathrm{d}\mathcal{H}^{N-1}. \tag{3.5}$$

It can be shown (see [28, sect. 3.3, p. 348]) that the various concepts of (first) derivatives with respect to a geometric variable depend on the direction of perturbation V only via its projection

$$F = \langle \vec{V}, \mathbf{n}_\Gamma\rangle \tag{3.6}$$

onto the normal direction to Γ. We therefore subsequently write $dE(\Gamma;F)$ instead of $dE(\Gamma;\vec{V})$ and likewise for the other types of derivatives.

3.2. The first variation of the energy. In the following, we frequently use the coordinate transformation $\mathbf{x} \mapsto \mathbf{y} = \Phi(\mathbf{x}) = \mathbf{x} + \mathbf{d}(\mathbf{x})$ to switch between representations on the transformed and the original configurations. Finding first variations of the functional (2.4) with respect to the geometry Γ requires differentiation with respect to Γ of functionals $\int_\Omega g \, \mathrm{d}\mathbf{x}$ and $\int_{\Phi(\Omega)} h \, \mathrm{d}\mathbf{x}$, respectively, where $\partial\Omega \subset \Gamma$. For integrals of the first type, the results of section 3 directly apply. Suppose that $T_t(\mathbf{x})$ is a flow map which defines a perturbation of Γ with corresponding vector field $\vec{V}$. Then the perturbation of Γ^Φ is given by the flow map $S_t(\mathbf{y}) = \Phi(T_t(\Phi^{-1}(\mathbf{y})))$. The corresponding perturbation vector field has the form $\vec{W}(\mathbf{y}) = (\nabla\Phi \cdot \vec{V})(\Phi^{-1}(\mathbf{y}))$. With this, we can apply the results of section 3 to integrals defined in the transformed configuration. For later use, we recall the following transformation formulas [20, sect. 1.7, pp. 37–41]:

$$\mathbf{n}_{\Gamma^\Phi}(\mathbf{y}) = \frac{\mathbf{Cof}\, \nabla\Phi(\mathbf{x}) \cdot \mathbf{n}_\Gamma(\mathbf{x})}{\|\mathbf{Cof}\, \nabla\Phi(\mathbf{x}) \cdot \mathbf{n}_\Gamma(\mathbf{x})\|}, \tag{3.7a}$$

$$F_\Phi := \langle \vec{W}(\mathbf{y}), \mathbf{n}_{\Gamma^\Phi}(\mathbf{y}) \rangle = \frac{\det\nabla\Phi(\mathbf{x})}{\|\mathbf{Cof}\, \nabla\Phi(\mathbf{x}) \cdot \mathbf{n}_\Gamma(\mathbf{x})\|} \langle \vec{V}(\mathbf{x}), \mathbf{n}_\Gamma(\mathbf{x}) \rangle = \frac{\det\nabla\Phi(\mathbf{x})}{\|\mathbf{Cof}\, \nabla\Phi(\mathbf{x}) \cdot \mathbf{n}_\Gamma(\mathbf{x})\|} F, \tag{3.7b}$$

$$\int_{\Gamma^\Phi} g \, d\mathcal{H}^{N-1} = \int_\Gamma g \circ \Phi \, \|\mathbf{Cof}\, \nabla\Phi \cdot \mathbf{n}_\Gamma\| \, d\mathcal{H}^{N-1}. \tag{3.7c}$$

Proposition 1 (see also [57, sect. 2.31, p. 112]) implies that

$$\begin{aligned} d\hat{E}\,((\Gamma, \mathbf{d}); F) = {} & \frac{1}{2} \int_\Gamma \Big([\![|R(\Gamma) - R_0|^2]\!] + \mu [\![|\nabla R(\Gamma)|^2]\!] \Big) F \, d\mathcal{H}^{N-1} \\ & + \frac{1}{2} \int_{\Gamma^\Phi} \Big([\![|T(\Gamma, \mathbf{d}) - T_0|^2]\!] \, | + \mu [\![|\nabla T(\Gamma, \mathbf{d})|^2]\!] \Big) F_\Phi \, d\mathcal{H}^{N-1} \\ & + \left\langle \frac{\partial E}{\partial R}, R'_F \right\rangle_{(H^1(D\setminus\Gamma))^*, H^1(D\setminus\Gamma)} \\ & + \left\langle \frac{\partial E}{\partial T}, T'_{F_\Phi} \right\rangle_{(H^1(D\setminus\Gamma^\Phi))^*, H^1(D\setminus\Gamma^\Phi)} + \int_\Gamma \kappa \, F \, d\mathcal{H}^{N-1}, \end{aligned} \tag{3.8}$$

where $[\![\cdot]\!]$ denotes magnitudes of jump discontinuities across Γ (from inside to outside) and across Γ_Φ, respectively. As above, κ is the additive curvature of Γ, and R'_F and T'_{F^Φ} are the shape derivatives of R and T in the direction of the perturbation given by F and F_Φ, respectively.

We give a brief explanation of the terms occurring in (3.8). The integrals which compose the reduced cost functional (2.7) can be written as sums of integrals over the individual connected components Ω_i of $D \setminus \Gamma$. Thus, the shape sensitivity analysis of (2.7) can be performed by calculating the sensitivities of integrals over each Ω_i and adding up the results. For each component Ω_i of $D \setminus \Gamma$, boundary integrals of the form $\int_{\partial\Omega_i} f \, \langle \vec{V}, \mathbf{n} \rangle \, d\mathcal{H}^{n-1}$ with appropriate f's occur in the respective shape sensitivities due to Proposition 1. Here the vector $\mathbf{n}$ is the *exterior* normal vector to $\partial\Omega_i$. We have $\partial\Omega_i \subset \Gamma \cup \partial D$ and $\Gamma \cup \partial D = \cup_i \partial\Omega_i$. Thus, the sum of integrals over $\partial\Omega_i$ can be rewritten as an integral over Γ (the contributions on ∂D vanish since

$\vec{V}|_{\partial D} = 0$). Since each component of Γ separates two connected components Ω_i and Ω_j with opposite exterior normal vectors, the jumps of the respective quantities across Γ occur in (3.8).

It can be shown (interpreting the integrals in the weak equations (2.9) and (2.10) as shape functionals and applying Proposition 1; see also [57, sect. 3.2 and 2.29]) that R'_F is the solution to the inhomogeneous Neumann-type boundary value problems

$$\int_{\tilde{\Omega}} \left(\mu \langle \nabla R'_F, \nabla \varphi \rangle + R'_F\, \varphi\right) \mathrm{d}\mathbf{x} = -\int_{\partial\tilde{\Omega}} \left(\mu \langle \nabla_\Gamma R_F, \nabla_\Gamma \varphi \rangle + (R_F - R_0)\, \varphi\right) F\, d\mathcal{H}^{N-1}$$

on each connected component $\tilde{\Omega}$ of $D \setminus \Gamma$ and for all $\varphi \in H^1(\tilde{\Omega})$. An elliptic regularity result then shows that the solution $R'_F|_{\tilde{\Omega}} \in H^1(\tilde{\Omega})$ for each connected component $\tilde{\Omega}$ of $D \setminus \Gamma$ and hence $R'_F \in H^1(D \setminus \Gamma)$. Analogously we obtain $T'_{F_\Phi} \in H^1(D \setminus \Gamma^\Phi)$. Here we need to assume that the transformation Φ is sufficiently smooth. Consequently, we can use R'_F and T'_{F_Φ} as test functions in (2.8) to conclude that the terms including the shape derivatives R'_F and T'_{F_Φ} in (3.8) vanish. Transforming all expressions in (3.8) onto the undeformed configuration (using (3.7)) and replacing Φ by $\mathbf{id} + \mathbf{d}$ yields

(3.9)
$$\begin{aligned} d\hat{E}\left((\Gamma, \mathbf{d}); F\right) =& \frac{1}{2} \int_\Gamma \left([\![|R(\Gamma) - R_0|^2]\!] + \mu [\![|\nabla R(\Gamma)|^2]\!] \right) F\, d\mathcal{H}^{N-1} \\ &+ \frac{1}{2} \int_\Gamma \Big([\![|T(\Gamma, \mathbf{d}) - T_0|^2 \circ (\mathbf{id} + \mathbf{d})]\!] \\ &\qquad + \mu\, [\![|\nabla T(\Gamma, \mathbf{d})|^2 \circ (\mathbf{id} + \mathbf{d})]\!] \Big)\, |\det(I + \nabla \mathbf{d})|\, F\, d\mathcal{H}^{N-1} \\ &+ \alpha \int_\Gamma \kappa\, F\, d\mathcal{H}^{N-1}. \end{aligned}$$

We now consider variation with respect to the displacement $\mathbf{d}$. The cost functional depends on $\mathbf{d}$ via the domain of integration $D \setminus \Gamma^{\mathbf{d}}$ and implicitly via $T(\Gamma, \mathbf{d})$. The perturbation of the geometry has the form $\Gamma^{\mathbf{d}+t\,\boldsymbol{\delta}} = \Gamma^{\mathbf{d}} + t\,\boldsymbol{\delta}(\Gamma) = \Gamma^{\mathbf{d}} + t\left(\boldsymbol{\delta} \circ (\mathbf{id} + \mathbf{d})^{-1}(\Gamma^{\mathbf{d}})\right)$. It can be shown [28, Chap. 7] that this perturbation is equivalent to a perturbation of $\Gamma^{\mathbf{d}}$ with the velocity vector field $\boldsymbol{\delta} \circ (\mathbf{id} + \mathbf{d})^{-1}$. We can therefore apply the results in section 3.1 to obtain

$$\begin{aligned} \left\langle \frac{\partial \hat{E}}{\partial \mathbf{d}}, \boldsymbol{\delta} \right\rangle =& \left\langle \frac{\partial E}{\partial T}, T'\left(\Gamma^{\mathbf{d}}, \boldsymbol{\delta} \circ (\mathbf{id} + \mathbf{d})^{-1}\right) \right\rangle_{(H^1(D\setminus\Gamma^{\mathbf{d}}))^*, H^1(D\setminus\Gamma^{\mathbf{d}})} \\ &+ \frac{1}{2} \int_{\Gamma^{\mathbf{d}}} \left([\![|T(\Gamma, \mathbf{d}) - T_0|^2]\!] + \mu [\![|\nabla T(\Gamma, \mathbf{d})|^2]\!] \right) \\ &\cdot \langle \boldsymbol{\delta} \circ (\mathbf{id} + \mathbf{d})^{-1}, \mathbf{n}_{\Gamma^{\mathbf{d}}} \rangle\, d\mathcal{H}^{N-1} + \nu \left\langle \frac{\partial E_{\mathrm{reg}}}{\partial \mathbf{d}}, \boldsymbol{\delta} \right\rangle. \end{aligned}$$

As above, we argue that $T'\left(\Gamma^{\mathbf{d}}, \boldsymbol{\delta} \circ (\mathbf{id} + \mathbf{d})^{-1}\right) \in H^1(D \setminus \Gamma^{\mathbf{d}})$, and the first term vanishes due to (2.8). If the regularization term (2.7) is used, the Fréchet derivative of $\hat{E}$ with respect to $\mathbf{d}$ in direction $\boldsymbol{\delta}$ reads as

$$\begin{aligned} \left\langle \frac{\partial \hat{E}}{\partial \mathbf{d}}, \boldsymbol{\delta} \right\rangle =& \frac{1}{2} \int_{\Gamma^{\mathbf{d}}} \left([\![|T(\Gamma, \mathbf{d}) - T_0|^2]\!] + \mu [\![|\nabla T(\Gamma, \mathbf{d})|^2]\!] \right) \\ &\cdot \langle \boldsymbol{\delta} \circ (\mathbf{id} + \mathbf{d})^{-1}, \mathbf{n}_{\Gamma^{\mathbf{d}}} \rangle\, d\mathcal{H}^{N-1} + \nu \int_D \nabla \mathbf{d} : \nabla \boldsymbol{\delta}\, \mathrm{d}\mathbf{x}, \end{aligned} \tag{3.10}$$

where ":" stands for the matrix tensor product, i.e., the scalar product corresponding to the Frobenius norm.

We have not yet specified the function space for the displacement field $\mathbf{d}$. The natural choice corresponding to the choice of the regularization E_{reg} would be $\mathbf{d} \in \mathbf{H}_0^1(D) = H_0^1(D, \mathbb{R}^n)$. For the application of shape sensitivity results, however, we need more regularity for the transformation $\Gamma \mapsto (\mathbf{id} + \mathbf{d})(\Gamma)$. This issue will be discussed later on when we define updates for the displacement $\mathbf{d}$.

4. Choice of a descent direction. We now address the question of finding an appropriate descent direction, i.e., a direction $\boldsymbol{\delta}_d \in \mathbf{H}_0^1(D)$ and a scalar function F_d defined on Γ such that

$$\left\langle \frac{\partial \hat{E}}{\partial \mathbf{d}}, \boldsymbol{\delta}_d \right\rangle < 0 \quad \text{and} \quad d\hat{E}\left((\Gamma, \mathbf{d}); F_d\right) < 0. \tag{4.1}$$

The descent directions $\boldsymbol{\delta}_d$ and F_d are found by minimizing the linear approximations over the unit spheres of appropriate function spaces in which the admissible directions are chosen. This idea is closely related to the approach presented in [21, 22], where the descent direction is determined with respect to a regularizing metric. In our case the metric is the scalar product of the adequately chosen Hilbert space. We leave the question of choice of the correct function space for F_d open for the moment and start with finding a descent direction with respect to $\mathbf{d}$. We say that $\boldsymbol{\delta}_d$ is the direction of steepest descent for $\hat{E}$ with respect to $\mathbf{d}$ and the metric induced by the $\mathbf{H}_0^1(D)$-norm if and only if $\boldsymbol{\delta}_d$ is a solution to the constrained optimization problem

$$\min_{\substack{\boldsymbol{\delta} \in \mathbf{H}_0^1(D) \\ \|\boldsymbol{\delta}\|_{\mathbf{H}_0^1(D)} = 1}} \left\langle \frac{\partial \hat{E}}{\partial \mathbf{d}}(\Gamma, \mathbf{d}), \boldsymbol{\delta} \right\rangle. \tag{4.2}$$

Note that $\left\langle \frac{\partial \hat{E}}{\partial \mathbf{d}}(\Gamma, \mathbf{d}), \cdot \right\rangle$ defines a bounded linear functional on $\mathbf{H}_0^1(D)$ provided that $\mathbf{d}$ is smooth enough. To solve (4.2), we introduce the Lagrange function

$$\begin{aligned}
&\mathcal{L}_f(\boldsymbol{\delta}, \lambda_f) = \\
&\quad \frac{1}{2} \int_{\Gamma_{\mathbf{d}}} \left([\![|T(\Gamma, \mathbf{d}) - T_0|^2]\!] + \mu [\![|\nabla T(\Gamma, \mathbf{d})|^2]\!] \right) \langle \boldsymbol{\delta} \circ (\mathbf{id} + \mathbf{d})^{-1}, \mathbf{n}_{\Gamma_{\mathbf{d}}} \rangle \, d\mathcal{H}^{N-1} \\
&\qquad + \nu \int_D \nabla \mathbf{d} : \nabla \boldsymbol{\delta} \, \mathrm{d}\mathbf{x} + \lambda_f \left(\int_D |\nabla \boldsymbol{\delta}|^2 \, \mathrm{d}\mathbf{x} - 1 \right).
\end{aligned}$$

The optimality system for (4.2) reads as $\frac{\partial \mathcal{L}_f}{\partial \boldsymbol{\delta}}(\boldsymbol{\delta}_d, \lambda_d) = 0$ and $\frac{\partial \mathcal{L}_f}{\partial \lambda_f}(\boldsymbol{\delta}_d, \lambda_d) = 0$. Therefore, the direction of steepest descent $\boldsymbol{\delta}_d$ is found as the solution to

$$\begin{aligned}
\int_D \nabla \boldsymbol{\delta}_d : \nabla \boldsymbol{\xi} \, \mathrm{d}\mathbf{x} = -\frac{1}{\lambda_f} \Bigg(&\nu \int_D \nabla \mathbf{d} : \nabla \boldsymbol{\xi} \, \mathrm{d}\mathbf{x} \\
&+ \frac{1}{2} \int_{\Gamma_{\mathbf{d}}} \left([\![|T(\Gamma, \mathbf{d}) - T_0|^2]\!] + \mu [\![|\nabla T(\Gamma, \mathbf{d})|^2]\!] \right) \langle \boldsymbol{\xi} \circ (\mathbf{id} + \mathbf{d})^{-1}, \mathbf{n}_{\Gamma_{\mathbf{d}}} \rangle \, d\mathcal{H}^{N-1} \Bigg)
\end{aligned} \tag{4.3}$$

for all $\boldsymbol{\xi} \in \mathbf{H}_0^1(D)$, where the multiplier λ_f is chosen such that $\|\boldsymbol{\delta}_d\|_{\mathbf{H}_0^1(D)} = 1$.

Alternatively one might want to allow $\boldsymbol{\delta}_d \in \mathbf{H}^1$ instead of prescribing homogeneous Dirichlet conditions, which can be particularly important in case of large

translations between the reference and the template image. Then $\boldsymbol{\delta}_d$ is given as the solution to

$$\begin{aligned}(4.4)\quad \int_D \big(\langle \boldsymbol{\delta}_d, \boldsymbol{\xi}\rangle + \nabla\boldsymbol{\delta}_d : \nabla\boldsymbol{\xi}\big)\,\mathrm{d}\mathbf{x} = -\frac{1}{\lambda_f}\Bigg(\nu\int_D \nabla\mathbf{d} : \nabla\boldsymbol{\xi}\,\mathrm{d}\mathbf{x} \\ + \frac{1}{2}\int_{\Gamma^{\mathbf{d}}} \big(\llbracket |T(\Gamma,\mathbf{d}) - T_0|^2\rrbracket + \mu\llbracket|\nabla T(\Gamma,\mathbf{d})|^2\rrbracket\big)\,\langle \boldsymbol{\xi}\circ(\mathbf{id}+\mathbf{d})^{-1}, \mathbf{n}_{\Gamma^{\mathbf{d}}}\rangle\, d\mathcal{H}^{N-1}\Bigg)\end{aligned}$$

for all $\boldsymbol{\xi} \in \mathbf{H}^1(D)$.

Application of the transformation rules (3.7c) and (3.7b) to the surface integral on the right-hand side of (4.3) and (4.4) yields

$$(4.5)\quad \frac{1}{2}\int_\Gamma \big(\llbracket |T(\Gamma,\mathbf{d}) - T_0|^2\rrbracket + \mu\llbracket|\nabla T(\Gamma,\mathbf{d})|^2\rrbracket\circ(\mathbf{id}+\mathbf{d})\big)\,\langle \boldsymbol{\xi}, \mathbf{Cof}\,\nabla\mathbf{d}\cdot\mathbf{n}_\Gamma\rangle\, d\mathcal{H}^{N-1}$$

for these terms.

We now make a few comments concerning the regularity of the displacement $\mathbf{d}$. The update $\boldsymbol{\delta}_d$ which solves (4.3) is a function in $\mathbf{H}^1_0(D)$ which—in general—does not possess much additional regularity since the source term is a distribution which is localized on Γ, thus introducing a singularity $\boldsymbol{\delta}_d$ along Γ. For the above shape sensitivity results to hold, we require that the displacement be smooth in every step. To circumvent this difficulty, we can replace $\boldsymbol{\delta}_d$ by a smooth approximation $\boldsymbol{\delta}_d^\epsilon$ for the actual update of the transformation. If the approximation is close enough in the $\mathbf{H}^1_0(D)$-norm, the descent property (4.1) will still be satisfied and the theoretical arguments are justified. In the numerical realization it turns out that smoothing of the transformation is not necessary.

To find a descent direction for the geometrical variable Γ, we first have to specify the function space and the corresponding metric for the update direction F_d. By choosing the update direction $\boldsymbol{\delta}_d \in \mathbf{H}^1_0(D)$, the movement of the transformed geometry $\Gamma^{\mathbf{d}} \mapsto \Gamma^{\mathbf{d}+t\boldsymbol{\delta}_d}$ corresponds to a movement in normal direction with speed function given by

$$F_{\boldsymbol{\delta}_d} = \langle \boldsymbol{\delta}_d, \mathbf{n}_{\Gamma^{\mathbf{d}}}\rangle \in H^{\frac{1}{2}}(\Gamma^{\mathbf{d}}).$$

It is therefore natural to choose the descent direction F also with respect to the $H^{\frac{1}{2}}$-norm on Γ. This choice should give a good balance between the descents achieved by moving the geometrical variable Γ and the functional variable $\mathbf{d}$, respectively. More precisely, we choose the descent direction F_d as the solution to the problem

$$(4.6)\quad \min_{\substack{F\in H^{\frac{1}{2}}(\Gamma)\\ \|F\|_{H^{\frac{1}{2}}(\Gamma)}=1}} dE\big((\Gamma,\mathbf{d}), F\big).$$

We introduce again a Lagrange function

$$\mathcal{L}_g(F,\lambda_g) = d\hat{E}\,((\Gamma,\mathbf{d}); F) + \lambda_g\left(\|F\|^2_{H^{\frac{1}{2}}(\Gamma)} - 1\right).$$

The optimality system for F_d then has the form

$$(4.7)\quad (F_d, G)_{H^{\frac{1}{2}}(\Gamma)} = -\frac{1}{\lambda_g} d\hat{E}\,((\Gamma,\mathbf{d}); G)$$

for all $G \in H^{\frac{1}{2}}(\Gamma)$. To evaluate the inner product $(\cdot,\cdot)_{H^{\frac{1}{2}}(\Gamma)}$ we consider the boundary value problem

$$\begin{aligned} -\Delta v + v &= 0 \text{ on } \Omega, \\ \frac{\partial v}{\partial n}\Big|_\Gamma &= H \in H^{-\frac{1}{2}}(\Gamma). \end{aligned} \tag{4.8}$$

Here $\Omega \subset D$ is chosen such that $\Gamma = \partial\Omega$. In the level-set context below, Ω can be chosen as the set of all points with negative function values of the level-set function. The weak formulation for (4.8) is given by

$$\int_\Omega \left(\langle \nabla v, \nabla \varphi\rangle + v\,\varphi\right) \mathrm{d}\mathbf{x} = \left\langle H, \varphi|_\Gamma \right\rangle_{H^{-\frac{1}{2}}(\Gamma), H^{\frac{1}{2}}(\Gamma)} \tag{4.9}$$

for all $\varphi \in H^1(\Omega)$. We define the Neumann-to-Dirichlet map for the operator $-\Delta +$ id on Ω as the linear operator $\mathcal{N} : H^{-\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma)$, which maps H in (4.8) to the Dirichlet trace $v|_\Gamma$ of the solution to (4.8). The operator $\mathcal{N}$ is invertible with $\mathcal{N}^{-1}$ given by the composition of the solution operator of the Dirichlet problem with inhomogeneous boundary conditions and the Neumann trace operator. By the closed graph theorem, $\mathcal{N}^{-1}$ is bounded; hence $\mathcal{N}$ constitutes an isomorphism between the spaces $H^{-\frac{1}{2}}(\Gamma)$ and $H^{\frac{1}{2}}(\Gamma)$. An inner product on $H^{\frac{1}{2}}(\Gamma)$ can be defined as

$$(F, G)_{H^{\frac{1}{2}}(\Gamma)} = \langle \mathcal{N}^{-1} F, G\rangle_{H^{-\frac{1}{2}}(\Gamma), H^{\frac{1}{2}}(\Gamma)}.$$

With this, we can write (4.7) as

$$\langle \mathcal{N}^{-1} F_d, G\rangle_{H^{-\frac{1}{2}}(\Gamma), H^{\frac{1}{2}}(\Gamma)} = -\frac{1}{\lambda_g} d\hat{E}\left((\Gamma, \mathbf{d}); G\right)$$

for all $G \in H^{\frac{1}{2}}(\Gamma)$. If we use (3.9), we obtain

$$\begin{aligned} F_d &= -\frac{1}{\lambda_g}\mathcal{N} F_g \text{ with} \\ F_g &= \frac{1}{2}\Big([\![|R(\Gamma) - R_0|^2]\!] + \mu [\![|\nabla R(\Gamma)|^2]\!] \\ &\quad + \left([\![|T(\Gamma, \mathbf{d}) - T_0|^2 \circ (\mathbf{id} + \mathbf{d})]\!] + \mu [\![|\nabla T(\Gamma, \mathbf{d})|^2 \circ (\mathbf{id} + \mathbf{d})]\!] \right) |\det(I + \nabla \mathbf{d})| \Big) \\ &\quad + \alpha\kappa \end{aligned} \tag{4.10}$$

on Γ.

5. Description of the algorithm. Let us now assemble all the discussed main building blocks into a regularized shape gradient descent algorithm within the level-set framework:

Step 1 Choose an initial level-set function u_0, choose an initial transformation $\mathbf{d}_0$. Set $u_c = u_0$, $\mathbf{d}_c = \mathbf{d}_0$.

Step 2 For the current level-set function u_c set $\Omega = \{\mathbf{x} \in D : u_c(\mathbf{x}) < 0\}$, $\Gamma = \{\mathbf{x} \in D : u_c(\mathbf{x}) = 0\}$. Solve (2.9) and (2.10) for R and T, respectively, for the current Γ and $\mathbf{d}_c$.

Step 3 Evaluate the expression F_g in (4.10).

Step 4 Solve the elliptic equation (4.8) with Neumann data given by F_g. Evaluate the Dirichlet trace to get F_d on Γ.
Step 5 Extend F_d to a function F_d^{ext} which is defined on a narrow band around Γ.
Step 6 Solve (4.3) for $\boldsymbol{\delta}_d$.
Step 7 Solve the level-set equation

$$u_t + F_d^{\text{ext}}|\nabla u| = 0 \quad \text{with } u(\cdot,0) = u_c.$$

Set the new $u_c = u(\cdot,\tau)$. Set $\mathbf{d}_c = \mathbf{d}_c + \tau\boldsymbol{\delta}_d$. Choose the step size τ according to a line search procedure.
Step 8 Stopping criterion. Else go to Step 2.

The finite element approximations of the functions R and T and the auxiliary variable v in (4.9) on the irregular domains $D \setminus \Gamma$, $D \setminus \Gamma^{\mathbf{d}}$, and Ω are done using composite finite elements (cf. [34]). The transformation vector field $\mathbf{d}$ is discretized using standard finite elements.

5.1. Step 2. Equations (2.9) and (2.10) are solved using CFEs for the solution of the second order elliptic equations on the variable and irregular domains $D \setminus \Gamma$ and $D \setminus \Gamma^{\mathbf{d}}$. The CFE code takes as input the function values of a level-set function, which defines the variable geometry, on a rectangular grid. For $D \setminus \Gamma$ the level-set function is given by u_c. For the transformed geometry $D \setminus \Gamma^{\mathbf{d}}$ a level-set function is given by $u_c^{\mathbf{d}} = u_c \circ (\mathbf{id} + \mathbf{d})^{-1}$. We introduce a triangulation $\mathcal{T}$ on D and approximate $\mathbf{d}$ by a piecewise affine transformation on $\mathcal{T}$.

5.2. Step 3. The data F_g are processed further in Step 4 as Neumann boundary data in (4.8). It follows from (4.9) that the data are used in the form $\int_\Gamma F_g\,\varphi_n\,d\mathcal{H}^{N-1}$ for all finite element basis functions φ_n. Hence, it is useful to determine the values of F_g on the intersection points of the rectangular finite element grid with Γ.

In F_g the jumps $R_i - R_e$ and $\nabla R_i - \nabla R_e$ occur, where R_i (interior) is the solution to (2.9) on Ω and R_e (exterior) is the solution to (2.9) on $D \setminus \overline{\Omega}$. We get the function values for R_i and R_e at the intersection points in a straightforward way from the respective finite element representations.

5.3. Step 4. In order to calculate F_d in (4.10), we solve

$$\langle \nabla\phi_i^{CFE}, \nabla\phi_j^{CFE}\rangle_{L^2(\Omega_1)}\bar{F}_{d,i} + \langle \phi_i^{CFE}, \phi_j^{CFE}\rangle_{L^2(\Omega_1)}\bar{F}_{d,i} = -\frac{1}{\lambda_g}\langle F_g, \phi_i\rangle_{L^2(\partial\Omega_1)},$$

where ϕ_i^{CFE} denotes the basis functions of the finite element space (cf. section 6), and $\bar{F}_{d,i}$ denotes the ith component of the vector $\bar{F}_d$, i.e., the coefficient vector of F_d with respect to the chosen basis.

5.4. Step 5. We now extend F_d given from the discrete contour Γ_h to a function F_d^{ext} defined on a neighborhood of Γ_h by solving the following transport equation:

$$\langle \nabla F_d^{\text{ext}}, \nabla d_\Gamma\rangle = 0 \quad \text{on} \quad \Omega \qquad \text{and} \qquad F_d^{\text{ext}} = F_d \quad \text{on} \quad \Gamma. \tag{5.1}$$

Here d_Γ denotes the signed distance function to Γ. Note that d_Γ and F_d^{ext} can be computed simultaneously by a modified fast marching method for solving the eikonal equation $|\nabla d| = 1$ (cf. [1] for a description of the algorithm).

5.5. Step 7. The discretization of the level-set equation

$$\partial_t u + F_d^{\text{ext}}|\nabla u| = 0 \quad \text{on} \quad \Omega \tag{5.2}$$

is carried out using an explicit upwind scheme. In our computations we have applied a third order accurate ENO-scheme (cf. [51, sect. 3.4, p. 33]).

6. Composite finite elements and multigrid. In this section we will briefly describe the spatial discretization of the H^1 function spaces on Ω_i, which are divided by the contour Γ, i.e., the zero level-set of u. Furthermore, we outline a multigrid method for the solution of (2.9), (2.10), (4.3), and (4.9). We use CFEs introduced by Hackbusch and Sauter [34]. Instead of resolving the Ω_i using a retriangulation or local adaptive refinement, we confine ourselves to a uniform quadrilateral (resp., hexahedral) grid $\mathcal{T}$ and define the triangulations $\mathcal{T}_i \subset \mathcal{T}$ for Ω_i with $\Omega_i \subseteq \overline{\bigcup_{T\in\mathcal{T}_i} T}$ by the following overlap condition:

$$T \in \mathcal{T}_i \iff T \in \mathcal{T}, \quad T \cap \Omega_i \neq \emptyset. \tag{6.1}$$

Let us denote by $V_h(\Omega_{\mathcal{T}})$ the usual finite element space given by the condition that for $U \in V_h(\Omega_{\mathcal{T}})$, $U\big|_T$ is a multilinear function for each $T \in \mathcal{T}$. The corresponding CFE space is then given by the restriction of the functions in $V_h(\Omega_{\mathcal{T}_i})$ to the domain Ω_i, i.e.,

$$V_h^{CFE}(\Omega_{\mathcal{T}_i}) := \{U|_{\Omega_i} \,|\, U \in V_h(\Omega_{\mathcal{T}_i})\}. \tag{6.2}$$

Hence, a basis $(\varphi_i^{CFE})_i$ of V_h^{CFE} is given by $\varphi_i^{CFE} := \varphi_i\big|_{\Omega_{\mathcal{T}_i}}$, where $(\varphi_i)_i$ denotes a basis of the space $V_h(\Omega_{\mathcal{T}_i})$.

For the assembling of the mass matrix $M_i = (\int_{\Omega_i} \varphi_i^{CFE}\varphi_j^{CFE}\,\mathrm{d}\mathbf{x})_{ij}$ and stiffness matrix $L_i = (\int_{\Omega_i} \nabla\varphi_i^{CFE}\nabla\varphi_j^{CFE}\,\mathrm{d}\mathbf{x})_{ij}$ we need to apply quadrature rules for functions on $T \cap \Omega_i$. On each cell T, which is crossed by the zero level-set of u, we generate on-the-fly a partition of $T \cap \Omega_i$ into simplices and apply a barycenter quadrature rule on each simplex.

In order to apply a multigrid method, we generate a sequence of nested CFE spaces by applying an appropriate coarsening process on the CFE triangulation on the finest level $l_{\max}$ ($\Omega_{\mathcal{T}_i}^{l_{\max}} := \Omega_{\mathcal{T}_i}$), i.e.,

$$\Omega_i \subset \Omega_{\mathcal{T}_i}^{l_{\max}} \subset \Omega_{\mathcal{T}_i}^{l_{\max}-1} \subset \cdots \subset \Omega_{\mathcal{T}_i}^{0}, \tag{6.3}$$

leading to correspondingly nested CFE spaces $V_h^{CFE}(\Omega_{\mathcal{T}_i}^{l})$, $0 \le l \le l_{\max}$. Prolongations and restrictions naturally have to be defined with respect to the CFE discretization; hence prolongation onto level l is defined by evaluation of the basis functions $\varphi_i^{CFE,l-1}$ for Lagrange nodes on level l. Convergence analysis for multigrid algorithms using CFEs has been presented by Hackbush and Sauter [34] and Warnke [61], and we refer to [33] for a comprehensive overview of geometric multigrid methods.

7. Numerical experiments. We have tested our approach in different scenarios. Figure 1 shows a synthetic image pair, which was designed to test the method in cases where only *very little common information* is contained in the images. The rotated shape on the upper left of the top figure is supposed to be fitted into the structure on the bottom left, which is hence determined only by the four small objects in the corners of the image. After 75 steps of the gradient descent, a deformation is found which rotates the propeller-like shape, and the resulting push-forward of the contour matches quite well to the rounded corners in the second structure in the bottom of the top figure. Hence this example shows the capability of a model-based inpainting, where the shape information of the inpainted contour is transformed from a reference image. Thus, complementary information originating from the first image, which is not present in the second image, is used for the segmentation of the second image. We assume that the deviations from the obvious solution of a pure rotation result from

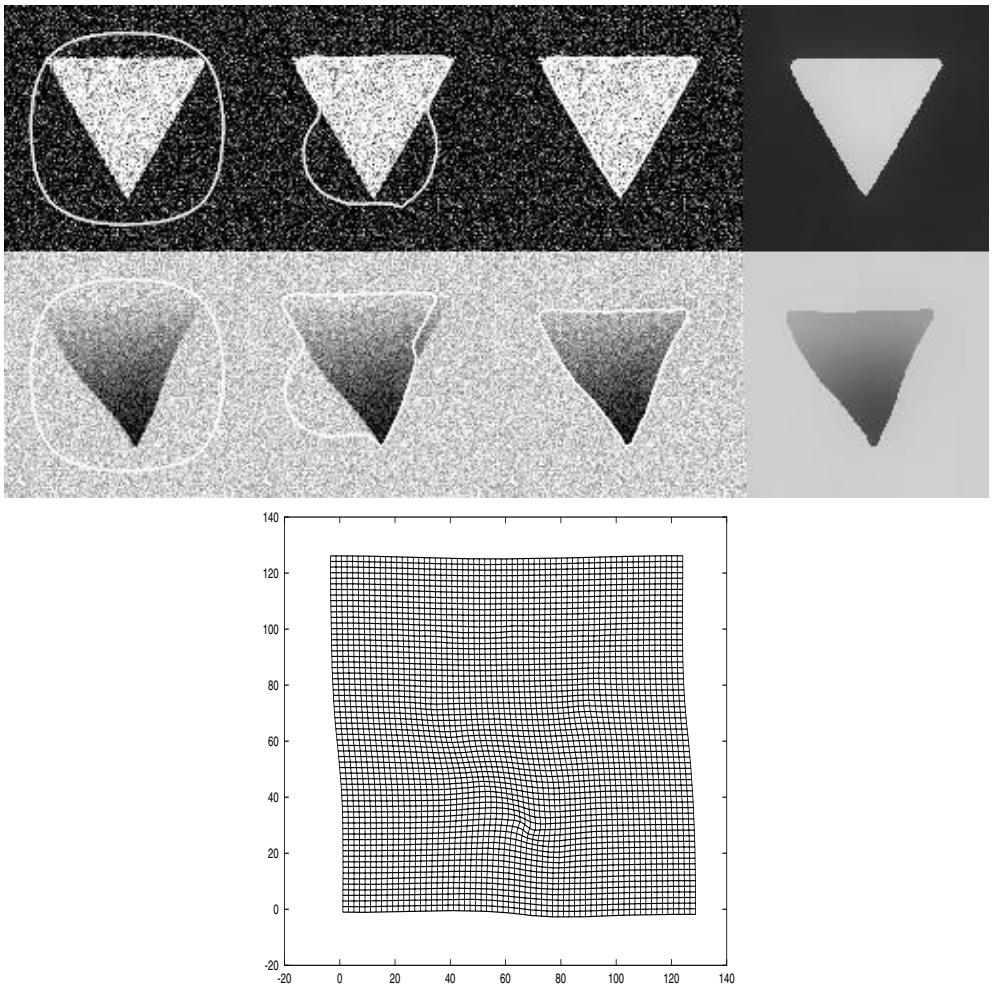


FIG. 2. *Simultaneous segmentation, registration, and denoising of two artificial test images. The smoothed reconstructions of both images are shown on the right-hand side of the top part of the figure.*

the fact that rigid transformations are not in the kernel of our regularization energy. We think that the result can be improved using a different regularization method, for example, a higher order method [46].

Figure 2 shows again the ability of the method to use complementary information from both images. In this situation, both images are contaminated by noise. The weak upper edge of the triangle in the second image is found accurately using information from the first image, where the corresponding edge is rather strong. As a by-product we also obtain smoothed reference and template images shown on the right-hand side of Figure 2 (top figure), where the edges detected by the segmentation are strongly enhanced. Note that the smoothed second image has a clearly recognizable upper edge.

In Figure 3 we have applied the algorithm to a pair of brain images. The top row shows a positron density (PD) scan, while the bottom row shows a T1-weighted

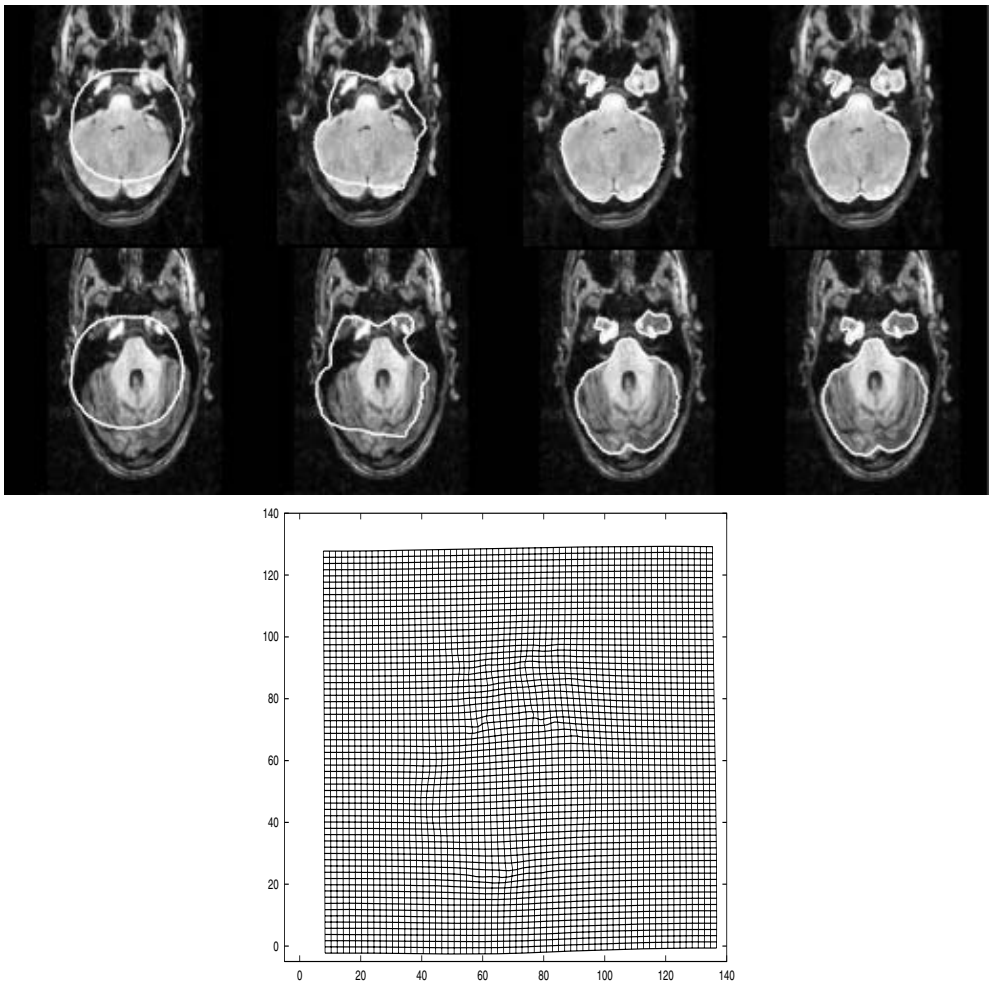


FIG. 3. *Top row: Reference image R (PD image of a human brain) and Γ. Bottom row: Template image T ($T1$-weighted MR image). The sequence shows the gradient descent for the iteration numbers $0, 10, 80$, and 180. The parameters were chosen as $\mu = 50$, $\alpha = 20$, $\nu = 500$. Images from the MPI of Cognitive Neuroscience, Leipzig, Germany.*

magnetic resonance image of the same patient. The initial difference of the image pair consists mainly of a translation of about 8-9 pixels. The algorithm finds the brain structure in both images well after about 80 steps, and the resulting deformation consists mainly of a shift enhanced by some minor locally detailed deformations. This example underlines the practicability of the level-set approach: After a few steps the initial contour splits into three different components which are henceforth independently mapped onto the corresponding segments in the template. Note also that the strong contrast between the brain and the other tissue in the first image helps to segment the outer brain boundary in the second image, where the contrast is not very strong and other edges are found in the vicinity of the boundary of the brain.

The last example in Figure 4 demonstrates the competing effect of the regularization and the energy contributions which pull the contour towards the edges. We can

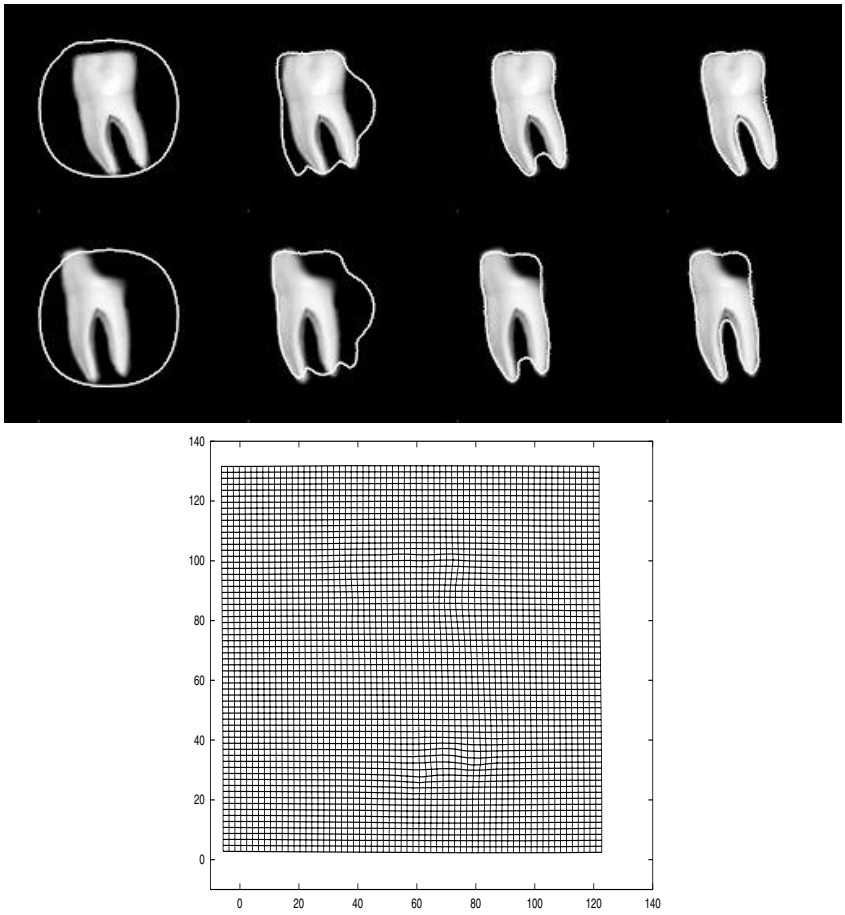


FIG. 4. *Matching a reference to a template with a destroyed region. The image in the bottom row of the top figure also differs nonrigidly from the image in the top row. The edge set of the reference is matched onto the contours of the template, where the alignment does not strongly decrease the regularity of the deformation. The parameters were chosen as* $\mu = 50$, $\alpha = 200$, *and* $\nu = 5000$. *The bottom image shows the deformation plot.*

exploit this in order to map an original reference shape (top row) to a given object, where the shape is partially corrupted (bottom row). Apart from the destroyed region, the shapes differ also by a nonrigid deformation plus a translation. This can be well observed in the second column. Here the deformation is still close to the identity, and hence the contours are aligning to the edges in the vicinity first until in subsequent iterations the deformation evolves in such a way that the contours map to the true edges in both images except on the border of the destroyed region. At this stage, the regularization dominates and prohibits the contour in the bottom row evolving towards the "visible" edge and prefers to adopt the contour from the reference image. This yields a reconstruction of the destroyed shape, which is optimal with respect to the regularization energy.

In this last example the optimal contours found by the algorithm depend significantly on the choice of the regularization parameter ν. The results shown in Figure 4 are obtained with a rather large value $\nu = 5000$, i.e., with a strong penalty on the

rigidity of the transformation Φ. If ν is chosen smaller, both contours become aligned with the edges in the images, which is only possible with a locally large deformation in the vicinity of the missing part of the tooth in the template image.

For all examples, the parameter μ does not have much influence on the segmentation and on the obtained transformation. It does, however, influence the character of the piecewise smooth approximations of the reference and template images R and T, respectively. Large μ yields approximations which are almost piecewise constant, whereas small μ allows a larger intensity variation within each segmented component. Finally, the algorithm seems to be rather insensitive with respect to the choice of the length parameter α.

8. Discussion. We have presented a level-set based algorithm for simultaenous segmentation and registration of images by incorporating a Mumford–Shah type energy on the reference image as well as on the template image, where the contour is transformed into the template image by a regularized deformation. The work presented here is motivated by the fact that, given an exact registration of two images of different modality, edge extraction and segmentation can be enhanced considerably by combining complementary feature information from both modalities. On the other hand, the process of registering a pair of images may rely on segmentations and feature extractions of both images, which often is a very tedious process, especially if, in some areas, the feature information is very weak. Due to the coupling of the edge sets by the smooth deformation, the edge in such areas is driven towards its correct shape.

We have demonstrated a further important application of this method, namely that this approach may also be used to perform a model-based reconstruction and inpainting of destroyed regions, without having to explicitly mark the region where the object is destroyed as long as there are no prominently dominating edges. Although the results are already very promising, there is still room for further conceptual modelling, e.g., to avoid competition of the broken edge and the reference edge along the boundary of the destroyed region.

Due to the regularization of the gradient flow, the minimization process has turned out to be stable and requires only a relatively small number of iterations until convergence. On the other hand, the regularization and necessity of determining the solutions of the Helmholtz equations in the regions Ω_1 and Ω_2 make each individual step rather expensive. In order to make the method efficient we have applied multigrid techniques which lead to an enormous speed-up of the algorithm.

We have performed all calculations using only the first variations of the energy. In further studies, one might investigate Levenberg–Marquart (cf. [10]) or pure Newton-type methods to further accelerate the minimization process.

Acknowledgments. The authors would like to thank Martin Rumpf for many helpful comments and Florian Liehr for his valuable advice concerning the composite finite element method. Furthermore, we want to thank Carlo Schaller for many stimulating discussions on medical imaging and computer-aided surgery.

REFERENCES

[1] D. Adalsteinsson and J. A. Sethian, *The fast construction of extension velocities in level set methods*, J. Comput. Phys., 148 (1999), pp. 2–22.

[2] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford University Press, New York, 2000.

[3] L. Ambrosio and V. M. Tortorelli, *Approximation of functionals depending on jumps by elliptic functionals via* Γ*-convergence*, Comm. Pure Appl. Math., 43 (1990), pp. 999–1036.
[4] L. Ambrosio and V. M. Tortorelli, *On the approximation of free discontinuity problems*, Boll. Un. Mat. Ital. B (7), 6 (1992), pp. 105–123.
[5] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing. Partial Differential Equations and the Calculus of Variations*, Springer-Verlag, New York, 2002.
[6] J. M. Ball, *Global invertibility of Sobolev functions and the interpenetration of matter*, Proc. Roy. Soc. Edinburgh Sect. A, 88 (1988), pp. 315–328.
[7] A. Bonnet, *On the regularity of the edge set of Mumford-Shah minimizers*, Progr. Nonlinear Differential Equations Appl., 25 (1996), pp. 93–103.
[8] B. Bourdin and A. Chambolle, *Implementation of an adaptive finite-element approximation of the Mumford-Shah functional*, Numer. Math., 85 (2000), pp. 609–646.
[9] A. Braides, *Approximation of Free-Discontinuity Problems*, Lecture Notes in Math. 1694, Springer-Verlag, Berlin, 1998.
[10] M. Burger, *Levenberg-Marquardt level set methods for inverse obstacle problems*, Inverse Problems, 20 (2004), pp. 259–282.
[11] M. Burger, B. Hackl, and W. Ring, *Incorporating topological derivatives into level set methods*, J. Comput. Phys., 194 (2004), pp. 344–362.
[12] V. Caselles, F. Catté, T. Coll, and F. Dibos, *A geometric model for active contours in image processing*, Numer. Math., 66 (1993), pp. 1–31.
[13] V. Caselles and B. Coll, *Snakes in movement*, SIAM J. Numer. Anal., 33 (1996), pp. 2445–2456.
[14] V. Caselles, R. Kimmel, and G. Sapiro, *Geodesic active contours*, Internat. J. Comput. Vision, 22 (1997), pp. 61–79.
[15] A. Chambolle, *Image segmentation by variational methods: Mumford and Shah functional and the discrete approximations*, SIAM J. Appl. Math., 55 (1995), pp. 827–863.
[16] T. F. Chan and L. A. Vese, *Image Segmentation Using Level Sets and the Piecewise Constant Mumford-Shah Model*, UCLA CAM Report 00-14, University of California, Los Angeles, 2000.
[17] T. F. Chan and L. A. Vese, *Active contours without edges*, IEEE Trans. Image Process., 10 (2001), pp. 266–277.
[18] T. F. Chan and L. A. Vese, *A level set algorithm for minimizing the Mumford-Shah functional in image processing*, in IEEE/Computer Society Proceedings of the 1st IEEE Workshop on Variational and Level Set Methods in Computer Vision, 2001, pp. 161–168.
[19] Ph. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[20] P. G. Ciarlet, *Mathematical Elasticity. Vol.* I. *Three-Dimensional Elasticity*, Stud. Math. Appl. 20, North-Holland, Amsterdam, 1988.
[21] U. Clarenz, M. Droske, and M. Rumpf, *Towards fast non-rigid registration*, in Inverse Problems, Image Analysis and Medical Imaging, Contemp. Math. 313, AMS, Providence, RI, 2002, pp. 67–84.
[22] U. Clarenz, S. Henn, K. Rumpf, and M. Witsch, *Relations between optimization and gradient flow methods with applications to image registration*, in Proceedings of the 18th GAMM-Seminar Leipzig on Multigrid and Related Methods for Optimization Problems, 2002, pp. 11–30.
[23] D. Cremers, T. Kohlberger, and C. Schnörr, *Nonlinear shape statistics in Mumford-Shah based segmentation*, in Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Lecture Notes in Comput. Sci. 2351, A. Heyden et al., eds., Springer-Verlag, Berlin, 2002, pp. 93–108.
[24] D. Cremers, F. Tischhäuser, J. Weickert, and C. Schnörr, *Diffusion snakes: Introducing statistical shape knowledge into the Mumford-Shah functional*, Internat. J. Comput. Vision, 50 (2002), pp. 1364–1379.
[25] G. Dal Maso, J. M. Morel, and S. Solimini, *A variational method in image segmentation: Existence and approximation results*, Acta Math., 168 (1996), pp. 89–151.
[26] C. A. Davatzikos, R. N. Bryan, and J. L. Prince, *Image registration based on boundary mapping*, IEEE Trans. Medical Imaging, 15 (1996), pp. 112–115.
[27] E. De Giorgi, M. Carriero, and A. Leaci, *Existence theorem for a minimum problem with free discontinuity set*, Arch. Rational Mech. and Anal., 108 (1989), pp. 195–218.
[28] M. C. Delfour and J. P. Zolésio, *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*, Adv. Des. Control 4, SIAM, Philadelphia, 2001.
[29] M. Droske and M. Rumpf, *A variational approach to nonrigid morphological image registration*, SIAM J. Appl. Math., 64 (2004), pp. 668–687.

[30] S. Esedoglu and J. Shen, *Digital inpainting based on the Mumford-Shah-Euler image model*, European J. Appl. Math., 13 (2002), pp. 353–370.
[31] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.
[32] X. Feng and A. Prohl, *Analysis of gradient flow of a regularized Mumford-Shah functional for image segmentation and image inpainting*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 291–320.
[33] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, Berlin, Heidelberg, 1985.
[34] W. Hackbusch and S. Sauter, *Composite finite elements for the approximation of PDEs on domains with complicated micro-structures*, Numer. Math., 75 (1997), pp. 447–472.
[35] S. Henn, *A Levenberg-Marquardt scheme for nonlinear image registration*, BIT, 43 (2003), pp. 743–759.
[36] S. Henn and K. Witsch, *A multigrid approach for minimizing a nonlinear functional for digital image matching*, Computing, 64 (2000), pp. 339–348.
[37] M. Hintermüller and W. Ring, *An inexact Newton-CG-type active contour approach for the minimization of the Mumford-Shah functional*, J. Math. Imaging Vision, 20 (2004), pp. 19–42.
[38] B. K. P. Horn and B. G. Schunk, *Determining optical flow*, Artificial Intelligence, 17 (1981), pp. 185–204.
[39] S. C. Joshi and M. I. Miller, *Landmark matching via large deformation diffeomorphisms*, IEEE Trans. Medical Imaging, 9 (2000), pp. 1357–1370.
[40] T. Kapur, L. Yezzi, and L. Zöllei, *A variational framework for joint segmentation and registration*, in Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis, 2001, pp. 44–51.
[41] M. Kass, A. Witkin, and D. Terzopoulos, *Snakes: Active contour models*, Internat. J. Comput. Vision, 1 (1988), pp. 321–331.
[42] S. L. Keeling and W. Ring, *Medical image registration and interpolation by optical flow with maximal rigidity*, J. Math. Imaging Vision, 23 (2005), pp. 47–65.
[43] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, *Multi-modal volume registration by maximization of mutual information*, IEEE Trans. Medical Imaging, 16 (1997), pp. 187–198.
[44] R. Malladi, J. A. Sethian, and B. C. Vemuri, *Shape modeling with front propagation: A level set approach*, IEEE Trans. Pattern Anal. Mach. Intell., 17 (1995), pp. 158–175.
[45] M. I. Miller, A. Trouve, and L. Younes, *On the metrics and Euler-Lagrange equations of computational anatomy*, Ann. Rev. Biomed. Eng., 4 (2002), pp. 375–405.
[46] J. Modersitzki and B. Fischer, *Curvature based image registration*, J. Math. Imaging Vision, 18, 2003, pp. 81–85.
[47] P. Monasse, *Contrast invariant registration of images*, in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Vol. 6, Phoenix, Arizona, 1999, pp. 3221–3224.
[48] J. M. Morel and S. Solimini, *Variational Methods in Image Segmentation*, Birkäuser Boston, Boston, 1995.
[49] D. Mumford and J. Shah, *Optimal approximation by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math., 42 (1989), pp. 577–685.
[50] H. H. Nagel and W. Enkelmann, *An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences*, IEEE Trans. Pattern Anal. Mach. Intell., 8 (1986), pp. 565–593.
[51] S. J. Osher and R. P. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, Springer-Verlag, New York, 2002.
[52] S. J. Osher and N. Paragios, *Geometric Level Set Methods in Imaging, Vision, and Graphics*, Springer-Verlag, New York, 2003.
[53] S. J. Osher and J. A. Sethian, *Fronts propagating with curvature dependent speed: Algorithms based on Hamilton–Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.
[54] T. J. Richardson and S. K. Mitter, *A variational formulation-based edge focussing algorithm*, Sādhanā, 22 (1997), pp. 553–574.
[55] S. Sauter and N. Stahn, *Composite Finite Elements and Multi-grid. Part* I: *Convergence Theory in* 1-*d*, Technical report 11-01, University of Zürich, Zürich, 2001.
[56] J. A. Sethian, *Level Set Methods and Fast Marching Methods. Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, 2nd ed., Cambridge University Press, Cambridge, UK, 1999.
[57] J. Sokołowski and J-P. Zolésio, *Introduction to shape optimization*, Springer-Verlag, Berlin, 1992.

[58] A. Tsai, A. Yezzi, and A. Willsky, *Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpolation, and magnification*, IEEE Trans. Image Process., 10 (2001), pp. 1169–1186.

[59] G. Unal and G. Slabaugh, *Coupled PDEs for non-rigid registration and segmentation*, in Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 168–175.

[60] B. C. Vemuri, J. Ye, Y. Chen, and C. M. Leonard, *Image registration via level-set motion: Applications to atlas-based segmentation*, Med. Image Anal., 7 (2003), pp. 1–20.

[61] R. Warnke, *Schnelle Löser für Elliptische Randwertprobleme mit Springenden Koeffizienten*, Dissertation, Universtiy of Zürich, Zürich, 2003.

[62] W. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, *Multi-modal volume registration by maximization of mutual information*, Med. Image Anal., 1 (1996), pp. 35–51.

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 2149–2186

ON THE TEMPERATURE-JUMP PROBLEM IN RAREFIED GAS DYNAMICS: THE EFFECT OF THE CERCIGNANI–LAMPIS BOUNDARY CONDITION*

ROSENEI FELIPPE KNACKFUSS† AND LILIANE BASSO BARICHELLO‡

Abstract. An analytical version of the discrete-ordinates method (the ADO method) is used to establish a solution to the temperature-jump problem in the rarefied gas dynamics field. Kinetic models derived from the linearized Boltzmann equation are used to formulate the problem in the one gas case and for a binary gas mixture. The gas-surface interaction is described by the Cercignani–Lampis kernel, which is written in terms of two accommodation coefficients. The solution is found to be very accurate and fast. Numerical results are presented not only for the temperature-jump coefficient but also for the density and temperature profiles. In particular, the effect of both accommodation coefficients on the temperature-jump coefficient is analyzed.

Key words. rarefied gas dynamics, temperature-jump coefficient, binary gas mixture, Cercignani–Lampis kernel, discrete ordinates

AMS subject classifications. 76P05, 76M22, 65N35

DOI. 10.1137/050643209

1. Introduction.

The increased interest in research fields related to the development of technologies associated with micro-electro-mechanical systems (MEMS) [1, 2] as well as other micro- and nanoflow applications [3] has brought attention to the rarefied gas dynamics field [4, 5, 6]. In fact, for these microsystems, where the characteristic length of the system is of the order of a mean-free path, and the gas flow is in the transition regime, the Boltzmann equation [7] has to be used in order to describe correctly the state of the gas. In this case the Knudsen number (Kn), defined as the ratio of the molecular mean-free path to a characteristic size, is close to the unity.

For a gas in a moderate state of rarefaction (Kn < 0.1), however, in order to take into account the rarefaction effects by using simpler models, it is usual to use the continuum mechanics equations to define the problem of interest, along with the velocity-slip and temperature-jump boundary conditions [3, 4, 5, 6, 7]. In particular, for the case of evaluating the temperature distribution in a rarefied gas restricted by a solid surface, the temperature-jump boundary condition is used to take into account the noted difference (proportional to the gradient of the temperature in the normal direction to the wall) between the temperature of the wall and the temperature of the gas near to the wall. In this way, the temperature-jump boundary condition is defined in terms of the temperature-jump coefficient [7, 8, 9].

Although the temperature-jump coefficient may be evaluated by solving either the Boltzmann equation or model equations (derived from the Bolztmann equation with simplified collision operators) [4, 10] it is usual to find its definition, as derived by Maxwell, written in terms of the accommodation coefficient of the gas [3]. This

*Received by the editors October 20, 2005; accepted for publication (in revised form) July 5, 2006; published electronically October 24, 2006. This work was partially supported by CNPq of Brasil.
http://www.siam.org/journals/siap/66-6/64320.html

†Departamento de Matemática, Universidade Federal de Santa Maria, Av. Roraima, Prédio 13, 97105-900, Santa Maria, RS, Brazil (knackfuss@smail.ufsm.br).

‡Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves 9500, 91509-900, Porto Alegre, RS, Brazil (lbaric@mat.ufrgs.br).

approach is, according to the literature [11], just an estimation, since it is based on the assumption that the distribution function does not vary in the Knudsen layer. In fact, such an approach has been shown [11] to underestimate the value of the temperature-jump coefficient.

On the other hand, surface effects play a very important role in the analysis of the temperature in microscales, and so particular attention has to be given to the definition of the boundary conditions and to the choice of appropriate kernels to describe the gas-surface interaction [12, 13, 14].

In this work, we present a complete derivation of the solution of the temperature-jump problem for the one gas case (based on the S-model [15]) and a binary gas mixture (based on the McCormack model [16]). The solution is developed in terms of an analytical version of the discrete-ordinates method (the ADO method) [17]. In fact, the temperature-jump problem has been studied over the years, and previous results on this problem, based on numerical approaches, can be found in the literature [4, 6, 18, 19, 20, 21, 22, 23, 24]. The ADO method, which has been successfully applied to derive unified solutions for a wide class of problems in the rarefied gas dynamics field [25, 26, 27, 28], has been also applied to evaluate the temperature-jump coefficient [29, 30, 31, 32]. In particular, the ADO solution has been shown to be accurate and fast, in comparison with numerical based approaches.

Here, in addition to complete the class of problems solved by the ADO method, based on the kinetic S-model of the linearized Boltzmann equation [28], and to establish the relation between this formulation and its extension to the binary mixture case based on the McCormack model [32], we include, for the temperature-jump problem, a special treatment for the gas-surface interaction: the Cercignani–Lampis kernel [12]. In fact, considering the significance of the surface effect analysis on microflow applications and the good results obtained by the application of the ADO method to the solution of problems in this field, this kernel has been included in the treatment of channel problems by the ADO method [33, 34]. Differently from the commonly used Maxwell boundary condition [10], the Cercignani–Lampis kernel is defined in terms of two accommodation coefficients such that a better physical representation of the surface effects is allowed. In this work, a complete study of the dependence of the temperature-jump coefficient on different kinetic models as well as on the gas-surface interaction kernel is then carried out.

In this way, we develop in sections 2 and 3 the basic formulation for the one gas case, and we describe its ADO solution in section 4. We detail the McCormack model for a binary gas mixture in sections 5 and 6. In section 7 we present the discrete-ordinates solution for the mixture. We discuss computational aspects and numerical results in section 8 before presenting some concluding comments in section 9.

2. A model equation: The one gas case. To start this work, we follow Williams [10] and consider the steady-state nonlinear Boltzmann equation written, in a general form, as

$$\mathbf{v} \cdot \nabla_r f(\mathbf{r}, \mathbf{v}) = J(f', f), \tag{1}$$

where $f(\mathbf{r}, \mathbf{v})$ is the gas atom space and velocity distribution function (f and f' are associated with, respectively, before and after collision distributions) and J is the collision operator [10]. For the cases weakly far from the equilibrium, it is customary to write f as

$$f(\mathbf{r}, \mathbf{v}) = f_0(\mathbf{v})[1 + h(\mathbf{r}, \mathbf{v})], \tag{2}$$

where h is a perturbation caused, by the presence of the walls, to the absolute Maxwellian $f_0(\mathbf{v})$,

$$f_0(\mathbf{v}) = n_0(\lambda_0/\pi)^{3/2}\mathrm{e}^{-\lambda_0 v^2}, \quad \lambda_0 = m_0/(2kT_0). \tag{3}$$

Here k is the Boltzmann constant, T_0 is a reference temperature, m_0 is the mass, and n_0 is the equilibrium density of the gas. In this way, if we substitute (2) into (1) and use, along with properties of the collision operator [9], some physical considerations, we obtain, for the dimensionless velocity variable

$$\mathbf{c} = \mathbf{v}[(m/2kT_0)]^{1/2}, \tag{4}$$

the linearized Boltzmann equation written in terms of the perturbation function h as [10, 35]

$$c_y\frac{\partial}{\partial y}h(y,\mathbf{c}) + \varepsilon h(y,\mathbf{c}) = \varepsilon\pi^{-3/2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{e}^{-c'^2}h(y,\mathbf{c}')F(\mathbf{c}':\mathbf{c})\mathrm{d}c'_x\mathrm{d}c'_y\mathrm{d}c'_z. \tag{5}$$

Here, in addition to the three components of the velocity vector (c_x, c_y, c_z) which are expressed in dimensionless units, we consider the dimensionless (written in terms of a mean-free path l) spatial variable $y > 0$, and

$$\varepsilon = \sigma_0^2 n_0 \pi^{1/2} l, \tag{6}$$

where σ_0 is the collision diameter of the gas particles (in the rigid-sphere approximation).

For rigid spheres, the scattering kernel $F(\mathbf{c}':\mathbf{c})$ can be expanded in terms of Legendre functions [36]. However, even if we consider a truncated form of this expansion, the problem of solving the resulting approximation of the linearized Boltzmann equation is still difficult from a numerical point of view [35]. For this reason, keeping in mind mathematical properties [7], one seeks to approximate the true kernel by physically meaningful approximations that can be more easily handled by analytical tools and numerical algorithms. In this way, the resulting equations are known as "model (kinetic) equations." Here we follow Siewert [33] and express the kernel, in (5), such that two of the well-known constant collision frequency models of the rarefied gas dynamics are represented:

$$F(\mathbf{c}':\mathbf{c}) = 1 + 2(\mathbf{c}'.\mathbf{c}) + (2/3)(c'^2 - 3/2)(c^2 - 3/2) + \widehat{\beta}M(\mathbf{c}':\mathbf{c}) \tag{7a}$$

with

$$M(\mathbf{c}':\mathbf{c}) = (4/15)(\mathbf{c}'.\mathbf{c})(c'^2 - 5/2)(c^2 - 5/2), \tag{7b}$$

where the case $\widehat{\beta} = 0$ defines the well-known BGK model equation [37] and the case $\widehat{\beta} = 1$ defines the S-model equation [15]. We note that, for both constant collision frequency models we use in this work, if we choose a mean-free path based on the viscosity to evaluate (6), we obtain [28, 35]

$$\varepsilon = \varepsilon_p = 1. \tag{8}$$

On the other hand, if we use a mean-free path based on thermal conductivity to evaluate (6), we get, respectively, for the BGK and the S-model

$$\varepsilon = \varepsilon_t = 1 \quad \text{and} \quad \varepsilon = \varepsilon_t = 3/2. \tag{9}$$

Taking into account the values listed in (8) and (9) we see that the evaluation of $\varepsilon_p/\varepsilon_t$ for the case of the S-model leads to a correct value for the Prandtl number (equal to 2/3). This aspect is considered an advantage of that model in comparison with the BGK model.

To complete the formulation of the temperature-jump problem, we write the boundary condition in terms of the Cercignani–Lampis kernel [12, 33],

$$h(0, c_x, c_y, c_z) = \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty h(0, c'_x, -c'_y, c'_z) R(c'_x, -c'_y, c'_z : c_x, c_y, c_z) \mathrm{d}c'_x \mathrm{d}c'_z \mathrm{d}c'_y, \tag{10a}$$

where

$$R(c'_x, c'_y, c'_z : c_x, c_y, c_z) = \frac{2c'_y}{\pi \alpha_n \alpha_t (2 - \alpha_t)} T(c'_x : c_x) S(c'_y : c_y) T(c'_z : c_z) \tag{10b}$$

with

$$T(x : z) = \exp\left[-\frac{[(1 - \alpha_t)z - x]^2}{\alpha_t (2 - \alpha_t)} \right] \tag{10c}$$

and

$$S(x : z) = \exp\left[-\frac{[(1 - \alpha_n)^{1/2} z - x]^2}{\alpha_n} \right] \widehat{I}_0 \left[\frac{2(1 - \alpha_n)^{1/2} |xz|}{\alpha_n} \right]. \tag{10d}$$

For computational purposes, we write

$$\widehat{I}_0(w) = I_0(w) \mathrm{e}^{-w}, \tag{11a}$$

where $I_0(w)$ is the modified Bessel function,

$$I_0(w) = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{e}^{w \cos \phi} \mathrm{d}\phi. \tag{11b}$$

Differently from the usual case of Maxwell boundary condition [10] defined in terms of one accommodation coefficient α, where it is assumed that some fraction α of the particles are reflected diffusely and the rest $(1 - \alpha)$ is reflected specularly, we can see in (10a) to (10d) that the Cercignani–Lampis kernel is defined in terms of two accommodation coefficients: $\alpha_t \in [0, 2)$ the accommodation coefficient of tangential momentum and $\alpha_n \in [0, 1)$ the accommodation coefficient of energy corresponding to the normal component of velocity. The use of more than one accommodation coefficient allows a better physical representation of the gas-surface interaction. In the case of the Cercignani–Lampis kernel, that is the case, according to the literature, mainly for roughness surfaces [24].

Still, to complete the definition of the temperature-jump problem, we have to specify the behavior of the solution at infinity. We impose the Welander condition [8]

$$\lim_{y \to \infty} \frac{\mathrm{d}}{\mathrm{d}y} T(y) = K, \tag{12}$$

where K is known and here the temperature perturbation is given, in terms of h as

$$T(y) = \frac{2}{3}\pi^{-3/2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-c'^2} h(y, c_x, c_y, c_z)\left(c^2 - \frac{3}{2}\right)\mathrm{d}c_x\mathrm{d}c_y\mathrm{d}c_z. \tag{13}$$

In solving the temperature-jump problem, in addition to the temperature perturbation already defined, another quantity we seek to compute is the density perturbation

$$N(y) = \pi^{-3/2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-c'^2} h(y, c_x, c_y, c_z)\mathrm{d}c_x\mathrm{d}c_y\mathrm{d}c_z. \tag{14}$$

By looking at the above definitions, (13) and (14), we see that we do not need to obtain the complete distribution h, but only some related integrals (moments), to compute those quantities. In this way, in what follows, we develop a procedure in order to get simpler problems, in terms of those moments, for which we will develop a solution with the ADO method.

3. A vector formulation–the S-model. In order to develop simpler formulations to evaluate the quantities of interest, we multiply (5) first by

$$\phi_1(c_x, c_z) = \frac{1}{\pi}\mathrm{e}^{-(c_x^2+c_z^2)} \tag{15}$$

and, in a second step by

$$\phi_2(c_x, c_z) = \frac{1}{\pi}(c_x^2 + c_z^2 - 2)\mathrm{e}^{-(c_x^2+c_z^2)}, \tag{16}$$

and, in both cases, we integrate over all c_x and c_z, such that, if we define

$$h_1(y, \xi) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi_1(c_x, c_z)h(y, c_x, c_y, c_z)\mathrm{d}c_x\mathrm{d}c_z, \tag{17}$$

$$h_2(y, \xi) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi_2(c_x, c_z)h(y, c_x, c_y, c_z)\mathrm{d}c_x\mathrm{d}c_z \tag{18}$$

and introduce the new notation $\xi = c_y$, we find that

$$\xi\frac{\partial}{\partial y}h_1(y, \xi) + \varepsilon h_1(y, \xi) = \varepsilon\int_{-\infty}^{\infty}\psi(\xi')\Big[\kappa_{11}(\xi', \xi)h_1(y, \xi') + \kappa_{12}(\xi', \xi)h_2(y, \xi')\Big]\mathrm{d}\xi' \tag{19}$$

and

$$\xi\frac{\partial}{\partial y}h_2(y, \xi) + \varepsilon h_2(y, \xi) = \varepsilon\int_{-\infty}^{\infty}\psi(\xi')\Big[\kappa_{21}(\xi', \xi)h_1(y, \xi') + \kappa_{22}(\xi', \xi)h_2(y, \xi')\Big]\mathrm{d}\xi', \tag{20}$$

with

$$\psi(\xi) = \pi^{-1/2}\mathrm{e}^{-\xi^2} \tag{21}$$

and

(22)
$$\kappa_{11}(\xi',\xi) = 1 + \frac{2}{3}\left(\xi^2 - \frac{1}{2}\right)\left(\xi'^2 - \frac{1}{2}\right) + \widehat{\beta}\left[2\xi'\xi + \frac{4}{15}\left(\xi'^2 - \frac{3}{2}\right)\left(\xi^2 - \frac{3}{2}\right)\xi'\xi\right],$$

(23)
$$\kappa_{12}(\xi',\xi) = \frac{2}{3}\left(\xi^2 - \frac{1}{2}\right) + \widehat{\beta}\left[\left(\frac{4}{15}\xi^2 - \frac{2}{5}\right)\xi\xi'\right],$$

(24)
$$\kappa_{21}(\xi',\xi) = \frac{2}{3}\left(\xi'^2 - \frac{1}{2}\right) + \widehat{\beta}\left[\left(\frac{4}{15}\xi'^2 - \frac{2}{5}\right)\xi\xi'\right],$$

(25)
$$\kappa_{22}(\xi',\xi) = \frac{2}{3} + \widehat{\beta}\frac{4}{15}\xi\xi'.$$

The same procedure, initiated in (15) and (16), is applied to the boundary condition, (10a), to define the components h_1 and h_2 at the boundary, which are written in the form

(26)
$$h_1(0,\xi) = \int_0^\infty h_1(0,-\xi')f(\xi',\xi)\mathrm{d}\xi'$$

and

(27)
$$h_2(0,\xi) = (1-\alpha_t)^2\int_0^\infty h_2(0,-\xi')f(\xi',\xi)\mathrm{d}\xi',$$

for ξ, $\xi'\in(0,\infty)$ and

(28)
$$f(\xi',\xi) = \frac{2\xi'}{\alpha_n}\exp\left[-\frac{[(1-\alpha_n)^{1/2}\xi - \xi']^2}{\alpha_n}\right]\widehat{I}_0\left[\frac{2(1-\alpha_n)^{1/2}\xi'\xi}{\alpha_n}\right].$$

In this way, we let $\mathbf{H}(y,\xi)$ be the vector with components $h_1(y,\xi)$ and $h_2(y,\xi)$ and rewrite (19) to (28) in a more appropriate matrix form as

(29)
$$\xi\frac{\partial}{\partial y}\mathbf{H}(y,\xi) + \varepsilon\mathbf{H}(y,\xi) = \varepsilon\int_{-\infty}^{\infty}\psi(\xi')\boldsymbol{K}(\xi',\xi)\mathbf{H}(y,\xi')\mathrm{d}\xi',$$

for $y>0$ and $\xi\in(-\infty,\infty)$. Here we note (21) to (25) and the definitions of

(30)
$$\mathbf{H}(y,\xi) = \begin{bmatrix} h_1(y,\xi) \\ h_2(y,\xi) \end{bmatrix}$$

and

(31)
$$\boldsymbol{K}(\xi',\xi) = \begin{bmatrix} \kappa_{11}(\xi',\xi) & \kappa_{12}(\xi',\xi) \\ \kappa_{21}(\xi',\xi) & \kappa_{22}(\xi',\xi \end{bmatrix}.$$

In regard to the boundary condition, we write

(32)
$$\mathbf{H}(0,\xi) = \mathbf{A}\int_0^\infty \mathbf{H}(0,-\xi')f(\xi',\xi)\mathrm{d}\xi',$$

with

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & (1-\alpha_t)^2 \end{bmatrix}, \tag{33}$$

for ξ, $\xi' \in (0,\infty)$.

Finally, we also use the vector notation to express, respectively, the temperature and density perturbations, given in (13) and (14), in the form

$$T(y) = \frac{2}{3}\int_{-\infty}^{\infty} \begin{bmatrix} \xi^2 - 1/2 \\ 1 \end{bmatrix}^T \mathbf{H}(y,\xi)\psi(\xi)\mathrm{d}\xi \tag{34}$$

and

$$N(y) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}^T \int_{-\infty}^{\infty} \mathbf{H}(y,\xi)\psi(\xi)\mathrm{d}\xi. \tag{35}$$

We now proceed to develop a discrete-ordinates solution for the vector problem. In what follows we chose to assume the S-model case ($\widehat{\beta} = 1$), to which the expression for the kernel and the development of the solution are more general. Much simpler expressions can be found for the BGK model, and the work of Barichello and Siewert [29] can be used to follow that derivation.

4. A discrete-ordinates solution (I). We now seek for a discrete-ordinates solution of the problem defined by (29) and (32) and the condition given by (12). In this sense, the analytical discrete-ordinates method we use in this work, the ADO method [17], is based on a half-range quadrature scheme, and so as a first step we write (29) in the form

$$\xi\frac{\partial}{\partial y}\mathbf{H}(y,\xi) + \varepsilon\mathbf{H}(y,\xi) = \varepsilon\int_0^{\infty} \psi(\xi')[\boldsymbol{K}(\xi',\xi)\mathbf{H}(y,\xi') + \boldsymbol{K}(-\xi',\xi)\mathbf{H}(y,-\xi')]\mathrm{d}\xi', \tag{36}$$

and we seek solutions of (36) of the form

$$\mathbf{H}(y,\xi) = \mathbf{\Phi}(\nu,\xi)\ \mathrm{e}^{-y\varepsilon/\nu}, \tag{37}$$

where the separation constants ν and the elementary (2×1 vector) solutions $\mathbf{\Phi}$ are to be determined. Substituting (37) into (36) we find that

$$\varepsilon(\nu-\xi)\mathbf{\Phi}(\nu,\xi) = \varepsilon\nu\int_0^{\infty} \psi(\xi')[\boldsymbol{K}(\xi',\xi)\mathbf{\Phi}(\nu,\xi') + \boldsymbol{K}(-\xi',\xi)\mathbf{\Phi}(\nu,-\xi')]\mathrm{d}\xi' \tag{38}$$

and

$$\varepsilon(\nu+\xi)\mathbf{\Phi}(\nu,-\xi) = \varepsilon\nu\int_0^{\infty} \psi(\xi')[\boldsymbol{K}(\xi',-\xi)\mathbf{\Phi}(\nu,\xi') + \boldsymbol{K}(-\xi',-\xi)\mathbf{\Phi}(\nu,-\xi')]\mathrm{d}\xi'. \tag{39}$$

Here we note, since

$$\boldsymbol{K}(\xi',-\xi) = \boldsymbol{K}(-\xi',\xi), \tag{40}$$

that

$$\mathbf{\Phi}(\nu,\xi) = \mathbf{\Phi}(-\nu,-\xi). \tag{41}$$

Continuing our development, we add and subtract (38) and (39), one from the other, to find that

$$\frac{1}{\xi^2}\left[\mathbf{V}(\nu,\xi) - \int_0^\infty \psi(\xi')\mathbf{P}(\xi',\xi)\mathbf{V}(\nu,\xi')\mathrm{d}\xi'\right] = \lambda\mathbf{V}(\nu,\xi) \tag{42}$$

and

$$\mathbf{U}(\nu,\xi) = \frac{\nu}{\xi}\left[\mathbf{V}(\nu,\xi) - \int_0^\infty \psi(\xi')[\boldsymbol{K}(\xi',\xi)\mathbf{V}(\nu,\xi') - \boldsymbol{K}(-\xi',\xi)\mathbf{V}(\nu,\xi')]\mathrm{d}\xi'\right], \tag{43}$$

where

$$\mathbf{U}(\nu,\xi) = \mathbf{\Phi}(\nu,\xi) + \mathbf{\Phi}(\nu,-\xi), \tag{44}$$

$$\mathbf{V}(\nu,\xi) = \mathbf{\Phi}(\nu,\xi) - \mathbf{\Phi}(\nu,-\xi), \tag{45}$$

and

$$\lambda = 1/\nu^2. \tag{46}$$

Still, we note that the separation constants ν occur in $(\pm)$ pairs, and

$$\begin{aligned}\mathbf{P}(\xi',\xi) = \frac{\xi}{\xi'}\Big[&[\boldsymbol{K}(\xi',\xi) + \boldsymbol{K}(-\xi',\xi)] + [\boldsymbol{K}(\xi',\xi) - \boldsymbol{K}(-\xi',\xi)]\Big] \\ -\int_0^\infty \psi(\xi'')\frac{\xi}{\xi''}&[\boldsymbol{K}(\xi',\xi'') - \boldsymbol{K}(-\xi',\xi'')][\boldsymbol{K}(\xi'',\xi) + \boldsymbol{K}(-\xi'',\xi)]\mathrm{d}\xi''.\end{aligned} \tag{47}$$

At this point, we introduce a half-range quadrature scheme, in $[0,\infty)$, defined by N nodes $\{\xi_k\}$ and weights $\{\omega_k\}$, and rewrite (42) and (43) evaluated at the quadrature points as

$$\frac{1}{\xi_i^2}\left[\mathbf{V}(\nu_j,\xi_i) - \sum_{k=1}^N \omega_k\psi(\xi_k)\mathbf{P}(\xi_k,\xi_i)\mathbf{V}(\nu_j,\xi_k)\right] = \lambda_j\mathbf{V}(\nu_j,\xi_i) \tag{48}$$

and

$$\mathbf{U}(\nu_j,\xi_i) = \frac{\nu_j}{\xi_i}\left[\mathbf{V}(\nu_j,\xi_i) - \sum_{k=1}^N \omega_k\psi(\xi_k)[\mathbf{K}(\xi_k,\xi_i)\mathbf{V}(\nu_j,\xi_k) - \mathbf{K}(-\xi_k,\xi_i)\mathbf{V}(\nu_j,\xi_k)]\right] \tag{49}$$

for $i = 1, 2, \ldots, N$. Once we solve our eigenvalue problem, defined by (48), we have the elementary solutions from

$$\mathbf{\Phi}(\nu_j,\xi_i) = \frac{1}{2}\Big[\mathbf{U}(\nu_j,\xi_i) + \mathbf{V}(\nu_j,\xi_i)\Big] \tag{50}$$

and

$$\mathbf{\Phi}(\nu_j,-\xi_i) = \frac{1}{2}\Big[\mathbf{U}(\nu_j,\xi_i) - \mathbf{V}(\nu_j,\xi_i)\Big]. \tag{51}$$

In (48) to (51) we use the subscript j to label the eigenvalues. For each j, (50) and (51) express $2N \times 1$ vectors.

Here it may be important to make a comment in regard to the eigenvalue problem defined in (42). In applying the ADO method we have found, for different model equations and for different classes of problems, specific eigenvalue problems. In most of those cases the problem is much simpler and expressed in a much more concise form than in the case of this work. Even the case of diagonal + rank one type matrix can be mentioned [17, 38]. In fact, for the specific derivation presented above, some properties of the kernel can also be used to derive a simpler eigenvalue problem, as shown by Scherer [39]. We chose, however, to keep the derivation as it is, in order to establish more close connections and analogies with the mixtures case we will present later on in this text.

Continuing, we consider (50) and (51) along with (46), we take the positive root ν_j, and we write the general discrete-ordinates solution to (29) as

$$\mathbf{H}(y, \pm\xi_i) = \sum_{j=1}^{2N} [A_j \mathbf{\Phi}(\nu_j, \pm\xi_i) \mathrm{e}^{-y\varepsilon/\nu_j} + B_j \mathbf{\Phi}(\nu_j, \mp\xi_i) \mathrm{e}^{y\varepsilon/\nu_j}], \tag{52}$$

for $i = 1, 2, \ldots, N$. Here the arbitrary constants are to be determined from the boundary condition and the imposed condition at infinity. Before doing that, however, we note that the eigenvalue problem yields two separation constants, say ν_1 and ν_2, that become unbounded. Because of this, we rewrite the general solution as

$$\mathbf{H}(y, \pm\xi_i) = \mathbf{H}^*(y, \pm\xi_i) + \sum_{j=3}^{2N} [A_j \mathbf{\Phi}(\nu_j, \pm\xi_i) \mathrm{e}^{-y\varepsilon/\nu_j} + B_j \mathbf{\Phi}(\nu_j, \mp\xi_i) \mathrm{e}^{y\varepsilon/\nu_j}], \tag{53}$$

$i = 1, 2, \ldots, N$, where we then introduced, in (53), four linear independent exact solutions of (29). In other words,

$$\mathbf{H}^*(y, \xi) = A_1 \mathbf{H}_1 + A_2 \mathbf{H}_2(\xi) + B_1 \mathbf{H}_3(\xi) + B_2 \mathbf{H}_4(\xi), \tag{54}$$

where, for the S-model,

$$\mathbf{H}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{H}_2(\xi) = \begin{bmatrix} \xi^2 - 1/2 \\ 1 \end{bmatrix}, \quad \mathbf{H}_3(\xi) = \begin{bmatrix} \xi \\ 0 \end{bmatrix}, \tag{55}$$

and

$$\mathbf{H}_4(y, \xi) = y\mathbf{G}_1(\xi) + \mathbf{F}_1(\xi) \tag{56a}$$

with

$$\mathbf{G}_1(\xi) = \begin{bmatrix} \xi^2 - 3/2 \\ 1 \end{bmatrix} \tag{56b}$$

and

$$\mathbf{F}_1(\xi) = \frac{-3\xi}{2\varepsilon} \mathbf{G}_1(\xi). \tag{56c}$$

In fact, in the process of finding the $\mathbf{H}_4$ solution, once that (56b) is an exact solution of (29), we can follow analogous procedure as the one proposed by Siewert [32], and we substitute (56a) into (29) to find that

$$\mathbf{F}_1(\xi) = -\frac{1}{\varepsilon}\xi\mathbf{G}_1(\xi) + \int_{-\infty}^{\infty} \psi(\xi')\boldsymbol{K}(\xi',\xi)\mathbf{F}_1(\xi')\mathrm{d}\xi'. \tag{57}$$

The form of the inhomogeneous term, in (57), as well as the the form of $\boldsymbol{K}(\xi',\xi)$ suggest that we look for solutions of the form

$$\mathbf{F}_1(\xi) = \sum_{\alpha=0}^{3} P_\alpha(\xi)\mathbf{F}_{1,\alpha}, \tag{58}$$

where $\mathbf{F}_{1,\alpha}$ are constant vectors and $P_\alpha(\xi)$ are orthogonal polynomials given by

$$P_0(\xi) = 1\ ,\quad P_1(\xi) = \xi\ ,\quad P_2(\xi) = \xi^2 - 1/2, \quad \text{and} \quad P_3(\xi) = \xi(\xi^2 - 3/2). \tag{59}$$

Thus, if we first substitute (58) into (57), and then we multiply the equation by $\psi(\xi)P_k(\xi)$, for $k = 0, 1, 2, 3$, to finally integrate over all ξ, we obtain an algebraic system (with 8 equations and rank 6)

$$\sum_{\alpha=0}^{3} P_\alpha(\xi)\mathbf{F}_{1,\alpha} = -\frac{1}{\varepsilon}\xi\mathbf{G}_1(\xi) + \int_{-\infty}^{\infty} \psi(\xi')\boldsymbol{K}(\xi',\xi)\sum_{\alpha=0}^{3} P_\alpha(\xi')\mathbf{F}_{1,\alpha}\mathrm{d}\xi', \tag{60}$$

for which a solution can be found explicitly, as the one given in (56a).

So, at this point we can go back to the issue of determining, in (53), the $4N$ arbitrary constants, such that our discrete-ordinates solution will be completely established. To start, if we consider the way that the solution is required to diverge at infinity, we take $B_j = 0$ for $j = 3, 4, \ldots, 2N$. In addition, when considering the Welander condition, given by (12) (and noting (34)), we find that

$$B_2 = K. \tag{61}$$

Still, we note that the solution $\mathbf{H}_1$, (55), satisfies the homogeneous boundary condition given by (32), and so the A_1 coefficient cannot be determined from that equation. However, following previous works [29, 40] we can impose on our solution an additional normalization condition

$$\lim_{y\to\infty}\left[N(y) + T(y)\right] = 0 \tag{62}$$

from where we get that

$$A_2 = -A_1. \tag{63}$$

In this way, we rewrite the general solution, (53), for $i = 1, 2, \ldots, N$, as

$$\mathbf{H}(y,\pm\xi_i) = A_2\mathbf{G}_1(\pm\xi_i) + B_1\mathbf{H}_3(\pm\xi_i) + K\mathbf{H}_4(y,\pm\xi_i) + \sum_{j=3}^{2N} A_j\boldsymbol{\Phi}(\nu_j,\pm\xi_i)\mathrm{e}^{-y\varepsilon/\nu_j}. \tag{64}$$

To determine the remaining $2N$ arbitrary constants, we substitute (64) into the discrete-ordinates version of the boundary condition given by (32)

$$\mathbf{H}(0,\xi_i) = \mathbf{A}\sum_{k=1}^{N}\omega_k \mathrm{e}^{-\xi_k{}^2}\mathbf{H}(0,-\xi_k)f(\xi_k,\xi_i), \tag{65}$$

for $i = 1, 2, \ldots, N$. We then obtain the following $2N \times 2N$ linear algebraic system:

$$\begin{aligned} A_2\mathbf{C^1}(\xi_i) + B_1\mathbf{C^2}(\xi_i) + \sum_{j=3}^{2N} A_j\left\{\mathbf{\Phi}_+(\nu_j) - \sum_{k=1}^{N}\omega_k f(\xi_k,\xi_i)\mathbf{A}^*\mathbf{\Phi}_-(\nu_j)\right\} \\ = K\left\{\sum_{k=1}^{N}\omega_k f(\xi_k,\xi_i)\mathbf{A}^*\mathbf{B}^*(-\xi_k) - \mathbf{B}^*(\xi_i)\right\}, \end{aligned} \tag{66}$$

where the components of the vectors $\mathbf{\Phi}_+(\nu_j)$ and $\mathbf{\Phi}_-(\nu_j)$ are given by (50) and (51),

$$\mathbf{C^1}(\xi_i) = \sum_{k=1}^{N}\omega_k f(\xi_k,\xi_i)\mathbf{A}^*\mathbf{R}^*(-\xi_k) - \mathbf{R}^*(\xi_i), \tag{67}$$

and

$$\mathbf{C^2}(\xi_i) = \mathbf{H}_3^*(\xi_i) - \sum_{k=1}^{N}\omega_k f(\xi_k,\xi_i)\mathbf{A}^*\mathbf{H}_3^*(-\xi_k). \tag{68}$$

Here $\mathbf{R}^*(\xi)$ is a $2N \times 1$ vector where each 2×1 component is

$$\mathbf{R}(\xi) = -\mathbf{G}_1(\xi) \tag{69}$$

for $\mathbf{G}_1(\xi)$ defined in (56b); $\mathbf{A}^*$ is a $2N \times 2N$ block diagonal matrix

$$\mathbf{A}^* = \mathrm{diag}\Big\{\mathbf{A}, \mathbf{A}, \ldots, \mathbf{A}, \mathbf{A}\Big\} \tag{70}$$

with $\mathbf{A}$ given in (33); $\mathbf{H}_3^*(\xi)$ and $\mathbf{B}^*(\xi)$ are $2N \times 1$ vectors where each 2×1 component is, respectively, defined in (55) and (56c). In addition, $f(\xi_k, \xi_i)$ is defined in (28).

Considering now the quantities we want to evaluate, we substitute (53) into the discrete-ordinates version of (34) and (35) to write the temperature perturbation

$$T(y) = Ky + A_2 + 2/3\sum_{j=3}^{2N} A_j\ (\mathrm{e}^{-y\varepsilon/\nu_j})M_1(\nu_j) \tag{71}$$

and the density perturbation

$$N(y) = -Ky - A_2 + \sum_{j=3}^{2N} A_j\ (\mathrm{e}^{-y\varepsilon/\nu_j})M_2(\nu_j), \tag{72}$$

where M_1 and M_2 are, respectively, the components of the vector given by

$$\mathbf{M}^*(\nu_j) = \pi^{-1/2}\sum_{k=1}^{N}\omega_k e^{-\xi_k{}^2}\begin{bmatrix} \xi_k^2 - 1/2 & 1 \\ 1 & 0 \end{bmatrix}[\mathbf{\Phi}(\nu_j,\xi_k) + \mathbf{\Phi}(\nu_j,-\xi_k)]. \tag{73}$$

Still, if we look now at the linear component of the temperature perturbation expression

$$T^*(y) = Ky + A_2 \tag{74}$$

we can define another quantity of interest, which relates the temperature perturbation at the wall with the gradient of the temperature, the temperature-jump coefficient ζ, as

$$T^*(0) = \zeta \frac{\mathrm{d}}{\mathrm{d}y} T^*(y)\Big|_{y=0} \tag{75}$$

such that

$$\zeta = A_2/K. \tag{76}$$

To be clear, we note that the normalization constant introduced in (62) does not affect the temperature-jump coefficient or the temperature perturbation. In fact, another choice of that constant would change only the density perturbation by the addition of a constant factor [30].

5. The McCormack model and a mixture of gases. In regard to a binary gas mixture, in this work we base our discussion on the McCormack model, as introduced by McCormack [16], following an appropriate notation proposed by Siewert [32]. In this way, to derive the balance equation, we consider the functions $h_\alpha(x, \mathbf{v})$ for the two types of particles ($\alpha = 1$ and 2) which denote perturbations from Maxwellian distributions for each species,

$$f_\alpha(y^*, \mathbf{v}) = f_{\alpha,0}(\mathbf{v})[1 + h_\alpha(y^*, \mathbf{v})], \tag{77}$$

where

$$f_{\alpha,0}(\mathbf{v}) = n_\alpha(\lambda_\alpha/\pi)^{3/2}\mathrm{e}^{-\lambda_\alpha v^2}, \quad \lambda_\alpha = m_\alpha/(2kT_0). \tag{78}$$

Here, again, k is the Boltzmann constant, m_α is the mass, n_α is the equilibrium density of the αth species, T_0 is a reference temperature, y^* is the spatial variable, and the vector $\mathbf{v}$, with magnitude v, is the particle velocity with components v_x, v_y, and v_z. It is found [16] that the balance equation is of the form

$$c_y \frac{\partial}{\partial y^*} h_\alpha(y^*, \mathbf{c}) + w_\alpha \gamma_\alpha h_\alpha(y^*, \mathbf{c}) = w_\alpha \gamma_\alpha \mathcal{L}_\alpha\{h_1, h_2\}(y^*, \mathbf{c}), \quad \alpha = 1, 2, \tag{79}$$

where $\mathbf{c}$, with magnitude c and components c_x, c_y, and c_z, is now a dimensionless velocity variable. In obtaining (79), we followed Siewert [32] and expressed the dimensionless variable $\mathbf{c}$ differently in the two equations: for the case $\alpha = 1$ we defined $\mathbf{c} = w_1\mathbf{v}$, and, analogously, for the case $\alpha = 2$ we defined $\mathbf{c} = w_2\mathbf{v}$, where

$$w_\alpha = [m_\alpha/(2kT_0)]^{1/2}, \quad \alpha = 1, 2. \tag{80}$$

Still, in (79), the γ_α are to be defined later, and the collision operator is written as

$$\mathcal{L}_\alpha\{h_1, h_2\}(y^*, \mathbf{c}) = \frac{1}{\pi^{3/2}} \sum_{\beta=1}^{2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-c'^2} h_\beta(y^*, \mathbf{c}') \boldsymbol{K}_{\beta,\alpha}(\mathbf{c}', \mathbf{c}) \mathrm{d}c'_x \mathrm{d}c'_y \mathrm{d}c'_z, \tag{81}$$

with the kernels expressed, in a general form, as [32]

(82)
$$\boldsymbol{K}_{\beta,\alpha} = \boldsymbol{K}^{(1)}_{\beta,\alpha}(\mathbf{c}',\mathbf{c}) + \boldsymbol{K}^{(2)}_{\beta,\alpha}(\mathbf{c}',\mathbf{c}) + \boldsymbol{K}^{(3)}_{\beta,\alpha}(\mathbf{c}',\mathbf{c}) + \boldsymbol{K}^{(4)}_{\beta,\alpha}(\mathbf{c}',\mathbf{c}), \quad \alpha,\beta = 1,2.$$

To avoid a heavy notation in the middle of the text, explicit expressions for each of the components in (82) are listed in Appendix A.

As in the one gas case, in (5), we introduce a dimensionless spatial variable $y = y^*/l_0$, where now l_0 is the mean-free path suggested by Sharipov and Kalempa [41], based on viscosity, defined as

(83)
$$l_0 = \frac{\mu v_0}{P_0},$$

for

(84)
$$v_0 = \left(\frac{2kT_0}{m}\right)^{1/2}$$

and

(85)
$$m = \frac{n_1 m_1 + n_2 m_2}{n_1 + n_2}.$$

In regard to (81), we continue to follow Sharipov and Kalempa [41] and Siewert [32] and express the viscosity of the mixture in terms of the pressures P_α and the collision frequencies γ_α as

(86)
$$\mu = \frac{P_1}{\gamma_1} + \frac{P_2}{\gamma_2},$$

where

(87)
$$\frac{P_\alpha}{P_0} = \frac{n_\alpha}{n_1 + n_2},$$

(88a)
$$\gamma_1 = \frac{\Psi_1\Psi_2 - \nu^{(4)}_{1,2}\nu^{(4)}_{2,1}}{\Psi_2 + \nu^{(4)}_{1,2}},$$

and

(88b)
$$\gamma_2 = \frac{\Psi_1\Psi_2 - \nu^{(4)}_{1,2}\nu^{(4)}_{2,1}}{\Psi_1 + \nu^{(4)}_{2,1}}.$$

In addition,

(89a)
$$\Psi_1 = \nu^{(3)}_{1,1} - \nu^{(4)}_{1,1} + \nu^{(3)}_{1,2},$$

(89b)
$$\Psi_2 = \nu^{(3)}_{2,2} - \nu^{(4)}_{2,2} + \nu^{(3)}_{2,1}.$$

Again, we used above some definitions listed in Appendix A.

At this point, we introduce [32]

$$\sigma_\alpha = \gamma_\alpha \omega_\alpha l_0 \tag{90}$$

or, more explicitly,

$$\sigma_\alpha = \gamma_\alpha \frac{n_1/\gamma_1 + n_2/\gamma_2}{n_1 + n_2} \left(\frac{m_\alpha}{m}\right)^{1/2}, \tag{91}$$

such that, in the next section, we rewrite (79), in order to specifically define the temperature-jump problem based on the McCormack model, that we want to solve in this work, including the Cercignani–Lampis kernel.

6. A binary gas mixture. Taking into account the development presented in the previous section, our starting point is now, analogously to the one gas case, (5), the kinetic equation for the McCormack model written as

$$c_y \frac{\partial}{\partial y} h_\alpha(y, \mathbf{c}) + \sigma_\alpha h_\alpha(y, \mathbf{c}) = \sigma_\alpha \mathcal{L}_\alpha\{h_1, h_2\}(y, \mathbf{c}) \tag{92}$$

for $y > 0$, the dimensionless variable (measured in terms of the mean-free path l_0), σ_α given in (90), and the collision operator defined in (82) and in Appendix A. For defining the temperature-jump problem, we supplement (92) with boundary condition, written in terms of the Cercignani–Lampis kernel as

(93)
$$h_\alpha(0, c_x, -c_y, c_z) = \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty h_\alpha(0, c'_x, c'_y, c'_z) R_\alpha(c'_x, c'_y, c'_z : c_x, -c_y, c_z) \mathrm{d}c'_x \mathrm{d}c'_z \mathrm{d}c'_y,$$

for $c_y > 0$, all c_x, c_z, and $\alpha = 1, 2$. Here

(94a)
$$R_\alpha(c'_x, c'_y, c'_z : c_x, c_y, c_z) = \frac{2c'_y}{\pi a_{n\alpha} a_{t\alpha}(2 - a_{t\alpha})} T_\alpha(c'_x : c_x) S_\alpha(c'_y : c_y) T_\alpha(c'_z : c_z),$$

$$T_\alpha(x : z) = \exp\left[-\frac{[(1 - a_{t\alpha})z - x]^2}{a_{t\alpha}(2 - a_{t\alpha})}\right], \tag{94b}$$

$$S_\alpha(x : z) = \exp\left[-\frac{[(1 - a_{n\alpha})^{1/2} z - x]^2}{a_{n\alpha}}\right] \widehat{I}_0\left[\frac{2(1 - a_{n\alpha})^{1/2} |xz|}{a_{n\alpha}}\right], \tag{94c}$$

and $\widehat{I}_0(z)$ is defined as in (11a). In addition, $a_{t\alpha}$ are the tangential momentum accommodation coefficients for each species α ($\alpha = 1, 2$), and $a_{n\alpha}$ are the normal accommodation coefficients for each species α ($\alpha = 1, 2$). The formulation of the temperature-jump problem will be complete with the use of the condition, at infinity, on the temperature perturbation, as we show later on in this text.

In addition to the temperature perturbation

$$T_\alpha(y) = \frac{2}{3}\pi^{-3/2} \int_{-\infty}^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \mathrm{e}^{-c'^2} h_\alpha(y, c_x, c_y, c_z) \left(\mathbf{c}^2 - \frac{3}{2}\right) \mathrm{d}c_x \mathrm{d}c_y \mathrm{d}c_z \tag{95}$$

we seek to compute also the density perturbation

$$N_\alpha(y) = \pi^{-3/2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{e}^{-c'^2} h_\alpha(y, c_x, c_y, c_z) \mathrm{d}c_x \mathrm{d}c_y \mathrm{d}c_z. \tag{96}$$

Since we do not have to compute the complete distribution h, as we did for the one gas case, we develop a simpler formulation, by multiplying successively (92) by (15) and (16) and defining

$$g_{2\alpha-1}(y, \xi) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi_1(c_x, c_z) h_\alpha(y, c_x, \xi, c_z) \mathrm{d}c_x \mathrm{d}c_z \tag{97}$$

and

$$g_{2\alpha}(y, \xi) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi_2(c_x, c_z) h_\alpha(y, c_x, \xi, c_z) \mathrm{d}c_x \mathrm{d}c_z, \tag{98}$$

for $\alpha = 1, 2$. We then obtain four coupled equations, which we write in a matrix form, for the components of $\mathbf{G}(y, \xi)$ given by (97) and (98),

$$\xi \frac{\partial}{\partial y} \mathbf{G}(y, \xi) + \mathbf{\Sigma} \mathbf{G}(y, \xi) = \mathbf{\Sigma} \int_{-\infty}^{\infty} \psi(\xi') \boldsymbol{K}_M(\xi', \xi) \mathbf{G}(y, \xi') \mathrm{d}\xi', \tag{99}$$

with

$$\mathbf{\Sigma} = \mathrm{diag}\{\sigma_1, \sigma_1, \sigma_2, \sigma_2\} \tag{100}$$

and $\psi(\xi)$ given in (21). In regard to (99), we list the elements $k_{i,j}(\xi', \xi)$ of the (matrix) kernel $\boldsymbol{K}_M(\xi', \xi)$ in Appendix B.

To establish the matrix form of the boundary condition, once more, we use the projections defined by (15) and (16) along with (97) and (98) to obtain, from (93),

$$\mathbf{G}(0, \xi) = \mathbf{D} \int_0^{\infty} \mathbf{F}(\xi', \xi) \mathbf{G}(0, -\xi) \mathrm{d}\xi', \quad \xi > 0, \tag{101}$$

where

$$\mathbf{D} = \mathrm{diag}\left\{1, (1 - a_{t1})^2, 1, (1 - a_{t2})^2\right\} \tag{102}$$

and

$$\mathbf{F}(\xi', \xi) = \mathrm{diag}\left\{f_1(\xi', \xi), f_1(\xi', \xi), f_2(\xi', \xi), f_2(\xi', \xi)\right\}. \tag{103}$$

Here

$$f_1(\xi', \xi) = \frac{2\xi'}{a_{n1}} \exp\left[-\frac{[(1 - a_{n1})^{1/2}\xi - \xi']^2}{a_{n1}}\right] \widehat{I}_0\left[\frac{2(1 - a_{n1})^{1/2}\xi'\xi}{a_{n1}}\right] \tag{104}$$

and

$$f_2(\xi', \xi) = \frac{2\xi'}{a_{n2}} \exp\left[-\frac{[(1 - a_{n2})^{1/2}\xi - \xi']^2}{a_{n2}}\right] \widehat{I}_0\left[\frac{2(1 - a_{n2})^{1/2}\xi'\xi}{a_{n2}}\right], \tag{105}$$

for $\xi, \xi' \in (0, \infty)$.

Once we solve the vector problem, we can compute the density and the temperature perturbations, respectively, given by

$$N_\alpha(y) = \int_{-\infty}^{\infty} \psi(\xi) g_{2\alpha-1}(y,\xi)\mathrm{d}\xi, \quad \alpha = 1,2, \tag{106}$$

and

$$T_\alpha(y) = \frac{2}{3}\int_{-\infty}^{\infty} \psi(\xi)\left[\left(\xi^2 - \frac{1}{2}\right) g_{2\alpha-1}(y,\xi) + g_{2\alpha}(y,\xi)\right]\mathrm{d}\xi, \quad \alpha = 1,2, \tag{107}$$

and we write, in this new notation, the condition imposed to the behavior of the solution at infinity [32]

$$\lim_{y\to\infty} \frac{\mathrm{d}}{\mathrm{dy}}\mathbf{T}(y) = K\begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{108}$$

Here K is considered known and $T_\alpha(y)$ are the components of the vector $\mathbf{T}(y)$.

In the next section we develop a discrete-ordinates solution for the problem defined by (99) to (108).

7. A discrete-ordinates solution (II). When seeking (4×1) solutions of (99) in the form

$$\mathbf{G}(y,\xi) = \mathbf{\Phi}(\nu,\xi)\ \mathrm{e}^{-y/\nu} \tag{109}$$

we follow the same steps described in section 4, when dealing with the one gas case. Then we do not repeat it here, but we write the eigenvalue problem relevant to this case:

$$(1/\xi_i^2)\left[\mathbf{\Sigma}^2\mathbf{V}(\nu_j,\xi_i) - \sum_{k=1}^{N}\omega_k\psi(\xi_k)\mathbf{P}(\xi_k,\xi_i)\mathbf{V}(\nu_j,\xi_k)\right] = \lambda_j\mathbf{V}(\nu_j,\xi_i) \tag{110}$$

with

$$\mathbf{U}(\nu_j,\xi_i) = (\nu_j/\xi_i)\mathbf{\Sigma}\Bigg[\mathbf{V}(\nu_j,\xi_i) - \sum_{k=1}^{N}\omega_k\psi(\xi_k)[\boldsymbol{K}_M(\xi_k,\xi_i)\mathbf{V}(\nu_j,\xi_k) - \boldsymbol{K}_M(-\xi_k,\xi_i)\mathbf{V}(\nu_j,\xi_k)]\Bigg] \tag{111}$$

for $i = 1,2,\ldots,N$. Here

$$\mathbf{P}(\xi',\xi) = (\xi/\xi')\mathbf{\Sigma}[\boldsymbol{K}_M(\xi',\xi) + \boldsymbol{K}_M(-\xi',\xi)]\mathbf{\Sigma} + \mathbf{\Sigma}^2[\boldsymbol{K}_M(\xi',\xi) - \boldsymbol{K}_M(-\xi',\xi)] - \int_0^{\infty}\psi(\xi'')(\xi/\xi'')\mathbf{\Sigma}[\boldsymbol{K}_M(\xi'',\xi) + \boldsymbol{K}_M(-\xi'',\xi)]\mathbf{\Sigma}[\boldsymbol{K}_M(\xi',\xi'') - \boldsymbol{K}_M(-\xi',\xi'')]\mathrm{d}\xi''. \tag{112}$$

We are also able to get the elementary solutions

$$\mathbf{\Phi}(\nu_j,\xi_i) = (1/2)\Big[\mathbf{U}(\nu_j,\xi_i) + \mathbf{V}(\nu_j,\xi_i)\Big] \tag{113}$$

and

$$\mathbf{\Phi}(\nu_j,-\xi_i) = (1/2)\Big[\mathbf{U}(\nu_j,\xi_i) - \mathbf{V}(\nu_j,\xi_i)\Big] \tag{114}$$

to write the general solution of the problem given by (99) as

$$\mathbf{G}(y,\pm\xi_i)=\sum_{j=1}^{4N}[A_j\mathbf{\Phi}(\nu_j,\pm\xi_i)\mathrm{e}^{-y/\nu_j}+B_j\mathbf{\Phi}(\nu_j,\mp\xi_i)\mathrm{e}^{y/\nu_j}] \tag{115}$$

for $i=1,2,\ldots,N$.

As noted in the one gas case, and as usual for conservative problems as these in the rarefied gas dynamics, we find again some unbounded separation constants: in this case the number is three. So, we rewrite the previous equation as

$$\mathbf{G}(y,\pm\xi_i)=\mathbf{G}^*(y,\pm\xi_i)+\sum_{j=4}^{4N}[A_j\mathbf{\Phi}(\nu_j,\pm\xi_i)\mathrm{e}^{-y/\nu_j}+B_j\mathbf{\Phi}(\nu_j,\mp\xi_i)\mathrm{e}^{y/\nu_j}] \tag{116}$$

for $i=1,2,\ldots,N$, where

$$\begin{aligned}\mathbf{G}^*(y,\xi)=&A_1\mathbf{G}_1+A_2\mathbf{G}_2+A_3\mathbf{G}_3+B_1\mathbf{G}_4(\xi)\\&+B_2[y\mathbf{H}_1(\xi)+\mathbf{F}_1(\xi)]+B_3[y\mathbf{H}_2(\xi)+\mathbf{F}_2(\xi)]\end{aligned} \tag{117}$$

with the exact solutions of (99) given by [32]

$$\mathbf{G}_1=\begin{bmatrix}1\\0\\0\\0\end{bmatrix},\qquad \mathbf{G}_2=\begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \tag{118}$$

$$\mathbf{G}_3(\xi)=\begin{bmatrix}\xi^2-1/2\\1\\\xi^2-1/2\\1\end{bmatrix}\quad\text{and}\quad \mathbf{G}_4(\xi)=\begin{bmatrix}r\xi\\0\\\xi\\0\end{bmatrix}. \tag{119}$$

Considering, in addition, the functions

$$\mathbf{H}_1(\xi)=\begin{bmatrix}-1+c_1(\xi^2-1/2)\\c_1\\c_1(\xi^2-1/2)\\c_1\end{bmatrix}\quad\text{and}\quad \mathbf{H}_2(\xi)=\begin{bmatrix}c_2(\xi^2-1/2)\\c_2\\-1+c_2(\xi^2-1/2)\\c_2\end{bmatrix} \tag{120}$$

with

$$r=(m_1/m_2)^{1/2},\quad c_1=(n_1/n),\quad c_2=(n_2/n),\quad\text{and}\quad n=n_1+n_2, \tag{121}$$

we can then find functions $\mathbf{F}_1(\xi)$ and $\mathbf{F}_2(\xi)$ such that

$$\mathbf{G}_5(y,\xi)=y\mathbf{H}_1(\xi)+\mathbf{F}_1(\xi)\quad\text{and}\quad \mathbf{G}_6(y,\xi)=y\mathbf{H}_2(\xi)+\mathbf{F}_2(\xi) \tag{122}$$

complete the set of exact solutions we are looking for. In general,

$$\mathbf{F}_\beta(\xi)=-\xi\mathbf{\Sigma}^{-1}\mathbf{H}_\beta(\xi)+\int_{-\infty}^{\infty}\psi(\xi')\boldsymbol{K}(\xi',\xi)\mathbf{F}_\beta(\xi')\mathrm{d}\xi', \tag{123}$$

and, as we did previously in section 4, we write

$$\mathbf{F}_\beta(\xi) = \sum_{\alpha=0}^{3} P_\alpha(\xi)\mathbf{F}_{\beta,\alpha} \tag{124}$$

to find the components of the vectors $\mathbf{F}_{\beta,\alpha}$ required in (124). Here the linear algebraic system with 16 equations, defined to find those components, is rank deficient (rank 12). Since equations (120) are solutions of (99), we can write [32]

$$\mathbf{F}_\beta(\xi) = \mathbf{U}_\beta P_1(\xi) + \mathbf{V}_\beta P_3(\xi), \tag{125}$$

where now the constant vectors $\mathbf{U}_\beta$ and $\mathbf{V}_\beta$ are solutions of the (rank 8) linear systems

$$(\mathbf{I} - \mathbf{A}_1^*)\mathbf{U}_1 - \mathbf{C}_1^*\mathbf{V}_1 = \Big[(c_2/\sigma_1) \quad -(c_1/\sigma_1) \quad -(c_1/\sigma_2) \quad -(c_1/\sigma_2)\Big]^T, \tag{126}$$

$$(\mathbf{I} - \mathbf{D}_1^*)\mathbf{V}_1 - \mathbf{B}_1^*\mathbf{U}_1 = \Big[-(c_1/\sigma_1) \quad 0 \quad -(c_1/\sigma_2) \quad 0\Big]^T \tag{127}$$

and

$$\Big[0 \quad 0 \quad 1 \quad 0\Big]\mathbf{U}_1 = 0, \tag{128}$$

for $\beta = 1$. In addition,

$$(\mathbf{I} - \mathbf{A}_1^*)\mathbf{U}_2 - \mathbf{C}_1^*\mathbf{V}_2 = \Big[-(c_2/\sigma_1) \quad -(c_2/\sigma_1) \quad (c_1/\sigma_2) \quad -(c_2/\sigma_2)\Big]^T, \tag{129}$$

$$(\mathbf{I} - \mathbf{D}_1^*)\mathbf{V}_2 - \mathbf{B}_1^*\mathbf{U}_2 = \Big[-(c_2/\sigma_1) \quad 0 \quad -(c_2/\sigma_2) \quad 0\Big]^T \tag{130}$$

and

$$\Big[0 \quad 0 \quad 1 \quad 0\Big]\mathbf{U}_2 = 0, \tag{131}$$

for $\beta = 2$. Here (128) and (131) were made part of the linear system to guarantee a unique solution. Still, in the above expressions $\mathbf{I}$ is the identity matrix, the superscript T denotes transpose operation, and

$$\mathbf{A}_1^* = 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(\xi')\psi(\xi)\boldsymbol{K}_M(\xi',\xi)P_1(\xi')P_1(\xi)\mathrm{d}\xi'\mathrm{d}\xi, \tag{132}$$

$$\mathbf{B}_1^* = (4/3)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(\xi')\psi(\xi)\boldsymbol{K}_M(\xi',\xi)P_1(\xi')P_3(\xi)\mathrm{d}\xi'\mathrm{d}\xi, \tag{133}$$

$$\mathbf{C}_1^* = 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(\xi')\psi(\xi)\boldsymbol{K}_M(\xi',\xi)P_3(\xi')P_1(\xi)\mathrm{d}\xi'\mathrm{d}\xi \tag{134}$$

and

$$\mathbf{D}_1^* = (4/3)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(\xi')\psi(\xi)\boldsymbol{K}_M(\xi',\xi)P_3(\xi')P_3(\xi)\mathrm{d}\xi'\mathrm{d}\xi. \tag{135}$$

However, (132) to (135) can be evaluated, by means of the quadrature scheme, such that we rewrite them (using notation listed in the appendix) in a final form

$$\mathbf{A}_1^* = \begin{bmatrix} 1-\eta_{1,2}^{(1)} & -(1/2)\eta_{1,2}^{(2)} & r\eta_{1,2}^{(1)} & (r^3/2)\eta_{1,2}^{(2)} \\ -(1/2)\eta_{1,2}^{(2)} & (2/5)\beta_1 & (r/2)\eta_{1,2}^{(2)} & (2/5)\eta_{1,2}^{(6)} \\ s\eta_{2,1}^{(1)} & (s^3/2)\eta_{2,1}^{(2)} & 1-\eta_{2,1}^{(1)} & -(1/2)\eta_{2,1}^{(2)} \\ (s/2)\eta_{2,1}^{(2)} & (2/5)\eta_{2,1}^{(6)} & -(1/2)\eta_{2,1}^{(2)} & (2/5)\beta_2 \end{bmatrix}, \tag{136}$$

$$\mathbf{B}_1^* = \begin{bmatrix} -(1/2)\eta_{1,2}^{(2)} & (2/5)\beta_1 & (r/2)\eta_{1,2}^{(2)} & (2/5)\eta_{1,2}^{(6)} \\ 0 & 0 & 0 & 0 \\ (s/2)\eta_{2,1}^{(2)} & (2/5)\eta_{2,1}^{(6)} & -(1/2)\eta_{2,1}^{(2)} & (2/5)\beta_2 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{137}$$

$$\mathbf{C}_1^* = \begin{bmatrix} -(3/4)\eta_{1,2}^{(2)} & 0 & (3r^3/4)\eta_{1,2}^{(2)} & 0 \\ (3/5)\beta_1 & 0 & (3/5)\eta_{1,2}^{(6)} & 0 \\ (3s^3/4)\eta_{2,1}^{(2)} & 0 & -(3/4)\eta_{2,1}^{(2)} & 0 \\ (3/5)\eta_{2,1}^{(6)} & 0 & (3/5)\beta_2 & 0 \end{bmatrix} \tag{138}$$

and

$$\mathbf{D}_1^* = \begin{bmatrix} (3/5)\beta_1 & 0 & (3/5)\eta_{1,2}^{(6)} & 0 \\ 0 & 0 & 0 & 0 \\ (3/5)\eta_{2,1}^{(6)} & 0 & (3/5)\beta_2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{139}$$

Thus, looking back to (116), the general discrete-ordinates solution for (99) can be written as

$$\mathbf{G}(y, \pm\xi_i) = \mathbf{G}^*(y, \pm\xi_i) + \sum_{j=4}^{4N} A_j \mathbf{\Phi}(\nu_j, \pm\xi_i) \mathrm{e}^{-y/\nu_j}, \tag{140}$$

taking into account that it is not allowed to diverge exponentially as y tends to infinity.

If we use (140) to evaluate (the discrete-ordinates version) (106) and (107) and we take the asymptotic part (the exponential terms were ignored) of the resulting expressions, we can write

$$N_1^*(y) = A_1 - B_2 y, \tag{141}$$

$$N_2^*(y) = A_2 - B_3 y, \tag{142}$$

$$T_1^*(y) = A_3 + (c_1 B_2 + c_2 B_3) y \tag{143}$$

and

$$T_2^*(y) = A_3 + (c_1 B_2 + c_2 B_3) y. \tag{144}$$

But again, as in the one gas case, two of the exact solutions, (118), satisfy the boundary condition, (101), such that the constants A_1 and A_2 cannot be determined from that boundary condition (density perturbations are not uniquely defined). As we did in the one gas case we make use of the condition given by (108) along with a normalization condition, similar to (62), to write

$$A_1 = -A_3, \tag{145}$$

$$A_2 = -A_3 \tag{146}$$

and

$$c_1 B_2 + c_2 B_3 = K. \tag{147}$$

In (147) we choose to write B_3 in terms of B_2 such that the general solution of the problem given by (99) is finally written as in (140) with

$$\begin{aligned}\mathbf{G}_*(y,\xi) = A_3\mathbf{R}(\xi) + B_1\mathbf{G}_4(\xi) + B_2\Big[[y\mathbf{H}_1(\xi) + \mathbf{F}_1(\xi)] - (c_1/c_2)[y\mathbf{H}_2(\xi) + \mathbf{F}_2(\xi)]\Big] \\ + (K/c_2)[y\mathbf{H}_2(\xi) + \mathbf{F}_2(\xi)].\end{aligned} \tag{148}$$

Here

$$\mathbf{R}(\xi) = \begin{bmatrix} \xi^2 - 3/2 \\ 1 \\ \xi^2 - 3/2 \\ 1 \end{bmatrix}. \tag{149}$$

The application of (140) in the discrete ordinates version of the boundary condition

$$\mathbf{G}(0,\xi_i) = \mathbf{D}\sum_{k=1}^{N}\omega_k\mathbf{F}(\xi_k,\xi_i)\mathbf{G}(0,-\xi_k) \tag{150}$$

results in the following linear system:

$$\begin{aligned}A_3\mathbf{C^3}(\xi_i) + B_1\mathbf{C^4}(\xi_i) + B_2\mathbf{C^5}(\xi_i) + \sum_{j=4}^{4N} A_j\left\{\mathbf{\Phi}_+(\nu_j) - \mathbf{D}^*\sum_{k=1}^{N}\omega_k\mathbf{F}^*(\xi_k,\xi_i)\mathbf{\Phi}_-(\nu_j)\right\} \\ = K\left\{-\mathbf{J}_1^*(\xi_i) + \mathbf{D}^*\sum_{k=1}^{N}\omega_k\mathbf{F}^*(\xi_k,\xi_i)\mathbf{J}_1^*(-\xi_k)\right\},\end{aligned} \tag{151}$$

where the components of the vectors $\mathbf{\Phi}_+(\nu_j)$ and $\mathbf{\Phi}_-(\nu_j)$ are given by (113) and (114),

$$\mathbf{C^3}(\xi_i) = \mathbf{R}^*(\xi_i) - \mathbf{D}^*\sum_{k=1}^{N}\omega_k\mathbf{F}^*(\xi_k,\xi_i)\mathbf{R}^*(-\xi_k), \tag{152}$$

$$\mathbf{C^4}(\xi_i) = \mathbf{G}_4^*(\xi_i) - \mathbf{D}^*\sum_{k=1}^{N}\omega_k\mathbf{F}^*(\xi_k,\xi_i)\mathbf{G}_4^*(-\xi_k) \tag{153}$$

and

$$\mathbf{C^5}(\xi_i) = \mathbf{K}_1^*(\xi_i) - \mathbf{D}^* \sum_{k=1}^{N} \omega_k \mathbf{F}^*(\xi_k, \xi_i)\mathbf{K}_1^*(-\xi_k). \tag{154}$$

Here $\mathbf{R}^*(\xi)$ is a $4N \times 1$ vector where each N component is of the type of (149); $\mathbf{K}_1^*(\xi)$ is a $4N \times 1$ vector defined by N components of the type

$$\mathbf{K}_1(\xi) = \mathbf{F}_1(\xi) - (c_1/c_2)\mathbf{F}_2(\xi), \tag{155}$$

and $\mathbf{J}_1^*(\xi)$ is a $4N \times 1$ vector where each N component is of the type

$$\mathbf{J}_1(\xi) = (1/c_2)\mathbf{F}_2(\xi). \tag{156}$$

In addition $\mathbf{D}^*$ and $\mathbf{F}^*(\xi_k, \xi)$ are $4N \times 4N$ diagonal matrices defined, looking back to (102) and (103), by

$$\mathbf{D}^* = \mathrm{diag}\Big\{\mathbf{D}, \mathbf{D}, \ldots, \mathbf{D}\Big\} \tag{157}$$

and

$$\mathbf{F}^*(\xi_k, \xi) = \mathrm{diag}\Big\{\mathbf{F}(\xi_1, \xi), \mathbf{F}(\xi_2, \xi), \ldots, \mathbf{F}(\xi_N, \xi)\Big\}. \tag{158}$$

Based on the derivation above we can write the density perturbation vector (two species) as

$$\mathbf{N}(y) = -\left[\begin{array}{c} A_3 + B_2 y \\ A_3 + \big(K/c_2 - (c_1/c_2)B_2\big)y \end{array}\right] + \sum_{j=4}^{4N} A_j \ \mathrm{e}^{-y/\nu_j} \ \mathbf{M}_1(\nu_j), \tag{159}$$

where

$$\mathbf{M}_1(\nu_j) = \pi^{-1/2} \sum_{k=1}^{N} \omega_k \mathrm{e}^{-{\xi_k}^2} \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right] [\mathbf{\Phi}(\nu_j, \xi_k) + \mathbf{\Phi}(\nu_j, -\xi_k)]. \tag{160}$$

The components $T_\alpha(y)$, $\alpha = 1, 2$, define the temperature perturbation vector

$$\mathbf{T}(y) = (A_3 + Ky)\left[\begin{array}{c} 1 \\ 1 \end{array}\right] + \frac{2}{3} \sum_{j=4}^{4N} A_j \ \mathrm{e}^{-y/\nu_j} \ \mathbf{M}_2(\nu_j), \tag{161}$$

where

$$\mathbf{M}_2(\nu_j) = \pi^{-1/2} \sum_{k=1}^{N} \omega_k \mathrm{e}^{-{\xi_k}^2} \left[\begin{array}{cccc} \xi_k^2 - 1/2 & 1 & 0 & 0 \\ 0 & 0 & \xi_k^2 - 1/2 & 1 \end{array}\right] [\mathbf{\Phi}(\nu_j, \xi_k) + \mathbf{\Phi}(\nu_j, -\xi_k)]. \tag{162}$$

From (161), and taking into account (143) and (144), along with definitions (for each species) analogous to (75), we define the temperature-jump coefficient

$$\zeta = A_3/K, \tag{163}$$

which we found to be the same for both species of the gas.

8. Computational aspects and numerical results. To implement the ADO solution, in order to define the quantities of interest, the first step is to define the associated quadrature scheme. In this sense, for most of the problems in the rarefied gas dynamics field that have been solved by this method [25, 26, 27, 28, 29, 30] the following approach has been shown to be adequate: to map the interval of integration into the interval $[0, 1]$ and then map the Gauss–Legendre points linearly to this interval. In general, the expression

$$u(\xi) = \mathrm{e}^{-\xi} \tag{164}$$

is the one used to map the interval $[0, \infty)$ into $[0, 1]$. After that, the eigenvalue problem which defines the separation constants and the elementary solutions can be formulated and solved. Next, the linear systems which define the exact solutions have to be considered. Continuing, the linear algebraic system for the arbitrary constants of the general solution is solved and then the quantities of interest are evaluated. In this way, the algorithm is fast and easy to implement.

The results presented in Tables 1 to 14 (which we believe to be correct for all digits listed, plus or minus one in the last digit) were obtained with $N = 60$ quadrature points for the one gas case and $N = 80$ quadrature points for the mixtures case (except for $\alpha_n = 0.01$, where $N = 120$). For both cases we consider $K = 1$. The developed (and not particularly optimized) Fortran code requires around four seconds to yield, on a 3.0 GHz Pentium 4 machine, all quantities of interest for a general mixture case.

For the one gas case we used $\varepsilon = \varepsilon_t$ (according to (8) and (9)), in general, in order to obtain the numerical results shown (otherwise we added the statement $\varepsilon = \varepsilon_p$ in the table). We also note the notation DS and CL for referring to the use of, respectively, diffuse-specular and Cercignani–Lampis boundary condition. In fact, as we have worked on trying to establish a very general and complete comparison on results for different problems, provided by different model equations, we list in this work some results we obtained by the ADO method, for the temperature-jump problem, based on the BGK and S-models with diffuse-specular and Cercignani–Lampis boundary condition, which were not available.

Continuing, we note that, in Tables 8 to 11, and in the text below, we use the notation $a_{n\alpha}$ and $a_{t\alpha}$, for the species $\alpha = 1, 2$, to refer to the accommodation coefficients, respectively, of energy and tangential momentum.

In order to generate numerical results, we consider three mixtures cases related to the following gases: (i) Ne-Ar, (ii) He-Xe and (iii) He-Ar, as follow:

- Ne-Ar: $m_1 = 20.183$, $m_2 = 39.948$, and $d_2/d_1 = 1.406$
  - case I $a_{t1} = 0.849$, $a_{t2} = 0.916$, $a_{n1} = 0.1$, $a_{n2} = 0.4$
  - case II $a_{t1} = 0.31$, $a_{t2} = 0.67$, $a_{n1} = 0.1$, $a_{n2} = 0.4$
  - case III $a_{t1} = 0.849$, $a_{t2} = 0.916$, $a_{n1} = 0.082$, $a_{n2} = 0.222$
- He-Xe: $m_1 = 4.0026$, $m_2 = 131.30$, and $d_2/d_1 = 2.226$
  - case IV $a_{t1} = 0.20$, $a_{t2} = 0.95$, $a_{n1} = 0.01$, $a_{n2} = 0.7$
  - case V $a_{t1} = 0.20$, $a_{t2} = 0.95$, $a_{n1} = 0.05$, $a_{n2} = 0.4$
  - case VI $a_{t1} = 0.882$, $a_{t2} = 1.014$, $a_{n1} = 0.01$, $a_{n2} = 0.7$
- He-Ar: $m_1 = 4.0026$, $m_2 = 39.948$, and $d_2/d_1 = 1.665$
  - case VII $a_{t1} = 0.20$, $a_{t2} = 0.916$, $a_{n1} = 0.01$, $a_{n2} = 0.4$
  - case VIII $a_{t1} = 0.882$, $a_{t2} = 0.91$, $a_{n1} = 0.01$, $a_{n2} = 0.4$
  - case IX $a_{t1} = 0.20$, $a_{t2} = 0.916$, $a_{n1} = 0.01$, $a_{n2} = 0.222$

The data set used, and listed above, regarding the mass and diameter of the particles, was reproduced from [32, 41, 42].

In addition, we used experimental data for defining the values of the tangential momentum accommodation coefficient: cases II, IV, and V were formulated in terms of data provided by Lord [43]; in cases I, III, and VI, the value of that coefficient was reproduced from [11], which followed the experimental work of Porodnov et al. [44]. Cases VII to IX were formulated based on the previous cases.

Since, to the best of our knowledge, experimental results for the (normal) energy accommodation coefficient are not available, we chose numerical values based (order of magnitude) on the thermal accommodation coefficient of the gases listed above. We used, in this way, the works of Thomas [45] and Thomas and Lord [46].

The results were tabulated, for the three mixtures, in terms of the molar concentration defined in terms of the first particle as

$$C = \frac{n_1/n_2}{1 + n_1/n_2}. \tag{165}$$

In addition to provide results, which we believe were not available in the literature, for this problem, based on a more general law for describing the gas-surface interaction, we have been able to perform many simulations from which we try to point out some aspects, commented below.

First of all, we think that an important aspect of providing results, based on models related to the normal energy accommodation coefficient, which may not be measured, is the possibility of using quantities we are able to evaluate, in connection with a procedure for estimating parameters [47], in order to estimate that coefficient numerically. In this sense, we include in Tables 1 to 14, in addition to the temperature-jump coefficient, results for temperature and density perturbations, for several values of accommodation coefficients, and for different values of concentration, in the mixture cases.

In regard to the one gas case, based on the experience with previous analysis for other classes of problems [28], the fact of getting results for the BGK and S-model in agreement, in general, in one or two digits (for a choice of the mean-free path), was, in fact, used as a way of having some confidence in our program. In addition, we confirm previous observation [32] that both models lead to the same value for the temperature-jump coefficient but slightly different values for the temperature and density perturbation; we also simulate (considering $n_1 = 0$, $n_2 = 0$ or $m_1 = m_2$ and $d_1 = d_2$) the one gas case, from the McCormack model, and the results agree perfectly with the S-model, as observed previously [32].

Having used the ADO method, which is an analytical approach to the spatial variable, to deal with the temperature-jump problem for a wide class of model equations, in the one gas case, with Cercignani–Lampis and Maxwell boundary condition, and having compared with available results [48] based on the linearized Boltzmann equation (LBE), we can say that in general the evaluation of this coefficient is not sensitive to the model equation to be used (see Table 7); comparisons between results (see Tables 1 to 5) obtained by the S-model and the LBE show agreement in one to two digits.

We generate results, showed in Tables 2 to 4, for a small value of α_n, which is, in general, hard to obtain accurately or even very much time consuming, from numerical approaches, but is consistent, in order of magnitude, with experimental values available for the total energy accommodation coefficient. It is noted that the

TABLE 1
The temperature-jump coefficient ζ.

α	BGK-DS	S-DS($\varepsilon = \varepsilon_p$)	S-DS	LBE-DS[31]
0.1	2.145012(1)	3.21752(1)	2.14501(1)	2.1349(1)
0.2	1.034747(1)	1.55212(1)	1.03475(1)	1.0251(1)
0.3	6.630514	9.94577	6.63051	6.5396
0.4	4.760333	7.14050	4.76033	4.6745
0.5	3.629125	5.44369	3.62912	3.5485
0.6	2.867615	4.30142	2.86761	2.7922
0.7	2.317534	3.47630	2.31753	2.2474
0.8	1.899741	2.84961	1.89974	1.8349
0.9	1.570264	2.35540	1.57026	1.5108
1.0	1.302716	1.95407	1.30272	1.2486

TABLE 2
The temperature-jump coefficient ζ, $\alpha_t = 0.25$.

α_n	BGK-CL	S-CL($\varepsilon = \varepsilon_p$)	S-CL	S-CL[24]	LBE-CL[48]
0.01	9.75601	1.46340(1)	9.75601	—	—
0.25	5.78950	8.68426	5.78950	8.684	5.7318
0.5	3.84176	5.76263	3.84176	5.763	3.7707
0.75	2.72408	4.08612	2.72408	4.087	2.6655
1.0	2.00553	3.00830	2.00553	3.009	1.9609

temperature-jump coefficient is very sensitive to the normal energy accommodation coefficient. In fact, from the cases shown in Figures 1 and 2, it is clear that the temperature-jump coefficient depends on both accommodation coefficients (the same is observed for the mixtures cases later on) and the significant variation occurs mainly when α_n and α_t are lower than 0.5. In these cases, in comparison with the results shown in Table 1 based on the Maxwell boundary condition, we see great increase in the jump values.

In Tables 12 to 14, we present results for the temperature-jump coefficient for the mixtures cases I to IX described previously in this section. We included the cases of concentration equal to one ($n_1 = 1$ and $n_2 = 0$) and equal to zero ($n_1 = 0$ and $n_2 = 1$), which reproduce exactly the results obtained for the one gas case (respectively, gas 1 and gas 2) and that may show how mixing a very small quantity of a different species can produce significant change in the jump coefficient. In fact, as observed for the case of Maxwell boundary conditions [32, 42], in Table 13, we note a small decrease on the value of the coefficient when we simulate the one gas case (species one). It seems that the (bigger) molecular mass ratio can be a reason for noticing this behavior; however, the same (decreasing) behavior is observed in Table 14, for the case VIII, and then, based on comparisons with cases VII and IX, one can also see for the same mixture the influence of the accommodation coefficients. However, one has to remember that the real values for α_n are unknown, and it has to be considered as a limitation for more specific conclusions.

Finally, in Figures 3 to 6, we tried to analyze the influence of each one of the accommodation coefficients, for the mixtures cases, based on cases I to VI. It really shows that the temperature-jump coefficient depends on both accommodation coefficients and mainly on the value of α_n. In addition, from Figures 4 and 6, the influence of the mass ratio in the magnitude order of the jump coefficient can be seen.

TABLE 3
The temperature-jump coefficient ζ, $\alpha_t = 0.5$.

α_n	BGK-CL	S-CL($\varepsilon = \varepsilon_p$)	S-CL	S-CL[24]	LBE-CL[48]
0.01	5.66914	8.50371	5.66914	—	—
0.25	3.88593	5.82890	3.88593	5.828	3.8696
0.5	2.78041	4.17061	2.78041	4.170	2.7282
0.75	2.05839	3.08759	2.05839	3.088	2.0010
1.0	1.55658	2.33487	1.55658	2.335	1.5015

TABLE 4
The temperature-jump coefficient ζ, $\alpha_t = 1.0$.

α_n	BGK-CL	S-CL($\varepsilon = \varepsilon_p$)	S-CL	S-CL[24]	LBE-CL[48]
0.01	4.20852	6.31278	4.20852	—	—
0.25	3.04475	4.56713	3.04475	4.567	3.0524
0.5	2.25090	3.37635	2.25090	3.376	2.2161
0.75	1.70032	2.55048	1.70032	2.551	1.6514
1.0	1.30272	1.95407	1.30272	1.954	1.2486

TABLE 5
Temperature $T(y)$ and density $N(y)$ perturbations, $\alpha = 0.5$.

	S-DS($\varepsilon = \varepsilon_p$)		S-DS		BGK-DS[30]		LBE-DS[31]	
y	$T(y)$	$N(y)$	$T(y)$	$N(y)$	$T(y)$	$N(y)$	$T(y)$	$N(y)$
0.0	4.37396	–4.61156	2.91597	–3.07437	2.91597	–3.07437	2.9250	–3.1153
0.1	4.72063	–4.92495	3.22570	–3.35355	3.18042	–3.31664	3.2342	–3.3647
0.2	4.94416	–5.12484	3.42167	–3.52947	3.36278	–3.48323	3.4238	–3.5257
0.3	5.13250	–5.29421	3.58709	–3.67966	3.52167	–3.62947	3.5831	–3.6654
0.4	5.30131	–5.44717	3.73615	–3.81653	3.66754	–3.76478	3.7268	–3.7944
0.5	5.45733	–5.58965	3.87478	–3.94514	3.80489	–3.89310	3.8605	–3.9167
0.6	5.60422	–5.72479	4.00614	–4.06814	3.93615	–4.01653	3.9874	–4.0345
0.7	5.74424	–5.85449	4.13216	–4.18706	4.06283	–4.13633	4.1093	–4.1491
0.8	5.87889	–5.98001	4.25410	–4.30291	4.18593	–4.25334	4.2275	–4.2612
0.9	6.00921	–6.10220	4.37281	–4.41637	4.30614	–4.36814	4.3428	–4.3715
1.0	6.13600	–6.22170	4.48894	–4.52793	4.42400	–4.48113	4.4558	–4.4802
2.0	7.29393	–7.33511	5.57466	–5.58912	5.52928	–5.55674	5.5196	–5.5251
3.0	8.36200	–8.38367	6.60474	–6.61081	6.57466	–6.58912	–	–
4.0	9.39637	–9.40840	7.61737	–7.62010	7.59758	–7.60560	–	–
5.0	1.04152(1)	–1.04221(1)	8.62318	–8.62448	8.61013	–8.61476	–	–
6.0	1.14260(1)	–1.14302(1)	9.62601	–9.62665	9.61737	–9.62011	–	–
7.0	1.24325(1)	–1.24350(1)	1.06274(1)	–1.06278(1)	1.06217(1)	–1.06234(1)	–	–
8.0	1.34365(1)	–1.34380(1)	1.16282(1)	–1.16284(1)	1.16243(1)	–1.16254(1)	–	–
9.0	1.44390(1)	–1.44400(1)	1.26286(1)	–1.26287(1)	1.26260(1)	–1.26267(1)	–	–
10.0	1.54406(1)	–1.54412(1)	1.36288(1)	–1.36289(1)	1.36271(1)	–1.36275(1)	–	–
20.0	2.54436(1)	–2.54436(1)	2.36291(1)	–2.36291(1)	2.36291(1)	–2.36291(1)	–	–

TABLE 6
Temperature $T(y)$ and density $N(y)$ perturbations, $\alpha_t = 0.5$, $\alpha_n = 0.5$.

	BGK-CL		S-CL($\varepsilon = \varepsilon_p$)		S-CL	
y	$T(y)$	$N(y)$	$T(y)$	$N(y)$	$T(y)$	$N(y)$
0.0	2.10157	–2.56938	3.15236	–3.85407	2.10157	–2.56938
0.1	2.34143	–2.71400	3.46214	–4.02100	2.38212	–2.72494
0.2	2.51598	–2.83371	3.67398	–4.15056	2.57076	–2.84749
0.3	2.67076	–2.94749	3.85614	–4.27124	2.73298	–2.96295
0.4	2.81423	–3.05829	4.02135	–4.38743	2.88054	–3.07492
0.5	2.95014	–3.16724	4.17520	–4.50085	3.01854	–3.18477
0.6	3.08054	–3.27492	4.32081	–4.61238	3.14976	–3.29315
0.7	3.20674	–3.38167	4.46011	–4.72251	3.27590	–3.40045
0.8	3.32960	–3.48770	4.59441	–4.83155	3.39810	–3.50691
0.9	3.44976	–3.59315	4.72464	–4.93973	3.51717	–3.61267
1.0	3.56767	–3.69812	4.85151	–5.04718	3.63369	–3.71786
2.0	4.67538	–4.73195	6.01307	–6.09793	4.72288	–4.75046
3.0	5.72288	–5.75046	7.08432	–7.12570	5.75467	–5.76525
4.0	6.74708	–6.76147	8.12062	–8.14220	6.76804	–6.77248
5.0	7.76037	–7.76823	9.14056	–9.15235	7.77418	–7.77617
6.0	8.76804	–8.77248	1.01520(1)	–1.01587(1)	8.77716	–8.77810
7.0	9.77261	–9.77519	1.11589(1)	–1.11628(1)	9.77867	–9.77912
8.0	1.07754(1)	–1.07769(1)	1.21631(1)	–1.21654(1)	1.07795(1)	–1.07797(1)
9.0	1.17772(1)	–1.17781(1)	1.31657(1)	–1.31671(1)	1.17799(1)	–1.17800(1)
10.0	1.27783(1)	–1.27788(1)	1.41674(1)	–1.41683(1)	1.27801(1)	–1.27802(1)
20.0	2.27804(1)	–2.27804(1)	2.41705(1)	–2.41705(1)	2.27804(1)	–2.27804(1)

TABLE 7
The temperature-jump coefficient ζ, $\alpha = 0.5$ with Maxwell boundary condition.

Model	ζ
BGK	3.629125[30]
Williams	3.435960[30]
Rigid-sphere	3.476180[30]
S ($\varepsilon = \varepsilon_p$)	5.443688
S ($\varepsilon = \varepsilon_t$)	3.629125
LBE	3.5485[31]

TABLE 8

Density and temperature perturbations for the mixture Ne-Ar, $C = 0.3$, $a_{t1} = 0.31$, $a_{n1} = 0.10$, $a_{t2} = 0.67$, $a_{n2} = 0.40$, $\zeta = 5.609842$.

y	$N_1(y)$	$N_2(y)$	$T_1(y)$	$T_2(y)$
0.0	–6.094582	–5.020045	5.003068	3.628826
0.1	–6.093327	–5.196696	5.221407	4.100402
0.2	–6.114674	–5.343053	5.381062	4.408179
0.3	–6.147316	–5.482370	5.524230	4.664051
0.4	–6.187790	–5.617862	5.658380	4.889534
0.5	–6.234186	–5.750740	5.786684	5.094294
0.6	–6.285266	–5.881609	5.910838	5.283771
0.7	–6.340156	–6.010827	6.031871	5.461415
0.8	–6.398206	–6.138629	6.150460	5.629586
0.9	–6.458912	–6.265184	6.267077	5.789991
1.0	–6.521875	–6.390621	6.382064	5.943908
2.0	–7.225736	–7.600364	7.479400	7.273053
3.0	–7.990690	–8.758942	8.529533	8.421366
4.0	–8.775655	–9.889628	9.558614	9.498200
5.0	–9.567645	–1.100422(1)	1.057641(1)	1.054113(1)
6.0	–1.036200(1)	–1.210921(1)	1.158764(1)	1.156635(1)
7.0	–1.115696(1)	–1.320832(1)	1.259490(1)	1.258170(1)
8.0	–1.195190(1)	–1.430375(1)	1.359967(1)	1.359130(1)
9.0	–1.274662(1)	–1.539684(1)	1.460285(1)	1.459745(1)
10.0	–1.354108(1)	–1.648842(1)	1.560499(1)	1.560145(1)
20.0	–2.147838(1)	–2.738018(1)	2.560967(1)	2.560958(1)

TABLE 9

Density and temperature perturbations for the mixture Ne-Ar, $C = 0.8$, $a_{t1} = 0.31$, $a_{n1} = 0.10$, $a_{t2} = 0.67$, $a_{n2} = 0.40$, $\zeta = 8.594219$.

y	$N_1(y)$	$N_2(y)$	$T_1(y)$	$T_2(y)$
0.0	–8.750733	–7.120369	7.090958	5.242827
0.1	–8.783852	–7.506333	7.459928	6.084367
0.2	–8.835794	–7.786932	7.709797	6.587750
0.3	–8.897839	–8.033746	7.922874	6.982114
0.4	–8.966726	–8.260064	8.114582	7.312956
0.5	–9.040610	–8.471952	8.291754	7.600955
0.6	–9.118298	–8.672914	8.458194	7.857802
0.7	–9.198955	–8.865234	8.616298	8.090910
0.8	–9.281972	–9.050521	8.767701	8.305339
0.9	–9.366890	–9.229963	8.913578	8.504716
1.0	–9.453352	–9.404472	9.054812	8.691742
2.0	–1.036126(1)	–1.098707(1)	1.031358(1)	1.018410(1)
3.0	–1.129379(1)	–1.242872(1)	1.143514(1)	1.138000(1)
4.0	–1.222722(1)	–1.381496(1)	1.249962(1)	1.247343(1)
5.0	–1.315743(1)	–1.517507(1)	1.353609(1)	1.352266(1)
6.0	–1.408435(1)	–1.652151(1)	1.455762(1)	1.455032(1)
7.0	–1.500869(1)	–1.786030(1)	1.557073(1)	1.556658(1)
8.0	–1.593117(1)	–1.919461(1)	1.657891(1)	1.657646(1)
9.0	–1.685235(1)	–2.052622(1)	1.758410(1)	1.758262(1)
10.0	–1.777263(1)	–2.185615(1)	1.858746(1)	1.858654(1)
20.0	–2.695996(1)	–3.513052(1)	2.859404(1)	2.859402(1)

TABLE 10

Density and temperature perturbations for the mixture He-Xe, $C = 0.3$, $a_{t1} = 0.20$, $a_{n1} = 0.05$, $a_{t2} = 0.95$, $a_{n2} = 0.40$, $\zeta = 9.208216$.

y	$N_1(y)$	$N_2(y)$	$T_1(y)$	$T_2(y)$
0.0	–9.859661	–6.752228	8.842905	4.550386
0.1	–9.879711	–6.953128	9.108078	5.171764
0.2	–9.907475	–7.131582	9.294006	5.587051
0.3	–9.940046	–7.306519	9.454962	5.936198
0.4	–9.976171	–7.479783	9.601494	6.246081
0.5	–1.001510(1)	–7.651818	9.738338	6.528768
0.6	–1.005634(1)	–7.822732	9.868148	6.791054
0.7	–1.009952(1)	–7.992533	9.992600	7.037259
0.8	–1.014435(1)	–8.161199	1.011283(1)	7.270349
0.9	–1.019063(1)	–8.328707	1.022967(1)	7.492473
1.0	–1.023817(1)	–8.495034	1.034371(1)	7.705251
2.0	–1.075993(1)	–1.009326(1)	1.139936(1)	9.508343
3.0	–1.132934(1)	–1.158524(1)	1.239185(1)	1.099113(1)
4.0	–1.192301(1)	–1.299418(1)	1.336773(1)	1.231344(1)
5.0	–1.253117(1)	–1.434063(1)	1.434089(1)	1.353987(1)
6.0	–1.314884(1)	–1.564032(1)	1.531607(1)	1.470376(1)
7.0	–1.377312(1)	–1.690495(1)	1.629469(1)	1.582465(1)
8.0	–1.440213(1)	–1.814319(1)	1.727694(1)	1.691501(1)
9.0	–1.503463(1)	–1.936146(1)	1.826251(1)	1.798316(1)
10.0	–1.566972(1)	–2.056457(1)	1.925092(1)	1.903493(1)
20.0	–2.207846(1)	–3.225069(1)	2.921173(1)	2.919434(1)

TABLE 11

Density and temperature perturbations for the mixture He-Xe, $C = 0.8$, $a_{t1} = 0.20$, $a_{n1} = 0.05$, $a_{t2} = 0.95$, $a_{n2} = 0.40$, $\zeta = 16.699401$.

y	$N_1(y)$	$N_2(y)$	$T_1(y)$	$T_2(y)$
0.0	–1.683016(1)	–1.227521(1)	1.458384(1)	8.714419
0.1	–1.686467(1)	–1.286386(1)	1.500194(1)	9.979183
0.2	–1.690923(1)	–1.332670(1)	1.528673(1)	1.079195(1)
0.3	–1.696028(1)	–1.374765(1)	1.552854(1)	1.145073(1)
0.4	–1.701606(1)	–1.414201(1)	1.574502(1)	1.201558(1)
0.5	–1.707549(1)	–1.451664(1)	1.594411(1)	1.251425(1)
0.6	–1.713784(1)	–1.487549(1)	1.613030(1)	1.296271(1)
0.7	–1.720259(1)	–1.522113(1)	1.630641(1)	1.337133(1)
0.8	–1.726932(1)	–1.555540(1)	1.647441(1)	1.374735(1)
0.9	–1.733772(1)	–1.587970(1)	1.663570(1)	1.409609(1)
1.0	–1.740755(1)	–1.619513(1)	1.679132(1)	1.442160(1)
2.0	–1.815177(1)	–1.900579(1)	1.815832(1)	1.688453(1)
3.0	–1.893156(1)	–2.143660(1)	1.935084(1)	1.861137(1)
4.0	–1.972025(1)	–2.366427(1)	2.046522(1)	2.001624(1)
5.0	–2.050975(1)	–2.577354(1)	2.153768(1)	2.125641(1)
6.0	–2.129757(1)	–2.781055(1)	2.258546(1)	2.240502(1)
7.0	–2.208315(1)	–2.980205(1)	2.361789(1)	2.349988(1)
8.0	–2.286658(1)	–3.176423(1)	2.464036(1)	2.456193(1)
9.0	–2.364816(1)	–3.370713(1)	2.565619(1)	2.560336(1)
10.0	–2.442825(1)	–3.563718(1)	2.666750(1)	2.663148(1)
20.0	–3.219264(1)	–5.471661(1)	3.669732(1)	3.669609(1)

TABLE 12
The temperature-jump coefficient ζ for the mixture Ne-Ar with CL boundary condition.

C	case I	case II	case III
0.00	3.819923	4.165162	4.763477
0.10	4.026792	4.624545	4.884249
0.20	4.229354	5.105370	5.001942
0.30	4.427408	5.609842	5.116410
0.40	4.620807	6.140509	5.227559
0.50	4.809467	6.700354	5.335369
0.60	4.993409	7.292911	5.439917
0.70	5.172797	7.922448	5.541419
0.80	5.348012	8.594219	5.640298
0.90	5.519769	9.314559	5.737289
0.94	5.587756	9.560220	5.775806
1.00	5.689301	1.009305(1)	5.833604

TABLE 13
The temperature-jump coefficient ζ for the mixture He-Xe with CL boundary condition.

C	case IV	case V	case VI
0.00	2.699092	3.805644	2.694548
0.10	4.980593	5.930889	3.592355
0.20	6.918308	7.665855	4.242706
0.30	8.703196	9.208216	4.776130
0.40	1.043710(1)	1.066110(1)	5.253175
0.50	1.218908(1)	1.209149(1)	5.708463
0.60	1.401600(1)	1.355138(1)	6.165612
0.70	1.596817(1)	1.508366(1)	6.641839
0.80	1.806406(1)	1.669940(1)	7.139433
0.90	2.004926(1)	1.816252(1)	7.558874
0.94	2.041581(1)	1.835065(1)	7.560751
1.00	1.775163(1)	1.577430(1)	6.407679

TABLE 14
The temperature-jump coefficient ζ for the mixture He-Ar with CL boundary condition.

C	case VII	case VIII	case IX
0.00	3.819923	3.823209	4.763477
0.10	4.830295	4.049216	5.771355
0.20	5.902419	4.287304	6.829256
0.30	7.054171	4.539984	7.953144
0.40	8.305739	4.811649	1.031862(1)
0.50	9.680309	5.101938	1.047218(1)
0.60	1.120361(1)	5.418278	1.188326(1)
0.70	1.286814(1)	5.760482	1.348900(1)
0.80	1.476387(1)	6.118533	1.520249(1)
0.90	1.666506(1)	6.434606	1.690730(1)
0.94	1.731059(1)	6.503639	1.746042(1)
1.00	1.775163(1)	6.407678	1.775163(1)

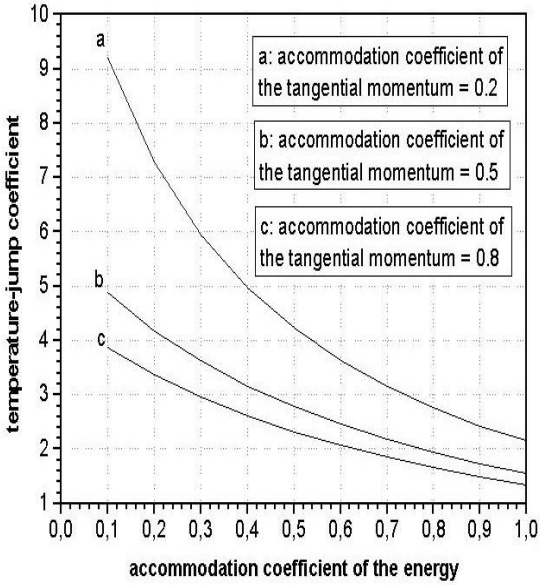


FIG. 1. *The temperature-jump coefficient: one gas case.*

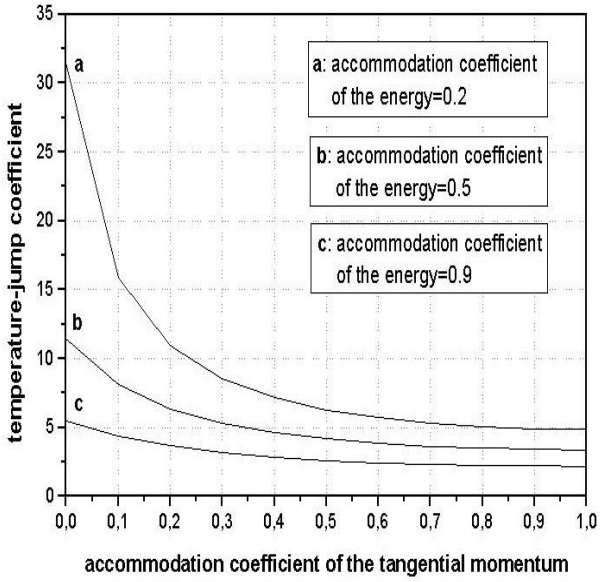


FIG. 2. *The temperature-jump coefficient: one gas case.*

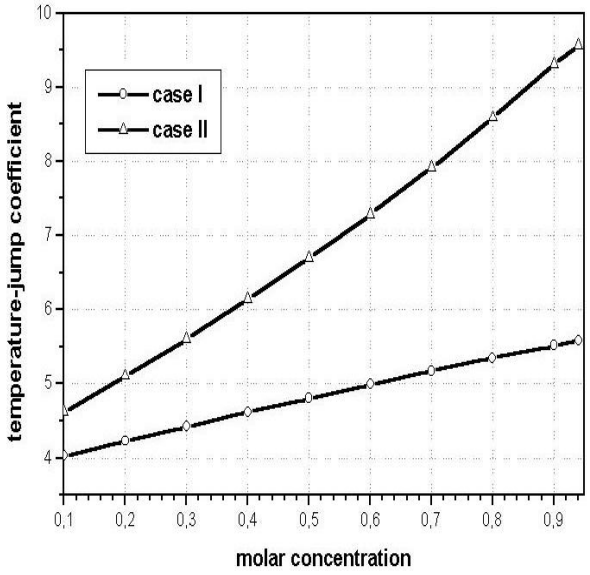


FIG. 3. *The temperature-jump coefficient: mixture Ne-Ar.*

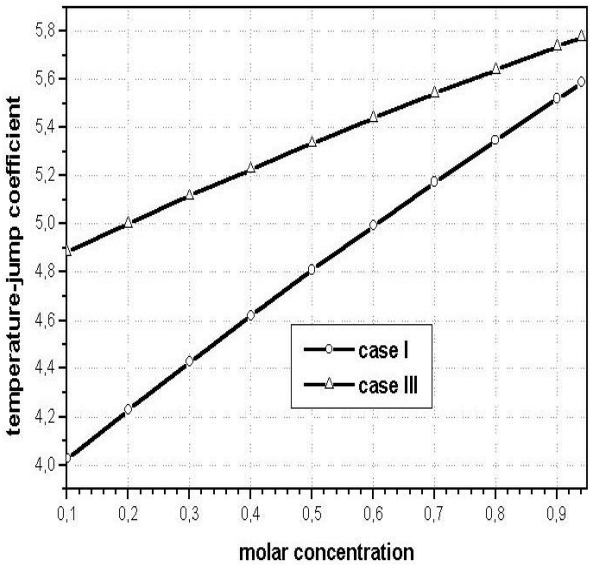


FIG. 4. *The temperature-jump coefficient: mixture Ne-Ar.*

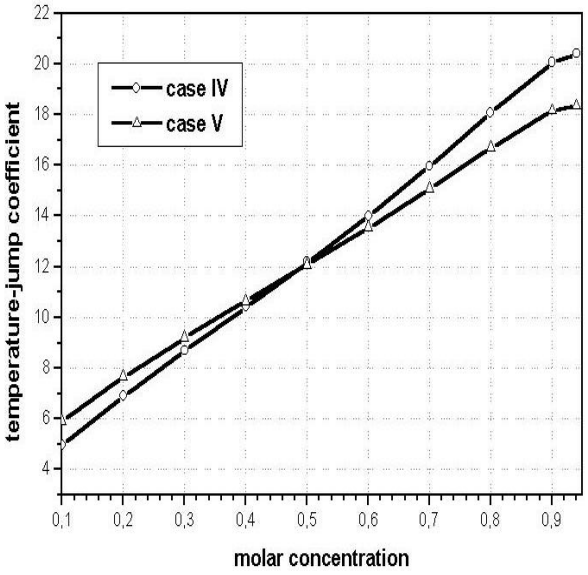


FIG. 5. *The temperature-jump coefficient: mixture He-Xe.*

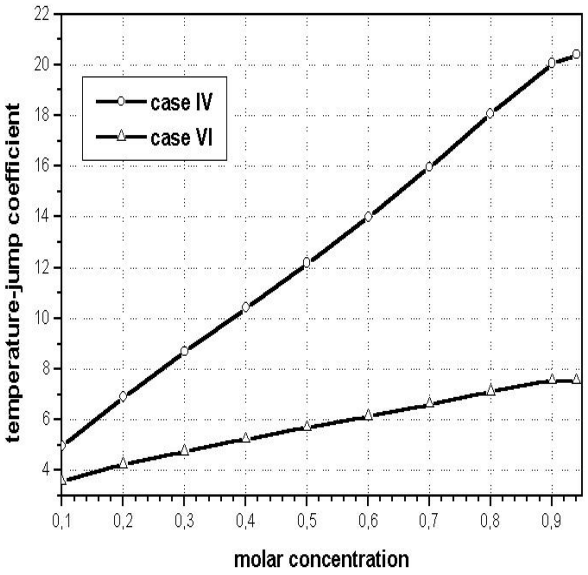


FIG. 6. *The temperature-jump coefficient: mixture He-Xe.*

9. Concluding comments. An analytical version of the discrete-ordinates method was used to develop a solution for the temperature-jump problem in rarefied gas dynamics, modeled by the S-model and the McCormack model, for the one gas and binary gas mixture cases, respectively. The gas-surface interaction was described by the Cercignani–Lampis boundary condition. The dependence of jump coefficient on both accommodation coefficients was shown, which points out the importance of special attention to the description of the gas-surface interaction. The evaluation of the jump coefficient for a binary mixture of gases shows more significant difference from the one gas case for a great mass ratio. Once more the discrete-ordinates solution was found to be very precise.

Appendix A. Basic definitions.
In (82),

$$\boldsymbol{K}^{(1)}_{1,1}(\mathbf{c}',\mathbf{c}) = 1 + \{2[1-\eta^{(1)}_{1,2}] - \eta^{(2)}_{1,2}(c'^2 - 5/2)\}\mathbf{c}'\cdot\mathbf{c}, \tag{A.1}$$

$$\boldsymbol{K}^{(2)}_{1,1}(\mathbf{c}',\mathbf{c}) = (2/3)[1-2r^*\eta^{(1)}_{1,2}](c'^2 - 3/2)(c^2 - 3/2), \tag{A.2}$$

$$\boldsymbol{K}^{(3)}_{1,1}(\mathbf{c}',\mathbf{c}) = 2\varpi_1[(\mathbf{c}'\cdot\mathbf{c})^2 - (1/3)c'^2c^2], \tag{A.3}$$

$$\boldsymbol{K}^{(4)}_{1,1}(\mathbf{c}',\mathbf{c}) = [(4/5)\beta_1(c'^2 - 5/2) - \eta^{(2)}_{1,2}](c^2 - 5/2)\mathbf{c}'\cdot\mathbf{c}, \tag{A.4}$$

$$\boldsymbol{K}^{(1)}_{2,1}(\mathbf{c}',\mathbf{c}) = r\{2\eta^{(1)}_{1,2} + \eta^{(2)}_{1,2}[r^2(c'^2 - 5/2) + c^2 - 5/2]\}\mathbf{c}'\cdot\mathbf{c}, \tag{A.5}$$

$$\boldsymbol{K}^{(2)}_{2,1}(\mathbf{c}',\mathbf{c}) = (4/3)r^*\eta^{(1)}_{1,2}(c'^2 - 3/2)(c^2 - 3/2), \tag{A.6}$$

$$\boldsymbol{K}^{(3)}_{2,1}(\mathbf{c}',\mathbf{c}) = 2\eta^{(4)}_{1,2}[(\mathbf{c}'\cdot\mathbf{c})^2 - (1/3)c'^2c^2], \tag{A.7}$$

$$\boldsymbol{K}^{(4)}_{2,1}(\mathbf{c}',\mathbf{c}) = (4/5)\eta^{(6)}_{1,2}(c'^2 - 5/2)(c^2 - 5/2)\mathbf{c}'\cdot\mathbf{c}, \tag{A.8}$$

$$\boldsymbol{K}^{(1)}_{2,2}(\mathbf{c}',\mathbf{c}) = 1 + \{2[1-\eta^{(1)}_{2,1}] - \eta^{(2)}_{2,1}(c'^2 - 5/2)\}\mathbf{c}'\cdot\mathbf{c}, \tag{A.9}$$

$$\boldsymbol{K}^{(2)}_{2,2}(\mathbf{c}',\mathbf{c}) = (2/3)[1-2s^*\eta^{(1)}_{2,1}](c'^2 - 3/2)(c^2 - 3/2), \tag{A.10}$$

$$\boldsymbol{K}^{(3)}_{2,2}(\mathbf{c}',\mathbf{c}) = 2\varpi_2[(\mathbf{c}'\cdot\mathbf{c})^2 - (1/3)c'^2c^2], \tag{A.11}$$

$$\boldsymbol{K}^{(4)}_{2,2}(\mathbf{c}',\mathbf{c}) = [(4/5)\beta_2(c'^2 - 5/2) - \eta^{(2)}_{2,1}](c^2 - 5/2)\mathbf{c}'\cdot\mathbf{c}, \tag{A.12}$$

$$\boldsymbol{K}^{(1)}_{1,2}(\mathbf{c}',\mathbf{c}) = s\{2\eta^{(1)}_{2,1} + \eta^{(2)}_{2,1}[s^2(c'^2 - 5/2) + c^2 - 5/2]\}\mathbf{c}'\cdot\mathbf{c}, \tag{A.13}$$

$$\boldsymbol{K}_{1,2}^{(2)}(\mathbf{c}',\mathbf{c}) = (4/3)s^*\eta_{2,1}^{(1)}(c'^2 - 3/2)(c^2 - 3/2), \tag{A.14}$$

$$\boldsymbol{K}_{1,2}^{(3)}(\mathbf{c}',\mathbf{c}) = 2\eta_{2,1}^{(4)}[(\mathbf{c}'\cdot\mathbf{c})^2 - (1/3)c'^2c^2] \tag{A.15}$$

and

$$\boldsymbol{K}_{1,2}^{(4)}(\mathbf{c}',\mathbf{c}) = (4/5)\eta_{2,1}^{(6)}(c'^2 - 5/2)(c^2 - 5/2)\mathbf{c}'\cdot\mathbf{c}. \tag{A.16}$$

In the definitions above it is used that

$$r = (m_1/m_2)^{1/2} \quad \text{and} \quad s = (m_2/m_1)^{1/2} \tag{A.17}$$

and

$$r^* = \frac{r^2}{1+r^2} \quad \text{and} \quad s^* = \frac{s^2}{1+s^2}. \tag{A.18}$$

In addition,

$$\varpi_1 = 1 + \eta_{1,1}^{(4)} - \eta_{1,1}^{(3)} - \eta_{1,2}^{(3)}, \tag{A.19}$$

$$\varpi_2 = 1 + \eta_{2,2}^{(4)} - \eta_{2,2}^{(3)} - \eta_{2,1}^{(3)}, \tag{A.20}$$

$$\beta_1 = 1 + \eta_{1,1}^{(6)} - \eta_{1,1}^{(5)} - \eta_{1,2}^{(5)}, \tag{A.21}$$

$$\beta_2 = 1 + \eta_{2,2}^{(6)} - \eta_{2,2}^{(5)} - \eta_{2,1}^{(5)}, \tag{A.22}$$

with

$$\eta_{i,j}^{(k)} = \frac{\nu_{i,j}^{(k)}}{\gamma_i}. \tag{A.23}$$

Continuing, following McCormack and Siewert's works [16, 32], we write

$$\nu_{\alpha,\beta}^{(1)} = \frac{16}{3}\frac{m_{\alpha,\beta}}{m_\alpha}n_\beta\Omega_{\alpha,\beta}^{11}, \tag{A.24}$$

$$\nu_{\alpha,\beta}^{(2)} = \frac{64}{15}\Big(\frac{m_{\alpha,\beta}}{m_\alpha}\Big)^2 n_\beta\Big(\Omega_{\alpha,\beta}^{12} - \frac{5}{2}\Omega_{\alpha,\beta}^{11}\Big), \tag{A.25}$$

$$\nu_{\alpha,\beta}^{(3)} = \frac{16}{5}\Big(\frac{m_{\alpha,\beta}}{m_\alpha}\Big)^2\frac{m_\alpha}{m_\beta}n_\beta\Big(\frac{10}{3}\Omega_{\alpha,\beta}^{11} + \frac{m_\beta}{m_\alpha}\Omega_{\alpha,\beta}^{22}\Big), \tag{A.26}$$

$$\nu_{\alpha,\beta}^{(4)} = \frac{16}{5}\Big(\frac{m_{\alpha,\beta}}{m_\alpha}\Big)^2\frac{m_\alpha}{m_\beta}n_\beta\Big(\frac{10}{3}\Omega_{\alpha,\beta}^{11} - \Omega_{\alpha,\beta}^{22}\Big), \tag{A.27}$$

$$\nu^{(5)}_{\alpha,\beta} = \frac{64}{15}\Big(\frac{m_{\alpha,\beta}}{m_\alpha}\Big)^3 \frac{m_\alpha}{m_\beta} n_\beta \Big[\Omega^{22}_{\alpha,\beta} + \Big(\frac{15m_\alpha}{4m_\beta} + \frac{25m_\beta}{8m_\alpha}\Big)\Omega^{11}_{\alpha,\beta} - \Big(\frac{m_\beta}{2m_\alpha}\Big)\Big(5\Omega^{12}_{\alpha,\beta} - \Omega^{13}_{\alpha,\beta}\Big)\Big] \tag{A.28}$$

and

$$\nu^{(6)}_{\alpha,\beta} = \frac{64}{15}\Big(\frac{m_{\alpha,\beta}}{m_\alpha}\Big)^3 \Big(\frac{m_\alpha}{m_\beta}\Big)^{3/2} n_\beta \Big[-\Omega^{22}_{\alpha,\beta} + \frac{55}{8}\Omega^{11}_{\alpha,\beta} - \frac{5}{2}\Omega^{12}_{\alpha,\beta} + \frac{1}{2}\Omega^{13}_{\alpha,\beta}\Big]. \tag{A.29}$$

Here

$$m_{\alpha,\beta} = \frac{m_\alpha m_\beta}{m_\alpha + m_\beta}. \tag{A.30}$$

Finally, the Ω functions are the Chapman–Cowling integrals [49, 50]

$$\Omega^{ij}_{\alpha,\beta} = \frac{(j+1)!}{8}\Big[1 - \frac{1+(-1)^i}{2(i+1)}\Big]\Big(\frac{\pi kT}{2m_{\alpha,\beta}}\Big)^{1/2}(d_\alpha + d_\beta)^2. \tag{A.31}$$

We note that here d_1 and d_2 are the diameters of the two types of particles and, as defined in the text, k is the Boltzmann constant and T_0 is a reference temperature.

Appendix B. The elements of the kernel.

We follow Siewert [32] and express the components of the matrix $\boldsymbol{K}_M(\xi',\xi)$ in (99) as

$$k_{1,1}(\xi',\xi) = 1 + f_{1,1}(\xi',\xi)\xi'\xi + (2/3)[1 - 2r^*\eta^{(1)}_{1,2} + 2\varpi_1](\xi'^2 - 1/2)(\xi^2 - 1/2), \tag{B.1}$$

$$k_{1,2}(\xi',\xi) = [(4/5)\beta_1(\xi^2 - 3/2) - \eta^{(2)}_{1,2}]\xi'\xi + (2/3)[1 - 2r^*\eta^{(1)}_{1,2} - \varpi_1](\xi^2 - 1/2), \tag{B.2}$$

$$k_{1,3}(\xi',\xi) = f_{1,3}(\xi',\xi)\xi'\xi + (4/3)[r^*\eta^{(1)}_{1,2} + \eta^{(4)}_{1,2}](\xi'^2 - 1/2)(\xi^2 - 1/2), \tag{B.3}$$

$$k_{1,4}(\xi',\xi) = [r^3\eta^{(2)}_{1,2} + (4/5)\eta^{(6)}_{1,2}(\xi^2 - 3/2)]\xi'\xi + (2/3)[2r^*\eta^{(1)}_{1,2} - \eta^{(4)}_{1,2}](\xi^2 - 1/2), \tag{B.4}$$

$$k_{2,1}(\xi',\xi) = [(4/5)\beta_1(\xi'^2 - 3/2) - \eta^{(2)}_{1,2}]\xi'\xi + (2/3)[1 - 2r^*\eta^{(1)}_{1,2} - \varpi_1](\xi'^2 - 1/2), \tag{B.5}$$

$$k_{2,2}(\xi',\xi) = (2/3)[1 - 2r^*\eta^{(1)}_{1,2}] + (1/3)\varpi_1 + (4/5)\beta_1\xi'\xi, \tag{B.6}$$

$$k_{2,3}(\xi',\xi) = [r\eta^{(2)}_{1,2} + (4/5)\eta^{(6)}_{1,2}(\xi'^2 - 3/2)]\xi'\xi + (2/3)[2r^*\eta^{(1)}_{1,2} - \eta^{(4)}_{1,2}](\xi'^2 - 1/2), \tag{B.7}$$

$$k_{2,4}(\xi',\xi) = (4/5)\eta_{1,2}^{(6)}\xi'\xi + (1/3)[4r^*\eta_{1,2}^{(1)} + \eta_{1,2}^{(4)}], \tag{B.8}$$

$$k_{3,1}(\xi',\xi) = f_{3,1}(\xi',\xi)\xi'\xi + (4/3)[s^*\eta_{2,1}^{(1)} + \eta_{2,1}^{(4)}](\xi'^2 - 1/2)(\xi^2 - 1/2), \tag{B.9}$$

$$k_{3,2}(\xi',\xi) = [s^3\eta_{2,1}^{(2)} + (4/5)\eta_{2,1}^{(6)}(\xi^2 - 3/2)]\xi'\xi + (2/3)[2s^*\eta_{2,1}^{(1)} - \eta_{2,1}^{(4)}](\xi^2 - 1/2), \tag{B.10}$$

$$k_{3,3}(\xi',\xi) = 1 + f_{3,3}(\xi',\xi)\xi'\xi + (2/3)[1 - 2s^*\eta_{2,1}^{(1)} + 2\varpi_2](\xi'^2 - 1/2)(\xi^2 - 1/2), \tag{B.11}$$

$$k_{3,4}(\xi',\xi) = [(4/5)\beta_2(\xi^2 - 3/2) - \eta_{2,1}^{(2)}]\xi'\xi + (2/3)[1 - 2s^*\eta_{2,1}^{(1)} - \varpi_2](\xi^2 - 1/2), \tag{B.12}$$

$$k_{4,1}(\xi',\xi) = [s\eta_{2,1}^{(2)} + (4/5)\eta_{2,1}^{(6)}(\xi'^2 - 3/2)]\xi'\xi + (2/3)[2s^*\eta_{2,1}^{(1)} - \eta_{2,1}^{(4)}](\xi'^2 - 1/2), \tag{B.13}$$

$$k_{4,2}(\xi',\xi) = (4/5)\eta_{2,1}^{(6)}\xi'\xi + (1/3)[4s^*\eta_{2,1}^{(1)} + \eta_{2,1}^{(4)}], \tag{B.14}$$

$$k_{4,3}(\xi',\xi) = [(4/5)\beta_2(\xi'^2 - 3/2) - \eta_{2,1}^{(2)}]\xi'\xi + (2/3)[1 - 2s^*\eta_{2,1}^{(1)} - \varpi_2](\xi'^2 - 1/2), \tag{B.15}$$

and

$$k_{4,4}(\xi',\xi) = (2/3)[1 - 2s^*\eta_{2,1}^{(1)}] + (1/3)\varpi_2 + (4/5)\beta_2\xi'\xi \tag{B.16}$$

with

$$f_{1,1}(\xi',\xi) = 2[1 - \eta_{1,2}^{(1)}] - \eta_{1,2}^{(2)}(\xi'^2 + \xi^2 - 3) + (4/5)\beta_1(\xi'^2 - 3/2)(\xi^2 - 3/2), \tag{B.17}$$

$$f_{3,3}(\xi',\xi) = 2[1 - \eta_{2,1}^{(1)}] - \eta_{2,1}^{(2)}(\xi'^2 + \xi^2 - 3) + (4/5)\beta_2(\xi'^2 - 3/2)(\xi^2 - 3/2), \tag{B.18}$$

$$\begin{aligned} f_{1,3}(\xi',\xi) = {} & 2r\eta_{1,2}^{(1)} + r\eta_{1,2}^{(2)}[r^2(\xi'^2 - 3/2) + \xi^2 - 3/2] \\ & + (4/5)\eta_{1,2}^{(6)}(\xi'^2 - 3/2)(\xi^2 - 3/2) \end{aligned} \tag{B.19}$$

$$\begin{aligned} f_{3,1}(\xi',\xi) = {} & 2s\eta_{2,1}^{(1)} + s\eta_{2,1}^{(2)}[s^2(\xi'^2 - 3/2) + \xi^2 - 3/2] \\ & + (4/5)\eta_{2,1}^{(6)}(\xi'^2 - 3/2)(\xi^2 - 3/2). \end{aligned} \tag{B.20}$$

Acknowledgments. The authors would like to thank K. Aoki, R. D. M. Garcia, M. Lampis, and V. Titarev for a special contribution in providing us some important references used in this work. We also thank to R. D. M. Garcia, C. E. Siewert, and F. Sharipov for helpful discussions in regard to the subject of this work.

REFERENCES

[1] C. Ho and Y. Tai, *Micro-electro-mechanical systems (MEMS) and fluid flows*, Annu. Rev. Fluid Mech., 30 (1998), pp. 579–612.
[2] M. Gad-El-Hak, *The MEMS Handbook*, 2nd ed., CRC Press, Boca Raton, 2006.
[3] G. Karniadakis, A. Beskok, and N. Aluru, *Microflows and Nanoflows*, Springer-Verlag, New York, 2005.
[4] C. Cercignani, *Mathematical Methods in Kinetic Theory*, Plenum Press, New York, 1969.
[5] C. Cercignani, *Rarefied Gas Dynamics From Basic Concepts to Actual Calculations*, Cambridge University Press, Cambridge, 2000.
[6] Y. Sone, *Kinetic Theory and Fluid Dynamics*, Birkhäuser, Boston, 2002.
[7] C. Cercignani, *The Boltzmann Equation and Its Applications*, Springer-Verlag, New York, 1988.
[8] P. Welander, *On the temperature jump in a rarefied gas*, Arkiv Fysik, 7 (1954), pp. 507–553.
[9] M. M. R. Williams, *Mathematical Methods in Particle Transport Theory*, Butterworth, London, 1971.
[10] M. M. R. Williams, *A review of the rarefied gas dynamics theory associated with some classical problems in flow and heat transfer*, Z. Angew. Math. Phys., 52 (2001), pp. 500–516.
[11] F. Sharipov, *Data on the velocity slip and temperature jump coefficients*, in Proceedings of the 5th Annual International Conference on Thermal and Mechanical Simulation Experiments in Micro-Electronics and Micro-Systems, Brussels, Belgium, 2004, pp. 243–249.
[12] C. Cercignani and M. Lampis, *Kinetic models for gas-surface interaction*, Transp. Theory Statist. Phys., 1 (1971), pp. 101–114.
[13] T. Klinc and I. Kuščer, *Slip coefficients for general gas-surface interaction*, Phys. Fluids, 15 (1972), pp. 1018–1022.
[14] R. Lord, *Some extensions of the Cercignani–Lampis gas-surface scattering kernel*, Phys. Fluids A, 3 (1991), pp. 706–710.
[15] E. M. Shakhov, *Generalization of the Krook kinetic relaxation equation*, Fluid Dynam., 3 (1968), pp. 142–145 (in Russian).
[16] F. J. McCormack, *Construction of linearized kinetic models for gaseous mixtures and molecular gases*, Phys. Fluids, 16 (1973), pp. 2095–2105.
[17] L. B. Barichello and C. E. Siewert, *A discrete-ordinates solution for a nongrey model with complete frequency redistribution*, JQSRT, 62 (1999), pp. 665–675.
[18] Y. Onishi, *Effects of accommodation coefficient on temperature and density fields in a slightly rarefied gas*, Trans. Japan Soc. Aero. Space Sci. 17, (1974), pp. 151–159.
[19] S. K. Loyalka, C. E. Siewert, and J. R. Thomas, Jr., *Temperature-jump problem with arbitrary accommodation*, Phys. Fluids, 21 (1978), pp. 854–855.
[20] S. K. Loyalka, *Temperature jump and thermal creep slip: Rigid sphere gas*, Phys. Fluids A, 1 (1989), pp. 403–408.
[21] Y. Sone, T. Ohwada, and K. Aoki, *Temperature jump and Knudsen layer in rarefied gas over a plane wall: Numerical analysis of the linearized Boltzmann equation for hard-sphere molecules*, Phys. Fluids A, 1 (1989), pp. 363–370.
[22] C. Cercignani and M. Lampis, *Variational calculation of the temperature jump for a binary mixture*, in Rarefied Gas Dynamics, A. E. Beylich, ed., VCH, Weinheim, Germany, 1991, pp. 1379–1384.
[23] S. K. Loyalka, *Temperature jump: Rigid-sphere gas with arbitrary gas/surface interaction*, Nucl. Sci. Eng., 108 (1991), pp. 69–73.
[24] F. Sharipov, *Application of the Cercignani-Lampis scattering kernel to calculations of rarefied gas flows.* II. *Slip and jump coefficients*, Eur. J. Mech. B Fluids, 22 (2003), pp. 133–143.
[25] L. B. Barichello, M. Camargo, P. Rodrigues, and C. E. Siewert, *Unified solutions to classical flow problems based on the BGK model*, Z. Angew. Math. Phys., 52 (2001), pp. 517–534.
[26] C. E. Siewert, *Poiseuille, thermal creep and Couette flow: Results based on the CES model of the linearized Boltzmann equation*, Eur. J. Mech. B Fluids, 21 (2002), pp. 579–597.
[27] M. Camargo and L. B. Barichello, *Unified approach for variable collision frequency models*

in rarefied gas dynamics, Transport Theory Statist. Phys., 33 (2004), pp. 227–260.

[28] L. C. Cabrera and L. B. Barichello, *Unified solutions to some classical problems in rarefied gas dynamics based on the one-dimensional linearized S-model equations*, Z. Angew. Math. Phys., 57 (2006), pp. 285–312.

[29] L. B. Barichello and C. E. Siewert, *The temperature-jump problem in rarefied gas dynamics*, European J. Appl. Math., 11 (2000), pp. 353–364.

[30] L. B. Barichello, A. C. R. Bartz, M. Camargo, and C. E. Siewert, *The temperature-jump problem for a variable collision frequency model*, Phys. Fluids, 14 (2002), pp. 382–391.

[31] C. E. Siewert, *The linearized Boltzmann equation: A concise and accurate solution of the temperature-jump problem*, JQSRT, 77 (2003), pp. 417–432.

[32] C. E. Siewert, *The McCormack model for gas mixtures: The temperature-jump problem*, Z. Angew. Math. Phys., 56 (2005), pp. 273–292.

[33] C. E. Siewert, *Generalized boundary conditions for the S-model kinetic equations basic to flow in a plane channel*, JQSRT, 72 (2002), pp. 75–88.

[34] R. F. Knackfuss and L. B. Barichello, *Surface effects in rarefied gas dynamics: An analysis based on the Cercignani–Lampis boundary condition*, Eur. J. Mech. B Fluids, 25 (2006), pp. 113–129.

[35] L. B. Barichello and C. E. Siewert, *Some comments on modeling the linearized Boltzmann equation*, JQSRT, 77 (2003), pp. 43–59.

[36] C. L. Pekeris and Z. Alterman, *Solution of the Boltzmann-Hilbert integral equation* II. *The coefficients of viscosity and heat conduction*, Proc. Nat. Acad. Sci. U.S.A., 43 (1957), pp. 998–1007.

[37] P. L. Bhatnagar, E. P. Gross, and M. Krook, *A model for collision processes in gases.* I. *Small amplitude processes in charged and neutral one-component systems*, Phys. Rev., 94 (1954), pp. 511–525.

[38] L. B. Barichello and C. E. Siewert, *A discrete-ordinates solution for Poiseuille flow in a plane channel*, Z. Angew. Math. Phys., 50 (1999), pp. 972–981.

[39] C. S. Scherer, *Kinetic Models of the Linearized Boltzmann Equation and a Heat-Transfer Problem in Microscale*, MSc. Dissertation, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil, 2005 (in Portuguese).

[40] Y. Onishi, *Kinetic theory analysis for temperature and density fields of a slightly rarefied binary gas mixture over a solid wall*, Phys. Fluids, 9 (1997), pp. 226–238.

[41] F. Sharipov and D. Kalempa, *Velocity slip and temperature jump coefficients for gaseous mixtures.* I. *Viscous slip coefficient*, Phys. Fluids, 15 (2003), pp. 1800–1806.

[42] F. Sharipov and D. Kalempa, *Velocity slip and temperature jump coefficients for gaseous mixtures.* IV. *Temperature jump coefficient*, Int. J. Heat and Mass Transfer, 48 (2005), pp. 1076–1083.

[43] R. G. Lord, *Tangential momentum accommodation coefficients of rare gases on polycrystalline metal surfaces*, in Rarefied Gas Dynamics, Progr. Astronautics and Aeronautics 51, Part I, J. L. Potter, ed., AIAA, New York, 1977, pp. 531–538.

[44] B. T. Porodnov, P. E. Suetin, S. F. Borisov, and V. D. Akinshin, *Experimental investigation of rarefied gas flow in different channels*, J. Fluid Mech., 64 (1974), pp. 417–437.

[45] L. B. Thomas, *A collection of some controlled surface thermal accommodation coefficient measurements*, in Rarefied Gas Dynamics, C. L. Brundin, ed., Academic Press, New York, 1967, pp. 155–162.

[46] L. B. Thomas and R. G. Lord, *Comparative measurements of tangential momentum and thermal accommodation on polished and on roughened steel spheres*, in Rarefied Gas Dynamics, K. Karamcheti, ed., Academic Press, New York, 1974, pp. 405–412.

[47] N. J. McCormick, *Gas-surface accommodation coefficients from viscous slip and temperature jump coefficients*, Phys. Fluids, 17 (2005), 107104.

[48] C. E. Siewert, *Viscous-slip, thermal-slip, and temperature-jump coefficients as defined by linearized Boltzmann equation and the Cercignani–Lampis boundary condition*, Phys. Fluids, 15 (2003), pp. 1696–1701.

[49] S. Chapman and T. G. Cowling, *The Mathematical Theory of Non-Uniform Gases*, Cambridge University Press, Cambridge, 1970.

[50] J. H. Ferziger and H. G. Kaper, *Mathematical Theory of Transport Processes in Gases*, North–Holland, Amsterdam, 1972.

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 2187–2203

AN ASYMPTOTIC ONE-DIMENSIONAL MODEL TO DESCRIBE FIRES IN TUNNELS III: THE TRANSIENT PROBLEM*

INGENUIN GASSER† AND HERBERT STEINRÜCK‡

Abstract. This is the third in a series of papers devoted to a model describing tunnel fire events (see [I. Gasser and J. Struckmeier, *Math. Methods Appl. Sci.*, 25 (2002), pp. 1231–1249] and [I. Gasser, *Math. Methods Appl. Sci.*, 26 (2003), pp. 1327–1347]). Here the transient behavior of the model is studied. In particular the stability of stationary solutions is analyzed. A good understanding of related bifurcation phenomena is obtained (at least numerically). This accelerates progress in the study of the stability of flow conditions in tunnels during a tunnel fire.

Key words. gasdynamics, low-Mach-number, fire in tunnels, stability

AMS subject classifications. 76D05, 76N15, 35Q30, 80A20

DOI. 10.1137/050624480

1. Introduction.

Due to some serious fire accidents in the recent past, tunnel fires have become an interesting topic not only for CFD engineers. In the last few decades various mathematical models based on a gas-dynamic description of the air in the tunnel have been proposed. A good overview of the modeling approaches is given in [BJL, P]. A recent survey of computer codes for tunnel fire simulations can be found in [OC] (and [F]). In this context the internet platform [FiT] should be mentioned also.

From a modeling point of view there are mainly three difficult issues:

- very low Mach numbers,
- large temperature differences such that significant heat transport takes place,
- the turbulent character of the flow.

In particular, the combination of the first two points induces a serious problem. The second point excludes a pure incompressible description (with the standard Boussinesq approximation for taking buoyancy effects into account). On the other hand, there is the known problem of fully compressible approaches in the low-Mach-number regime. Although these facts are known (even in the tunnel-fire literature; see [BJL]), they are often ignored due to the lack of alternatives.

From an application point of view a detailed knowledge of the three-dimensional flow is in general not necessary. The main question arising from an application can often be answered by one-dimensional models. In addition, one-dimensional approaches have important advantages over higher-dimensional models:

- there are simple ways to describe turbulence,
- numerical simulations are in general not very expensive,
- optimization strategies can be applied and solved in reasonable time.

Here we consider such a one-dimensional model, which was introduced in [GStr]. To our knowledge this is one of the first models derived from underlying fluid dynamics

*Received by the editors February 16, 2005; accepted for publication (in revised form) July 5, 2006; published electronically October 24, 2006.
http://www.siam.org/journals/siap/66-6/62448.html

†Department Mathematik, Universität Hamburg, Bundesstraße 55, 20146 Hamburg, Germany (gasser@math.uni-hamburg.de). This author was partially supported by the *Berufsfeuerwehr Bozen* and by the *EU network Hyperbolic and Kinetic Equations HYKE* (contract HPRN-CT-2002-00282).

‡Institut für Strömungslehre und Wärmeübertragung, Technische Universität Wien, Resselgasse 3, A-1040 Wien, Austria (herbert.steinrueck@tuwien.ac.at).

equations. The single dimension refers to the longitudinal spatial extension. This model is derived in such a way that it allows us to combine both a good description in the low-Mach-number regime and significant heat transport. The results obtained show that the model seems to reflect the main features of tunnel fires. A good (at least qualitative) agreement with results from experiments has been obtained (see [GStr, G4]).

Stationary solutions of the model equation were discussed in [G1]. However, under some critical conditions the steady flow through the tunnel may become unstable. Then the transient behavior will become important. The goal of this paper is twofold. First we study the dynamics of the one-dimensional initial-boundary value problem. Second we determine stability conditions for the stationary solutions.

The paper is organized as follows. In section 2 we briefly present the model. Then in section 3 we state the existence theory for the transient problem, and in section 4 we show numerical examples of the dynamics. The stability analysis is presented in section 5.

2. The model. The model we are going to study here was derived in [GStr]. The purpose of the model is to describe fire events in a long tunnel. The starting points are the mass, momentum, and energy balances for a flow of compressible gas (air or air-smoke mixture) in the tunnel. Due to the small ratio of the diameter $\tilde{d}$ of the tunnel cross section and the length $\tilde{L}$ of the tunnel, a one-dimensional description can be applied. In addition, the flow in the tunnel (even with fire) is a typical low-Mach-number flow. Therefore an asymptotic analysis with respect to low Mach numbers has been performed. Since in case of fire significant heat sources are involved the densities cannot be assumed to be constant, and the resulting flow is compressible.

However, in the low-Mach-number limit the flow has characteristic properties of an incompressible flow (infinite velocity of sound) and of a compressible flow (large density variations) as well.

We briefly outline the derivation of the model. Let ρ, u, p, T be the dimensionless density, velocity, pressure, and temperature of the flow in the tunnel, respectively. In order to distinguish dimensionless quantities from their dimensional counterparts, the latter are marked by $\tilde{}$. The thermodynamic coordinates are scaled by their corresponding values of the ambient air, $\tilde{p}_R$, $\tilde{\rho}_R$, $\tilde{T}_R$, respectively. An appropriate reference value for the velocities is given by

$$\tilde{u}_R = \frac{\tilde{Q}(\gamma - 1)}{\tilde{A}\tilde{p}_R}, \tag{1}$$

where $\tilde{Q}$ is the strength of all heat sources (fires) in the tunnel and $\tilde{A}$ is the cross section of the tunnel. The space coordinate x is naturally scaled with the length of the tunnel $\tilde{L}$, and the time t is scaled with $\tilde{L}/\tilde{u}_R$. The constant $\gamma = \tilde{c}_p/\tilde{c}_v$ is the ratio of the specific heat capacities for constant pressure and constant volume of air, respectively. Neglecting viscous stresses and heat conduction in the longitudinal direction, the continuity equation, equation of motion, and energy equation in dimensionless form read

$$\begin{aligned} \rho_t + (\rho u)_x &= 0, \\ u_t + u u_x + \left(\frac{1}{\gamma M^2}\right)\frac{1}{\rho} p_x &= -p_{dv}\frac{u|u|}{2} - f_d \sin\alpha, \\ (\rho T)_t + (u\rho T)_x + (\gamma - 1) p u_x &= q_f. \end{aligned} \tag{2}$$

These have to be supplemented by the equation of state for an ideal gas, $p = \rho T$. In the equation of motion we have considered the pressure loss $p_{dv}u|u|/2$ and the buoyancy force $f_d \sin\alpha$. The scaled pressure loss coefficient $p_{dv} = \tilde{L}/\tilde{d}\,\xi(Re)$, where $\xi(Re)$ is the pressure loss coefficient of a fully developed pipe flow and is a function of the Reynolds number and the roughness of the pipe wall only. The buoyancy parameter f_d is defined by $f_d = \tilde{L}\tilde{g}/\tilde{u}_R^2$, where $\tilde{g}$ is the gravity acceleration. The Mach number $M = \tilde{u}_R/\tilde{c}$ is the ratio of the reference velocity $\tilde{u}_R$ to the speed of sound of the reference state $\tilde{c}^2 = \gamma\tilde{p}_R/\tilde{\rho}_R$.

Typical values corresponding to the above scaling are

$$\gamma = \frac{7}{5}, \quad \tilde{c} = 341\,\mathrm{ms}^{-1}, \quad M^2 = 8.6 \cdot 10^{-6}, \quad Re = 6.7 \cdot 10^5, \quad Pr = 0.72. \tag{3}$$

Due to $M \approx 3 \cdot 10^{-3}$ we are in the low-Mach-number regime. It is important to note that the calculated Reynolds number indicates that the flow in the tunnel is turbulent.

The low Mach number is the reason for the above-mentioned difficulties in the numerical treatment of (2). However, in the present application the Mach number is always small (in time and in space). Therefore an expansion with respect to a low Mach number (setting $\varepsilon = \gamma M^2$),

$$p = p_0 + \varepsilon p_1 + \mathcal{O}(\varepsilon^2), \tag{4}$$

is reasonable. This leads to $p_0 = p_0(t)$. Since the tunnel is an open region, the leading order pressure (which corresponds to the mean outside pressure) will not change in time. Therefore $p_0 = const.$ Obviously we have $p_0 = 1$ (such that the unscaled leading order pressure is equal to p_r). Then in leading order we have $T = \frac{1}{\rho}$.

The leading order equations (in ϵ) are given by

$$\rho_t + u\rho_x = -\rho q, \tag{5}$$

$$u_t + uu_x + \frac{1}{\rho}p_x = -p_{dv}\frac{u|u|}{2} - f_d \sin\alpha, \tag{6}$$

$$u_x = q, \tag{7}$$

with $q = q(x,t)$ as time and space dependent (scaled) heat source. For more details on the derivation of the model and scaling, see [GStr, G4].

As far as boundary data is concerned we prescribe Dirichlet data for the pressure p at the entrance and the exit

$$p(t,0) = p_l(t), \qquad p(t,1) = p_r(t) \quad \forall t > 0. \tag{8}$$

Moreover, assuming no fire at the entrance and exit, we have $q = 0$ at the boundaries, and therefore homogenous Neumann conditions for the velocity,

$$u_x(t,0) = u_x(t,1) = 0 \quad \forall t > 0, \tag{9}$$

hold. For the density we assume standard inflow boundary conditions

$$\rho(t,0) = \rho_l(t) \text{ if } u(t,0) > 0, \qquad \rho(t,1) = \rho_r(t) \text{ if } u(t,1) < 0 \quad \forall t > 0. \tag{10}$$

Initial data are prescribed for the density and the velocity:

$$u(0,x) = u_0(x), \qquad \rho(0,x) = \rho_0(x) \quad \forall x \in [0,1]. \tag{11}$$

Thus, our model consists of (5)–(6), the boundary conditions (8)–(10), and the initial conditions (11).

A first problem lies in the fact that we have two boundary conditions for the pressure but only a first derivative of the pressure in the model. There are at least two ways to resolve this problem. One possibility is to rewrite the model in the following way [GStr]. Equation (7) is substituted by the x-derived velocity equation (6),

$$\rho_t + u\rho_x = -\rho q, \tag{12}$$

$$\left(\frac{1}{\rho}p_x\right)_x = -p_{dv}q|u| - q_x u - q^2 - q_t - f_d \cos\alpha\, \alpha_x, \tag{13}$$

$$u_t + uq + \frac{1}{\rho}p_x = -p_{dv}\frac{u|u|}{2} - f_d \sin\alpha. \tag{14}$$

Thus we have an elliptic equation for the pressure which is consistent with two boundary conditions (for the pressure). This formulation has the advantage that it can be extended easily to higher dimensions [GST1, GST2, G4].

An alternative reformulation—restricted to one space dimension—is the following. We eliminate the pressure by multiplying (6) by ρ and integrating over $x \in [0,1]$. This gives

$$\int_0^1 \rho u_t dx + \int_0^1 \rho u u_x dx + p_r - p_l = -\int_0^1 p_{dv}\rho\frac{u|u|}{2}dx - \int_0^1 f_d \sin\alpha \rho dx. \tag{15}$$

Equation (7) gives

$$u(x,t) = v(t) + \int_0^x q(y,t)dy = v(t) + Q(x,t), \tag{16}$$

where $Q(x,t)$ is a known function. Then we obtain the system

$$\rho_t + (v+Q)\rho_x = -\rho q, \tag{17}$$

$$Rv_t + R_q v + \int_0^1 p_{dv}\rho\frac{(v+Q)|v+Q|}{2}dx = -R_{Q_t+Qq+f_d\sin\alpha} - p_r + p_l, \tag{18}$$

for ρ and v, where R, R_q, $R_{Q_t+Qq+f_d\sin\alpha}$ denote functionals applied to $\rho(x,t)$ defined by (for general f)

$$R(t) = \int_0^1 \rho(x,t)dx, \tag{19}$$

$$R_f(t) = \int_0^1 \rho(x,t)f(x,t)dx. \tag{20}$$

System (17)–(18) consists of an ordinary differential equation (ODE) for v and a partial differential equation (PDE) for the density ρ. The only boundary conditions needed are the inflow conditions (10) for the continuity equation. The conditions on the pressure appear as parameters in (18). The condition on the velocity (9) is automatically fulfilled by (16). In this paper we use this last formulation (17)–(18).

3. Global existence and uniqueness. In this section we focus on the global existence and uniqueness of solutions of (17)–(18) with boundary conditions (10) and initial conditions (11).

We formulate the following conditions:

(A1) On the data in the equations:
  (i) Let $q = q(x,t) \geq 0$ be a smooth function on $[0,1]\times[0,\infty)$ with $\operatorname{supp} q(.,t) \subset (0,1)\ \forall t \in [0,\infty)$.
  (ii) Let $p_{dv} = p_{dv}(x)$ be a smooth function on $[0,1]$ with $p_{dv}(x) > 0\ \forall x \in [0,1]$.
  (iii) Let $\alpha = \alpha(x)$ be a smooth function on $[0,1]$ with $-\frac{\pi}{2} < \alpha(x) < \frac{\pi}{2}$ $\forall x \in [0,1]$.

(A2) On the boundary data:
  (i) Let $\rho_l = \rho_l(t) > 0$ and $\rho_r = \rho_r(t) > 0$ be in $C^1[0,T]\ \forall T > 0$.
  (ii) Let $p_l = p_l(t)$ and $p_r = p_r(t)$ be smooth functions for $t \in [0,\infty)$.

(A3) On the initial data:
  (i) Let $u_0 = u_0(x) = v_0 + \int_0^x q(y,0)dy$ for some constant v_0.
  (ii) Let $\rho_0 = \rho_0(x) > 0$ be a piecewise continuous differentiable function for $x \in [0,1]$.

Before presenting the main existence result we state a result concerning lower and upper bounds of the solutions of the continuity equation (17). In the following we use the notation $f_{max}(T) = \max_{(x,t)\in[0,1]\times[0,T]} f(x,t)$.

LEMMA 3.1. *Let* (A1)(i) *hold. Let* $T > 0$. *Then for a given* $v \in C[0,T]$ *the continuity equation* (17) *with initial* ρ_0 *and boundary data* (10) *has a unique piecewise continuously differentiable solution* ρ *on* $(x,t) \in [0,1]\times[0,T]$. *It satisfies*

$$0 < \rho_{min}(T) \leq \rho(x,t) \leq \rho_{max}(T) \qquad \forall (x,t) \in [0,1]\times[0,T], \tag{21}$$

with

$$\rho_{min}(T) = \min_{t\in[0,T],\, x\in[0,1]} (\rho_l(t), \rho_r(t), \rho_0(x))\, e^{-q_{max}(T)}, \tag{22}$$

$$\rho_{max}(T) = \max_{t\in[0,T],\, x\in[0,1]} (\rho_l(t), \rho_r(t), \rho_0(x)). \tag{23}$$

We also present the proof since it involves the characteristics on which the main ideas of the existence analysis are based.

Proof. Consider a time interval $[0,T]$. The characteristic curves $\eta = \eta(t, x_0, t_0)$ for $t \in [0,T]$, $x, x_0 \in [0,1]$, are defined by

$$\frac{d\eta(t,x_0,t_0)}{dt} = u(\eta(t,x_0,t_0),t) = v(t) + Q(\eta(t,x_0,t_0),t), \qquad \begin{cases} \eta(0,x_0,0) &= x_0, \\ \eta(t_0,0,t_0) &= 0, \\ \eta(t_0,1,t_0) &= 1. \end{cases} \tag{24}$$

Given a continuous $v = v(t)$, the right-hand side in (24) is continuous (in η and t) and Lipschitz-continuous in η. The Lipschitz constant L satisfies $L \leq q_{max}(T)$. Therefore the characteristic curves are uniquely defined at every point (x,t) and do not intersect.

The continuity equation (17) reads along the characteristics

$$\frac{d\rho(\eta(t,x_0,t_0),t)}{dt} = -q(\eta(t,x_0,t_0),t)\rho(\eta(t,x_0,t_0),t), \tag{25}$$

with

$$\begin{aligned} \rho(\eta(0,x_0,0),0) &= \rho_0(x_0) && \text{for} && 0 < \eta(t,x_0,t_0) < 1 \quad \forall t \in [0,T], \\ \rho(0,t) &= \rho(\eta(t,0,t),t) = \rho_l(t) && \text{for} && u(0,t) > 0, \\ \rho(1,t) &= \rho(\eta(t,1,t),t) = \rho_r(t) && \text{for} && u(1,t) < 0. \end{aligned}$$

The density does not increase along the characteristics; it decreases at most exponentially in the region with nonvanishing heat-source q. This implies the estimates (21), (22), (23). □

Now we state the main global existence result.

THEOREM 3.1. *Let* (A1)–(A3) *hold. Then for all* $T > 0$ *there exists a unique solution* (ρ, v)*—*ρ *piecewise continuous differentiable in* $[0,1] \times [0,T]$ *and* $v \in C^1[0,T]$*—to* (17)–(18) *with boundary conditions* (10) *and initial conditions* (ρ_0, v_0) *from* (11).

The idea of the proof is a fixed point argument in the ODE (18). Clearly, in every step we have to control the coefficients of the ODE (18), which involve the solution of the PDE (17). Therefore, the proof becomes technical. Details on the proof are given in [GSte2].

4. The transient behavior. Here we would like to discuss the transient behavior of the solutions of our model with respect to the applications. In [G1] the stationary problem was analyzed. The result was that even in simple (realistic) cases (and even without fire) multiple (realistic) solutions of the stationary problem may exist. This depends mainly on the pressure difference at the entrance and exit and on the slope profile in the tunnel. In addition, numerical simulations indicate that in some cases more than one of these (multiple) solutions seem to be stable.

Numerical simulations in the transient problem for realistic tunnel examples have been performed in [G4, GStr] in which data from experiments have been included. Depending on the initial data, completely different long term behavior can be observed. Therefore, the dynamics of the transient problem is all but trivial. Clearly, this is of particular interest for the application.

Here we present an example to give an idea of the dynamics of the model (see Figures 1–3). The example starts with initial data corresponding to a no-fire situation in a tunnel with slope (see Table 1 for details). The pressure difference induces a steady downward flow (in Figure 1, downwards means to the left) in the tunnel. Then (in the time interval $t \in [10, 15]$ minutes) a fire of strength $Q = 6\,\mathrm{MW}$ is ignited in the middle of the tunnel. Buoyancy forces reduce the driving force of the flow, but they are too weak to compensate for the external pressure difference, and thus the flow is still downward. The flow adjusts rapidly to a new stationary state. Below (in Figure 1, to the left of) the heat source there is a hot air-smoke mixture; above the fire there is the fresh cold air. Later (in the time interval $t \in [60, 65]$ minutes) the heat production of the fire is increased (8 MW heat source) and the transition to the corresponding stationary state takes place (slowly). Now the buoyancy forces

TABLE 1
Data of tunnel and fire.

Length	4 km
Cross section	$100\,\mathrm{m}^2$
Slope	3%
Pressure loss coefficient ξ	0.0956
Pressure-difference $p(L) - p(0)$	1013.55 mbar
Pressure-difference due to the altitude	14.4 mbar
Mean initial velocity	$-5.0\,\mathrm{ms}^{-1}$
Heat source $q(x,t) = Q(t)\delta(x)$	$Q(t) = \begin{cases} 0\,\mathrm{MW} & t < 10\,\mathrm{min} \\ \text{linear cont.} & 10\,\mathrm{min} < t < 15\,\mathrm{min} \\ 6\,MW & 15\,\mathrm{min} < t < 60\,\mathrm{min} \\ \text{linear cont.} & 60\,\mathrm{min} < t < 65\,\mathrm{min} \\ 8\,MW & 65\,\mathrm{min} < t \end{cases}$

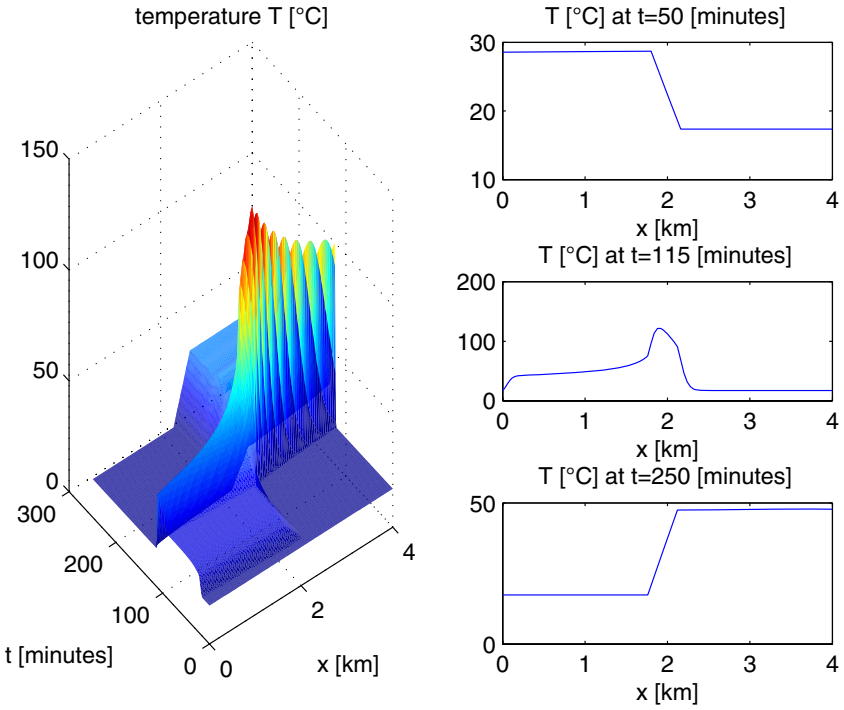


FIG. 1. *Temperature field (left) and the temperature at three time levels (right) in the transition between three different stationary states corresponding to the data in Table* 1.

acting on the hot air below the fire are strong enough to compensate for the outside pressure difference, and thus the flow velocity decreases and finally the flow changes its direction. When the flow velocity tends to zero the enthalpy produced by the fire is not removed from the fire, and thus the temperature at the fire rises dramatically, even enhancing the upward buoyancy forces. Then the air starts to flow upward, and buoyancy accelerates the flow. With the onset of the flow in the upward direction the fire is cooled again; however, the very hot air/smoke produced during the very slow flow phase is still in the tunnel. This hot gas induces a large upward buoyancy force, resulting in a positive velocity overshoot. From the left (lower) end, cold air is sucked in, and thus the temperature of the fire is cooled, resulting in a cooler hot air-smoke mixture above the fire, which in turn reduces the buoyancy and the upward velocity. In Figure 2 we can see the flow velocity. In the right of Figure 2 we observe the difference between the velocity at the entrance and the exit. This difference depends on the strength of the heat source and has its maximum value in the final state (8 MW heat source).

We want to point out the oscillating behavior that occurs as the fire is attaining the final state. The temperature distribution has a discontinuity due the inversion of the flow direction. This can be seen best in the contour plots of the temperature distribution (Figure 3). The flow changes direction shortly after the first particles of the air, which have been heated by the fire after its strength has been increased, reach the left end of the tunnel. After the inversion of the flow direction, cold air is sucked in. Therefore a contact discontinuity between the hot and the fresh cold air forms.

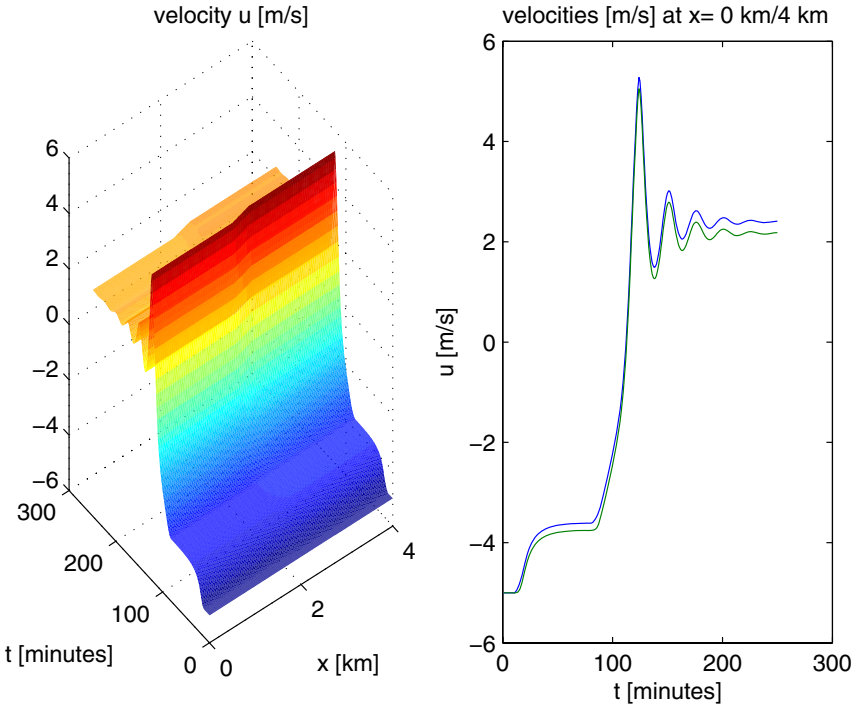


FIG. 2. *Velocity field (left) and the velocity at the entrance (upper curve) and exit (lower curve) of the tunnel as a function of time (right) in the transition between three stationary states corresponding to the data in Table* 1.

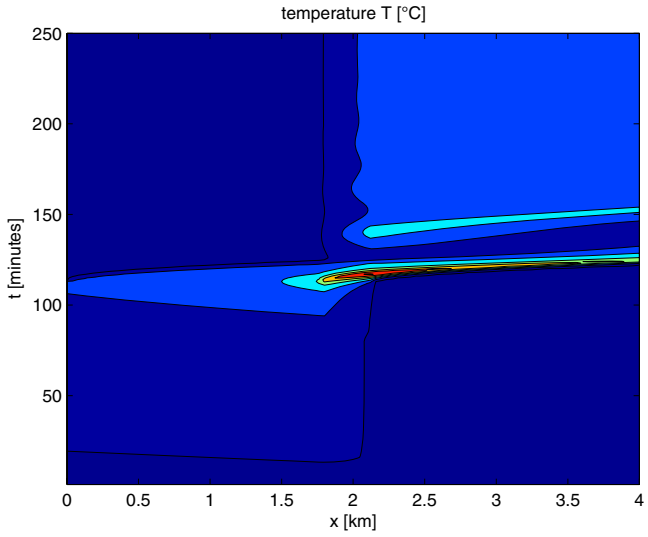


FIG. 3. *Contour plot of the temperature field in the transition between three stationary states corresponding to the data in Table* 1.

Obviously, the analysis in section 3 can be applied to such discontinuous solutions.

5. Stability of the stationary solutions. The starting point for the stability analysis is (17)–(18). The corresponding stationary problem was studied in [G1]. There the complete solution structure was analyzed. Depending on the data all the cases—no solution, exactly one solution, and multiple (nonvacuum) solutions—are possible. Especially from the application point of view it is important to know whether the solutions of the stationary problem are stable as solutions of the underlying transient problem. This is the main question in this section.

For this purpose we assume $q = q(x)$ and constant (in time) data ρ_l, ρ_r, p_l, p_r. Then we linearize (17)–(18) around a solution (ρ_s, v_s) of the stationary problem using

$$\rho(x,t) = \rho_s(x) + \rho_L(x,t), \qquad v(t) = v_s + v_L(t), \tag{26}$$

and obtain

$$\rho_{L;t} + (v_L + Q)\rho_{L;x} = -v_L\rho_{s;x} - q\rho_L, \tag{27}$$

$$R_s v_{L;t} + R_{s;q+p_{dv}(v_s+Q)} v_L = -R_{L;G}, \tag{28}$$

where we use the notation

$$R_{s;f}(t) = \int_0^1 \rho_s(x) f(x,t)dx, \qquad R_{L;f}(t) = \int_0^1 \rho_L(x,t) f(x,t)dx \tag{29}$$

and

$$G(x) = q(x)(v_s + Q(x)) + p_{dv}(x)\frac{(v_s + Q(x))|v_s + Q(x)|}{2} + f_d \sin\alpha(x). \tag{30}$$

The strategy is again to "solve" the continuity equation (27) and to evaluate the terms in (28) where the density is involved. Then one is left with an ODE for the velocity correction v_L only.

We now focus on the case of positive (constant) mass-flux $m = \rho_s(x)(v_s + Q(x))$. The case of negative mass-flux can be handled analogously. Remember that solutions with vanishing mass-flux imply (in case of nonvanishing heat-source) vacuum solutions which are not of interest in this application (see [G1]).

In case of a stationary solution the characteristic $\eta(t, x_0, t_0)$ of the continuity equation can be written in the form

$$\eta(t, x_0, t_0) = \eta^*(t - t_E(x_0, t_0)), \tag{31}$$

where $\eta^*(t) = \eta(t, 0, 0)$ and $t_E(x,t)$ is the time when the characteristic through x, t enters the interval $(0, 1)$.

Thus we have

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\eta^* = v_s + Q(\eta^*), \quad \eta^*(0) = 0, \tag{32}$$

$$t_E(x,t) := t - \int_0^x \frac{\mathrm{d}x'}{v_s + Q(x')} = t - t_l(x). \tag{33}$$

The time $t_l(x)$ is the time a particle needs to move from the entrance of the tunnel $(x = 0)$ to the position x. Along this characteristic curve the density evolves as

$$\rho_L(\eta^*(t - t_E), t) = -\int_{t_E}^t v_L(\tau)\rho_{s;x}(\eta^*(\tau - t_E)) \exp\left(-\int_\tau^t q(\eta^*(s - t_E))\,\mathrm{d}s\right)\mathrm{d}\tau \tag{34}$$

and

(35)
$$\rho_L(x,t) = -\int_{t_E(x,t)}^{t} v_L(\tau)\rho_{s;x}(\eta^*(\tau - t_E(x,t)))\exp\left(-\int_\tau^t q(\eta(s-t_E(t,x)))\,\mathrm{d}s\right)\mathrm{d}\tau.$$

Substituting

$$z = \eta^*(s - t + t_l(x)), \quad \mathrm{d}z = \eta^{*\prime}\,\mathrm{d}s = (v_s + Q(z))\,\mathrm{d}s \tag{36}$$

gives

(37)
$$\int_\tau^t q(\eta(s-t_E(t,x)))\,\mathrm{d}s = \int_{\eta^*(\tau-t+t_l(x))}^{\eta^*(t_l(x))=x} \frac{q(z)\mathrm{d}z}{v_s+Q(z)} = \ln\frac{v_s+Q(x)}{v_s+Q(\eta(\tau-t+t_l(x)))}$$

and

$$\begin{aligned}\rho_L(x,t) &= -\int_{t_0(x,t)}^{t} U_L(\tau)\rho_x(\eta(\tau-t_0(x,t)))\frac{v_s+Q(\eta(\tau-t_0(x,t)))}{v_s+Q(x)}\mathrm{d}\tau\\ &= -\frac{1}{v_s+Q(x)}\int_0^{t_l(x)} U_L(t-s)J(t_l(x)-s)\mathrm{d}s \qquad (38)\end{aligned}$$

with

$$J(t_l(x)-s) = \rho_x(\eta(t_l(x)-s))(v_s+Q(\eta(t_l(x)-s))). \tag{39}$$

Then we conclude

$$\begin{aligned}-\int_0^1 \rho_L(x,t)\,G(x)\,\mathrm{d}x &= \int_0^1\int_0^{t_l(x)} U(t-s)\frac{G(x)\,J(t_l(x)-s)}{v_s+Q(x)}\,\mathrm{d}s\,\mathrm{d}x\\ &= \int_0^{t_l(1)} U(t-s)\int_{\eta(s)}^1 \frac{G(x)\,J(t_l(x)-s)}{v_s+Q(x)}\,\mathrm{d}x\,\mathrm{d}s. \qquad (40)\end{aligned}$$

We summarize the resulting problem for v_L:

$$\frac{\mathrm{d}}{\mathrm{d}t}v_L = Av_L + \int_0^{t_l(1)} v_L(t-s)K(s)\mathrm{d}s, \tag{41}$$

$$A = -\frac{\int_0^1 \rho_s(q+\xi|v_s+Q|)\,\mathrm{d}x}{\int_0^1 \rho_s\,\mathrm{d}x} < 0, \tag{42}$$

$$K(s) = \frac{1}{\int_0^1 \rho_s\,\mathrm{d}x}\int_{\eta(s)}^1 \frac{G(x)}{v_s+Q(x)}J(t_l(x)-s)\,\mathrm{d}x. \tag{43}$$

This is a Volterra integro-differential equation. Before using standard results on such equations we remark the following.

Remark 5.1.

(i) We have

$$J(\tau) = (v_s+Q(\eta(\tau))\,\rho_{s;x}(\eta(\tau)) = -\rho_s(\eta(\tau)))\,q(\eta(\tau)) \le 0. \tag{44}$$

(ii) $G(x) > 0$ if $\alpha > 0$. In this case we have $K(s) \leq 0$.

Using Theorem 5.2.1 in [Bu], we can state the following result.

THEOREM 5.1. *If*

$$\int_0^{t_l} |K(\tau)| \, \mathrm{d}\tau < |A|, \tag{45}$$

then the trivial solution is asymptotically stable.

The proof goes along the lines of Theorem 5.2.1 in [Bu].

Remark 5.2. For appropriately "small" heat sources $||q||_1 < c$ the functions J and K are such that Theorem 5.1 holds.

Remark 5.3. For the case $\alpha \geq 0$ Theorem 5.1 does not take into account the fact that both A and K have the "good" negative sign. Therefore a stronger result can be expected.

5.1. Point-sources. For many applications (e.g., when the diameter of the region where the heat-source is active is small compared to the length of the tunnel) it might be useful to consider point-sources.

Here we discuss the case of a single point-source located at the position $x = x_q$, of the form

$$q(x,t) = q\delta(x - x_q). \tag{46}$$

Then we conclude

$$Q(x,t) = qH(x - x_q), \tag{47}$$

with H the Heavy-side function. We assume a time independent source q, since the aim is to study the stability of stationary solutions. Integrating the continuity equation over the support of $q(x,t)$, we obtain that the mass flux ρu is continuous at x_q:

$$\rho_s(x_{q-}, t)v = \rho_s(x_{q+}, t)(v + q) \qquad \text{at } x = x_q. \tag{48}$$

Then the continuity equation reads as

$$\begin{cases} \rho_t + v\rho_x = 0 & \text{for} \quad x < x_q, \\ \rho_t + (v+q)\rho_x = 0 & \text{for} \quad x > x_q. \end{cases} \tag{49}$$

Using the continuity of the mass flow at x_q, we can evaluate the integral

$$vR_q + R_{Qq} = \int_0^1 u\rho \, \mathrm{d}x = qv\rho(x_{q-})$$

and obtain the integrated momentum equation

$$Rv_t + q\rho_- v + \int_0^{x_q} p_{dv}\rho \frac{v|v|}{2} dx + \int_{x_q}^1 p_{dv}\rho \frac{(v+q)|v+q|}{2} dx = -R_{f_d \sin\alpha} - p_r + p_l. \tag{50}$$

Focusing again on the case of strictly positive mass fluxes, the stationary solution is given by

$$u_s = \begin{cases} v_s & \text{for} \quad x < x_q, \\ v_s + q & \text{for} \quad x > x_q, \end{cases} \tag{51}$$

and the stationary density results in

$$\rho_s = \begin{cases} \rho_{s-} &= \rho(0) & \text{for} \quad x < x_q, \\ \rho_{s+} &= \rho(0)\frac{v_s}{v_s+q} & \text{for} \quad x > x_q, \end{cases} \tag{52}$$

where v_s is a solution of the nonlinear equation

$$\begin{aligned} &q\rho_- v_s + p_{dv}\frac{\rho_- v_s}{2}\left(x_q|v_s| + (1-x_q)|v_s+q|\right) \\ &= \rho_- f_d\left(\int_0^{x_q}\sin(\alpha)\,\mathrm{d}x + \frac{v_s}{v_s+q}\int_{x_q}^1 \sin(\alpha)\,\mathrm{d}x\right) + p_l - p_r. \end{aligned} \tag{53}$$

In order to keep the calculations simple we consider the linearization for a constant p_{dv} and constant slope profile α. The solutions of the linearized equations are denoted by a subscript $_L$. The linearized continuity equation reads

$$\begin{cases} \rho_{L;t} + v_s\rho_{L;x} &= 0 & \text{for} \quad x < x_q, \\ v_L(t)\rho_{s-} + v_s\rho_{L,-} &= (v_s+q)\rho_{L+} + v_L\rho_{s+} & \text{for} \quad x = x_q, \\ \rho_{L;t} + (v_s+q)\rho_{L;x} &= 0 & \text{for} \quad x > x_q, \end{cases} \tag{54}$$

and the linearized equation (50) becomes

$$\begin{aligned} &(\rho_{s-}x_q + \rho_{s+}(1-x_q))v_{L;t} + (\rho_{s-}q + \rho_{s-}p_{dv}v_s x_q + \rho_{s+}p_{dv}(v_s+q)(1-x_q))v_L \\ &= -p_{dv}\frac{v_s^2}{2}\int_0^{x_q}\rho_L\mathrm{d}x - p_{dv}\frac{(v_s+q)^2}{2}\int_{x_q}^1\rho_L\mathrm{d}x - \int_{x_q}^1\rho_L f_d\sin\alpha\mathrm{d}x - \rho_{L-}qv_s. \end{aligned} \tag{55}$$

The solution of the continuity equation is given by (assuming vanishing density disturbances at the left boundary)

$$\rho_L = \begin{cases} 0 & & \text{for} \quad x < x_q, \\ \rho_L\left(x_q+, t-\frac{x-x_q}{v_s+q}\right) &= \frac{\rho_{s-}-\rho_{s+}}{v_s+q}v_L\left(t-\frac{x-x_q}{v_s+q}\right) & \text{for} \quad x > x_q, \end{cases} \tag{56}$$

for $t - \frac{x-x_q}{v_s+q} > 0$. Then with $\rho_{s-} - \rho_{s+} = \rho(0)\frac{q}{v_s+q}$ we obtain

$$\int_{x_q}^1 \rho_L dx = \rho(0)\frac{q}{v_s+q}\int_0^{\frac{1-x_q}{v_s+q}} v_L(t-s)ds. \tag{57}$$

Therefore (55) reads

$$v_{L;t} + av_L = b\int_0^{\frac{1-x_q}{v_s+q}} v_L(t-s)ds, \tag{58}$$

or with $t = \tau\frac{1-x_q}{v_s+q}$, $V(\tau) = v_l(\tau\frac{1-x_q}{v_s+q})$ we obtain

$$\begin{aligned} V_\tau + AV &= -B\int_0^1 V(\tau-\sigma)d\sigma, \\ A &= \frac{(v_s+q)(\rho_{s-}q + \rho_{s-}p_{dv}v_s x_q + \rho_{s+}p_{dv}(v_s+q)(1-x_q))}{(1-x_q)(\rho_{s-}x_q + \rho_{s+}(1-x_q))} > 0, \\ B &= \frac{(v_s+q)(p_{dv}\frac{(v_s+q)^2}{2} + f_d\sin\alpha)}{(1-x_q)(\rho_{s-}x_q + \rho_{s+}(1-x_q))}. \end{aligned} \tag{59}$$

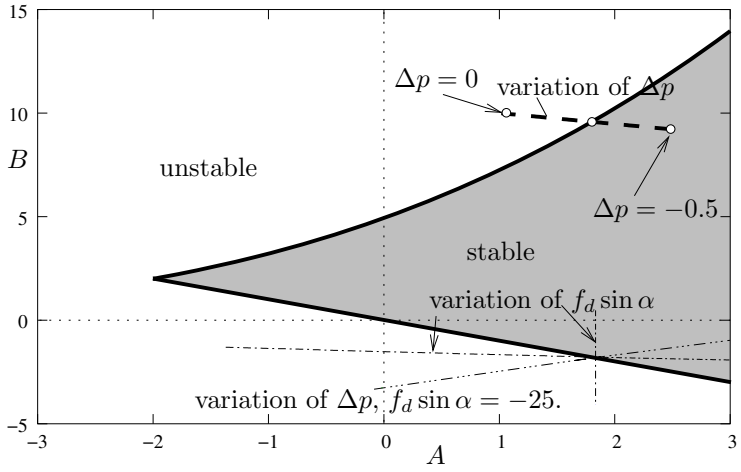


FIG. 4. *Stability diagram (detail). Parameter values A, B corresponding to the stationary solutions of the examples of section* 5.2 *are also indicated: variation of* Δp *for* $f_d = 0$*, dashed line; variation of* Δp *for* $f_d \sin\alpha = -25$*, dash-triple-dotted line; variation of* $f_d \sin\alpha$*, dash-dotted line.*

In order to find the stable and unstable parameter regions we make the ansatz $V(\tau) = e^{\zeta\tau}$ and obtain the characteristic equation

$$\zeta + A + B\frac{1 - e^{\zeta}}{\zeta} = 0. \tag{60}$$

Since we are interested in the location of the change of stability we look for solutions of (60) with vanishing real part. Thus we set $\zeta = is$ with s real, and compare the real and imaginary parts. This yields a parameterized curve in the A, B plane of the form

$$A = \frac{s \sin s}{1 - \cos s}, \qquad B = \frac{s^2}{1 - \cos s} \tag{61}$$

(a part of this curve is the upper border of the stable region in Figure 4). When crossing this curve at $s \neq 0$, the real part of a pair of complex conjugate eigenvalues changes sign. Thus we expect a Hopf-bifurcation to occur. It is easy to see that the integral equation (59) has nontrivial constant solutions if $A + B = 0$. This second curve is also indicated in the stability diagram Figure 4 (lower border of the stable region). Crossing this second curve, a bifurcation of the steady-state solutions is expected.

Figure 4 is a detail of the stability diagram (depending on the parameter values A and B). The formulation (60) allows us to reduce the stability analysis to two parameters A and B. These parameters combine many of the parameters of the model. The price of the reduction is that the parameters A and B cannot be interpreted easily from an application point of view.

In Figure 4 some possibilities for varying the original parameters are indicated. This will be discussed in the next section.

5.2. Numerical analysis. The main purpose here is to study in more detail the behavior of the solutions when stability is lost. The two examples that follow correspond to the situation on the upper and the lower borders, respectively, of the stable region in the stability diagram Figure 4.

As long as only solutions with positive flow velocities are involved, the density ρ has its boundary value ρ_l for $x < x_q$. Thus for solving the continuity equation we can reduce the domain of the numerical solution to the interval $(x_q, 1)$ and use the interface condition for ρ at x_q as an (inflow)-boundary condition. The discretization of the continuity and integrated momentum equation is just straightforward. We have used here a second order method.

For the numerical example we have taken $p_{dv} = 40$, $x_q = 0.5$, $\alpha = 0$, and $q = 1$, and varied $\Delta p = p_r - p_l$. It can be verified that the stationary solution is stable for $\Delta p < \Delta p^* = -0.254$ and unstable for $\Delta p > \Delta p^*$. This is in agreement with the linear stability analysis. The corresponding values of the parameters A and B are indicated in the stability diagram (Figure 4).

Taking the unstable stationary solution with a small perturbation as the initial condition, we observe that the transient solution oscillates with increasing amplitude around the stationary solution until negative velocities are obtained. In no case could a stable limit cycle with only positive velocities be obtained. This indicates that the Hopf-bifurcation at $\Delta p = \Delta p^*$ is subcritical.

In order to compute the (unstable) limit cycle we formulate a fixed point problem: We take the values of the density ρ_i at the grid points $x_i = x_q + (1 - x_q)i/N$ at $t = 0$ and the period t_p as unknowns. Thus the following equations have to be satisfied:

$$\rho(x_i, t_p) = \rho_i, \quad i = 1, \ldots, N, \quad \text{and} \quad v(t_p) = v(0). \tag{62}$$

The pressure difference Δp and the initial velocity $v(0)$ are prescribed appropriately. The resulting equations are solved using the NAGLIB routine C05NBF [NAG]. In the examples given here the number of grid points is N = 100 and 200, respectively.

The minimum and the maximum of the resulting limit cycle are indicated in Figure 5. Thus a subcritical Hopf-bifurcation is verified. However, the limit cycle with strictly positive velocity values exists only for a narrow pressure interval. Neglecting the modulus function in the friction term and the positivity of the density, this unstable limit cycle can be continued to negative velocities.

In order to obtain physically relevant solutions for negative velocities, a more sophisticated numerical analysis is necessary. Taking a conventional finite volume

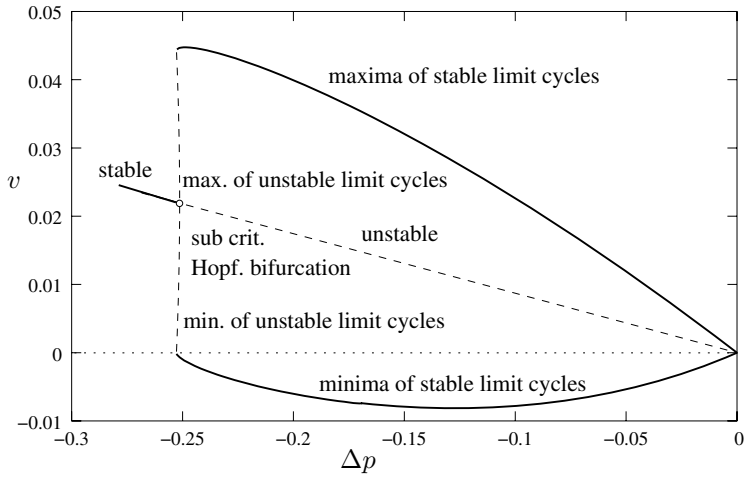


FIG. 5. *Numerical solution for $p_{dv} = 40$, $q = 1$, $\Delta x_0 = 0.5$, $u_L = 1$. The steady-state solution is stable for $\Delta p < \Delta p^*$. The limit cycles are represented by their maximum and minimum velocities.*

method or finite different scheme for the entire solution interval $x \in (0,1)$ may produce negative densities. In order to avoid this problem we employ the method of characteristics. Considering the jump condition of the density at $x = x_q$ for negative values of v which are larger than $-q$, we conclude that the only possibility of a nonnegative density is that $\rho(x_q-) = \rho(x_q+) = 0$ if $-q < v < 0$ holds. In that case the characteristics left of x_q point in the negative direction, and thus there will be a jump discontinuity of the density at a location $x_0(t) < x_q$, which moves with the fluid velocity $v(t)$ left of x_q. Thus $\rho(x,t) = 0$ for $x_0(t) < x < x_q$. For $0 < x < x_0(t)$ the density has the inflow value ρ_l at $x = 0$. Thus we construct the following numerical method: For the domain $0 < x < x_q$ the density distribution is completely described by $x_0(t)$, as long as $v > -q$. For $x > x_q$ we define the characteristics by

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{x}_i(t) = v + q, \quad \bar{x}_i(0) = x_i = x_q + (1 - x_q)i/N, \quad \rho_i = \rho(x_i, 0). \tag{63}$$

Every time a characteristic $\bar{x}_j(t)$ leaves the domain of consideration at $x = 1$ a new characteristic will be introduced at $x = x_q$, and ρ_j will be set to $\rho(x_q+, t)$, which is zero if $v < 0$, or $\rho(x_q-, t)v/(q + v)$ if $v > 0$. Note that $\rho(x_q-, t) = \rho_l$ if $x_0 = x_q$, or $\rho(x_q-, t) = 0$ if $x_0 < x_q$.

Employing this method in the case of the unstable stationary solution, the transient solution tends to a stable limit cycle which has time intervals with negative velocities. In Figure 5 the maximum and minimum values of the limit cycle are shown. This stable limit cycle connects to the unstable limit cycle where the minimum of the velocity is zero.

As a second example we consider an inclined tunnel. We assume p_{dv}, x_q, and q have the same values as in the first example. We vary $f_d \sin\alpha$ and $\Delta p = p_r - p_l$ such that $v_s = 0.023$ is a stationary solution. However, (53) for the stationary flows has in general more than one solution (as known from [G1]). We compute a second solution branch, which intersects the first (constant) solution branch when $A + B = 0$ holds; see Figure 4. The stability parameters A and B for both solution branches are indicated in the stability diagram, and we observe that at the intersection of the two branches a stability exchange occurs (Figure 6).

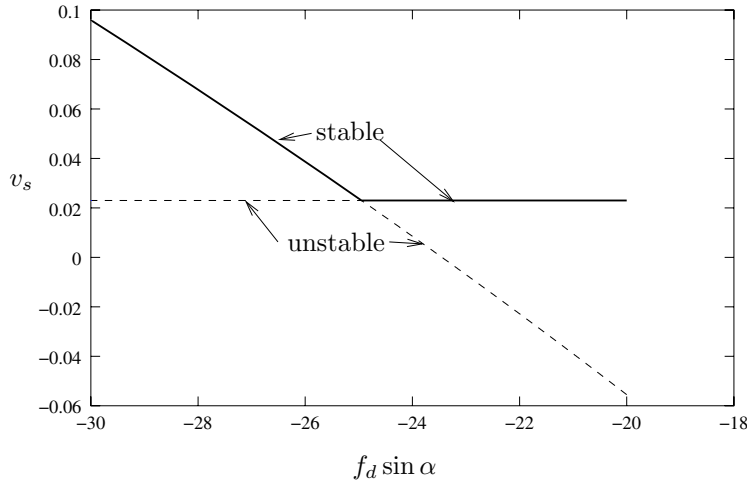


FIG. 6. *Variation of the buoyancy parameter $f_d \sin\alpha$. The pressure Δp is chosen such that $v_s = 0.023$ is a stationary solution. A second solution branch exists, which intersects the first.*

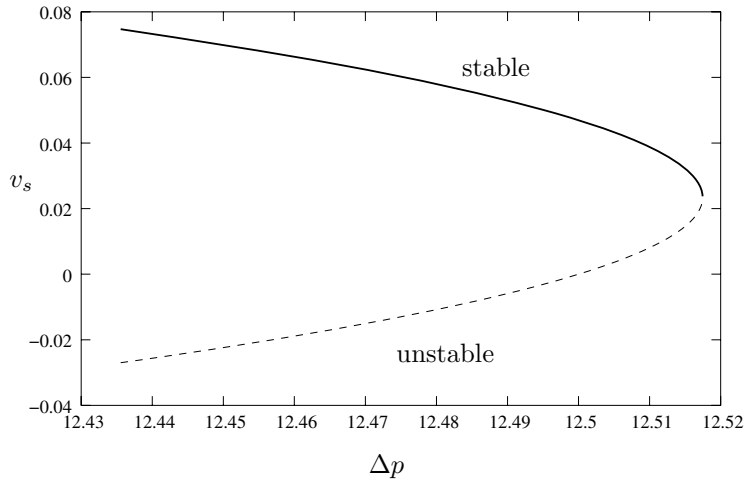


FIG. 7. *Variation of Δp for $f_d \sin\alpha = -25$. For Δp less than a critical value, two solution branches exist.*

In the third example we assume $f_d \sin\alpha = -25$ fixed and vary Δp. For Δp less than a limiting value Δp_c two solution branches exist. The upper one corresponds to a stable and the lower one to an unstable solution. At the limiting value Δp_c both solution branches are connected and terminate there (Figure 7).

Thus we have verified by numerical examples that, while crossing the upper limit of the stability domain, a subcritical Hopf-bifurcation occurs, and while crossing the lower bound of the stability domain, a bifurcation of stationary solution occurs.

6. Conclusions. We have studied the global dynamics for solutions of model equations for the description of tunnel fires. Also, the stability of the stationary solutions has been investigated. We see at least numerically that Hopf-like bifurcations occur and induce oscillation with large amplitudes and even, most dangerous for fire fighters, flow reversal can occur. The critical cases can be identified by a relatively small outside pressure difference. Since the Hopf-bifurcation seems to be subcritical, the large amplitude limit cycle can be attained even at conditions where the steady-state solution is still stable. Thus the presented analysis delivers a simple tool for determining the stability of flow conditions during a tunnel fire.

REFERENCES

[BJL] G. B. BRANDT, S. F. JAGGER, AND C. J. LEA, *Fires in tunnels*, Phil. Trans. R. Soc. London A, 356 (1998), pp. 2873–2906.

[Bu] T. A. BURTON, *Volterra Integral and Differential Equations*, Math. Sci. Engrg. 167, Academic Press, New York, 1983.

[F] R. FRIEDMAN, *An international survey of computer models for fire and smoke*, J. Fire Protect. Engrg., 4 (1992), pp. 81–92.

[FiT] *European Thematic Network on Fire in Tunnels*, web portal at http://www.etnfit.net/ (accessed 6th September, 2004).

[G1] I. GASSER, *An asymptotic-induced one-dimensional model to describe fires in tunnels* II: *The stationary model*, Math. Methods Appl. Sci., 26 (2003), pp. 1327–1347.

[G4] I. GASSER, *On the mathematics of tunnel fires*, GAMM-Mitt., 26 (2003), pp. 109–126.

[GSte] I. GASSER AND H. STEINRÜCK, *On the stability of solutions of a tunnel fire model*, Proc. Appl. Math. Mech., 3 (2003), pp. 444–445.

[GSte2] I. Gasser and H. Steinrück, *On the existence of transient solutions of a tunnel fire model*, Comm. Math. Sci., 4 (2006), pp. 609–619.

[GStr] I. Gasser and J. Struckmeier, *An asymptotic-induced one-dimensional model to describe fires in tunnels*, Math. Methods Appl. Sci., 25 (2002), pp. 1231–1249.

[GST1] I. Gasser, J. Struckmeier, and I. Teleaga, *Modelling and simulation of fires in vehicle tunnels*, Internat. J. Numer. Methods Fluids, 44 (2004), pp. 277–296.

[GST2] I. Teleaga, I. Gasser, and J. Struckmeier, *Modelling and simulation of fires in vehicle tunnels*, Proc. Appl. Math. Mech., 3 (2003), pp. 400–401.

[NAG] *Nag Fortran Library Manual*, Mark 21, The Numerical Algorithms Group, Oxford, UK, 2005; online at http://www.nag.com.

[OC] S. M. Olenick and D. J. Carpenter, *An updated international survey of computer models for fire and smoke*, J. Fire Protect. Engrg., 13 (2003), pp. 87–110.

[P] *Fire and Smoke Control in Road Tunnels*, Report of the PIARC Committee on Road Tunnels, World Road Association, La Defense, France, 1999.

SIAM J. APPL. MATH.
Vol. 66, No. 6, pp. 2204–2205

ERRATUM: GLOBAL STABILITY IN CHEMOSTAT-TYPE COMPETITION MODELS WITH NUTRIENT RECYCLING*

SHIGUI RUAN† AND XUE-ZHONG HE‡

Abstract. This note corrects some typos and errors in the paper [S. Ruan and X.-Z. He, *SIAM J. Appl. Math.*, 58 (1998), pp. 170–192].

Key words. competition, global stability, nutrient recycling

DOI. 10.1137/06065876X

The main error in the paper [1] is that conditions on the matrix B should be replaced by conditions on the matrix AB at four separate places.

1. On page 177, condition (iii) in Theorem 2.8 should read

(iii) *the matrix AB is semipositive definite.*

2. On page 182, there were a few typos in Theorem 3.8 and an error in condition (v). The corrected version of the theorem is as follows.

THEOREM 3.8. *Assume that*

(i) *system* (3.1) *has a positive equilibrium* $E^* = (S^*, N_1^*, N_2^*)$*;*

(ii) $D + D_i < m_i$, $b_i D_i < \mu_i p(S_i^*)$, $i = 1, 2$*;*

(iii) $T_f < \infty$, $T_i^* = (1/d_i^*) \int_0^\infty F(s)[e^{d_i^* s} - 1]\, ds < \infty$ *with* $d_i^* := (D + D_i) + \sum_{j=1}^2 \delta_{ij} m_j$, $i = 1, 2$*;*

(iv) $b_i D_i[(m_i + \sum_{j=1}^2 \delta_{ij} N_j^*) T_i^* + m_i T_f]/2 < \mu_i$, $i = 1, 2$*;*

(v) *The matrix AB is semipositive definite, where* $A = \mathrm{diag}(\alpha_1, \alpha_2)$ *with* $\alpha_i = [\mu_i p(S_i^*) - b_i D_i]/m_i$ $(i = 1, 2)$ *and* $B = (b_{ij})_{2\times 2}$, $b_{ij} \geq 0$ *defined by*

$$b_{ij} = \begin{cases} \delta_{ii} - \frac{T_f m_i}{2[\mu_i p(S^*) - b_i D_i] N_i^*} \sum_{j=1}^2 b_j D_j \delta_{ji} m_j & \text{if } i = j, \\ \delta_{ij} & \text{if } i \neq j. \end{cases} \tag{3.25}$$

Then E^* is global asymptotically stable.

3. On page 186, condition (iii) in Theorem 4.3 should read

(iii) *the matrix AB is semipositive definite.*

4. On page 188, Theorem 4.6 should read as follows.

THEOREM 4.6. *Assume that*

(i) *system* (4.11) *has a positive equilibrium* $E^* = (S^*, N_1^*, \ldots, N_n^*)$*;*

(ii) $D + D_i < m_i$, $b_i D_i < \mu_i p(S_i^*)$, $i = 1, 2, \ldots, n$*;*

(iii) $T_f < \infty$, $T_i^* = (1/d_i^*) \int_0^\infty F(s)[e^{d_i^* s} - 1]\, ds < \infty$, $i = 1, 2, \ldots, n$, *where* $d_i^* = (D + D_i) + \sum_{j=1}^n \delta_{ij} m_j$*;*

(iv) $b_i D_i[(m_i + \sum_{j=1}^n \delta_{ij} N_j^*) T_i^* + m_i T_f]/2 < \mu_i$, $i = 1, 2, \ldots, n$*;*

(v) *the matrix AB is semipositive definite, where* $A = \mathrm{diag}(\alpha_i)_{n\times n}$ *with* $\alpha_i =$

*Received by the editors May 3, 2006; accepted for publication (in revised form) August 7, 2006; published electronically October 24, 2006.

http://www.siam.org/journals/siap/66-6/65876.html

†Department of Mathematics, University of Miami, P.O. Box 249085, Coral Gables, FL 33124-4250 (ruan@math.miami.edu).

‡School of Finance and Economics, University of Technology, P.O. Box 123, Broadway NSW 2007, Sydney, Australia (tony.he1@uts.edu.au).

$[\mu_i p(S_i^*) - b_i D_i]/m_i$ $(i = 1, \ldots, n)$ *and* $B = (b_{ij})_{n\times n}$ *with* $b_{ij} \geq 0$ *defined as follows:*

$$b_{ij} = \begin{cases} \delta_{ii} - \frac{T_f m_i}{2[\mu_i p(S^*) - b_i D_i] N_i^*} \sum_{j=1}^n b_j D_j \delta_{ji} m_j & \text{if} \quad i = j, \\ \delta_{ij} & \text{if} \quad i \neq j. \end{cases}$$

Then E^* *is globally asymptotically stable.*

Acknowledgment. The authors would like to thank Karl Hadeler and Julia Hesseler for noticing the errors.

REFERENCE

[1] S. Ruan and X.-Z. He, *Global stability in chemostat-type competition models with nutrient recycling*, SIAM J. Appl. Math., 58 (1998), pp. 170–192.